The Marvelous World of tRNAs

From Accurate Mapping to Chemical Modifications

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Informatik

Vorgelegt

von Master of Science Bioinformatik Anne Hoffmann
geboren am 14.07.1987 in Halle (Saale)

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler, Universität Leipzig
2. Prof. Dr. Andrew Torda, Universität Hamburg

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 09.06.2020 mit dem Gesamtprädikat *summa cum laude*

Bibliographic Description

Title:	The Marvelous World of tRNAs
Subtitle:	From Accurate Mapping to Chemical Modifications
Type:	Dissertation
Author:	Anne Hoffmann
Year:	2020
Professional discipline:	Computer Science
Language:	English
Pages in the main part:	173
Chapter in the main part:	13
Number of Figures:	54
Number of Tables:	13
Number of Appendices:	2
Number of Citations:	423
Key Words:	tRNA, tRNA Modifications, Read Mapping, RNA-seq, tRNA Evolution, nm-tRNAs

This thesis is based on the following publications.

[1] A. Hoffmann*, L. Erber*, H. Betat, P. F. Stadler, M. Mörl, and J. Fallmann. "Changes of the tRNA modification pattern during the development of *Dictyostelium discoideum*". In: *RNA Biol.* under review (2020).

[2] A. Hoffmann*, C. Lorenz, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "Temperature Dependence of Bacterial tRNA Modifications". In: In preparation (2020).

[3] L. Erber, A. Hoffmann, J. Fallmann, H. Betat, S. Prohaska, P. F. Stadler, and M. Mörl. "*Dictyostelium discoideum*: Unusual occurrence of two active CCA-adding enzymes". In: *IJMS* accepted (2020).

[4] S. Hoser*, A. Hoffmann*, A. Meindl, M. Gamper, S. Bernhart, L. Müller, M. Misslinger, M. Hoelzl, K. Perfler, K. Singer, M. Ploner, H. Lindner, H. Schaal, P. F. Stadler, and A. Hüttenhofer. "Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing". In: *Genom Biol.* accepted (2020).

[5] L. Erber*, A. Hoffmann*, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "LOTTE-seq (Long hairpin oligonucleotide based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing". In: *RNA Biol.* 0 (2019), pp. 1–10. *DOI*: 10.1080/15476286.2019.1664250.

[6] A. Hoffmann, J. Fallmann, E. Vilardo, M. Mörl, P. F. Stadler, and F. Amman. "Accurate mapping of tRNA reads". In: *Bioinf.* 34.7 (2018), pp. 1116–1124. *DOI*: 10.1093/bioinformatics/btx756.

[7] S. J. Berkemer, A. Hoffmann, C. R. A. Murray, and P. F. Stadler. "SMORE: Synteny Modulator of Repetitive Elements". In: *Life* 7.4 (2017), p. 42. *DOI*: 10.3390/life7040042.

[8] C. A. Velandia-Huerto, S. J. Berkemer, A. Hoffmann, N. Retzlaff, L. C. Romero Marroquín, M. Hernández-Rosales, P. F. Stadler, and C. I. Bermúdez-Santana. "Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies". In: *BMC Genom.* 17.1 (2016), p. 617. *DOI*: 10.1186/s12864-016-2927-4.

*The authors share first authorship.

Abstract

Since the discovery of transfer RNAs (tRNAs) as decoders of the genetic code, life science has transformed. Particularly, as soon as the importance of tRNAs in protein synthesis has been established, researchers recognized that the functionality of tRNAs in cellular regulation exceeds beyond this paradigm. A strong impetus for these discoveries came from advances in large-scale RNA sequencing (RNA-seq) and increasingly sophisticated algorithms. Sequencing tRNAs is challenging both experimentally and in terms of the subsequent computational analysis. In RNA-seq data analysis, mapping tRNA reads to a reference genome is an error-prone task. This is in particular true, as chemical modifications introduce systematic reverse transcription errors while at the same time the genomic loci are only approximately identical due to the post-transcriptional maturation of tRNAs. Additionally, their multi-copy nature complicates the precise read assignment to its true genomic origin. In the course of the thesis a computational workflow was established to enable accurate mapping of tRNA reads. The developed method removes most of the mapping artifacts introduced by simpler mapping schemes, as demonstrated by using both simulated and human RNA-seq data. Subsequently, the resulting mapping profiles can be used for reliable identification of specific chemical tRNA modifications with a false discovery rate of only 2%. For that purpose, computational analysis methods were developed that facilitates the sensitive detection and even classification of most tRNA modifications based on their mapping profiles. This comprised both untreated RNA-seq data of various species, as well as treated data of *Bacillus subtilis* that has been designed to display modifications in a specific read-out in the mapping profile. The discussion focuses on sources of artifacts that complicate the profiling of tRNA modifications and strategies to overcome them. Exemplary studies on the modification pattern of different human tissues

and the developmental stages of *Dictyostelium discoideum* were carried out. These suggested regulatory functions of tRNA modifications in development and during cell differentiation.

The main experimental difficulties of tRNA sequencing are caused by extensive, stable secondary structures and the presence of chemical modifications. Current RNA-seq methods do not sample the entire tRNA pool, lose short tRNA fragments, or they lack specificity for tRNAs. Within this thesis, the benchmark and improvement of LOTTE-seq, a method for specific selection of tRNAs for high-throughput sequencing, exhibited that the method solves the experimental challenges and avoids the disadvantages of previous tRNA-seq protocols. Applying the accurate tRNA mapping strategy to LOTTE-seq and other tRNA-specific RNA-seq methods demonstrated that the content of mature tRNAs is highest in LOTTE-seq data, ranging from 90% in *Spinacia oleracea* to 100% in *D. discoideum*.

Additionally, the thesis addressed the fact that tRNAs are multi-copy genes that undergo concerted evolution which keeps sequences of paralogous genes effectively identical. Therefore, it is impossible to distinguish orthologs from paralogs by sequence similarity alone. Synteny, the maintenance of relative genomic positions, is helpful to disambiguate evolutionary relationships in this situation. During this thesis a workflow was computed for synteny-based orthology identification of tRNA genes. The workflow is based on the use of pre-computed genome-wide multiple sequence alignment blocks as anchors to establish syntenic conservation of sequence intervals. Syntenic clusters of concertedly evolving genes of different tRNA families are then subdivided and processed by cograph editing to recover their duplication histories. A useful outcome of this study is that it highlights the technical problems and difficulties associated with an accurate analysis of the evolution of multi-copy genes. To showcase the method, evolution of tRNAs in primates and fruit flies were reconstructed.

In the last decade, a number of reports have described novel aspects of tRNAs in terms of the diversity of their genes. For example, nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) have been reported whose presence provokes intriguing questions about their functionality. Within this thesis an annotation strategy was developed that led to the identification of 335 and 43 novel nm-tRNAs in human and mouse, respectively. Interestingly, downstream analyses showed that the localization of several nm-tRNAs in introns and the over-representation of conserved RNA-binding sites of proteins involved in splicing suggest a potential regulatory function of intronic nm-tRNAs in splicing.

Zusammenfassung

Nach der Entdeckung von Transfer-RNAs (tRNAs) als Dekodierer des genetischen Codes, wurde postuliert, dass die Funktionalität von tRNAs über dieses Paradigma hinausgeht. Ein Anstoß dafür war die Entwicklung der RNA-Sequenzierung (RNA-seq) und zunehmend ausgereiftere Algorithmen. Die Sequenzierung von tRNAs ist sowohl experimentell als auch im Hinblick auf die anschließende computergestützte Analyse eine Herausforderung. Hinsichtlich der RNA-seq Datenanalyse ist das Mapping der gelesenen tRNA Fragmente (Reads) auf das Referenzgenom anfällig für Fehler. Dies gilt insbesondere dann, wenn chemische Modifikationen systematische Fehler bei der reversen Transkription einführen, während gleichzeitig die Genloci aufgrund der posttranskriptionellen Reifung von tRNAs nur annähernd identisch sind. Da tRNAs in mehreren Genkopien vorliegen können, erschwert dies zusätzlich das zielgenaue Mapping von tRNA-Reads zu ihrem genomischen Ursprung. Im Rahmen dieser Dissertation wurde ein computergestützter Arbeitsablauf etabliert, der ein präzises tRNA-Read Mapping ermöglicht. Die entwickelte Methode beseitigt die meisten Mapping-Artefakte, die durch einfachere Mapping-Schemata eingeführt werden, wie die Verwendung von simulierten und menschlichen RNA-seq Daten demonstrierte. Anschließend können die resultierenden Mapping-Profile für die zuverlässige Identifizierung spezifischer chemischer tRNA-Modifikationen verwendet werden, mit einer Falscherkennungsrate von nur 2%. Zu diesem Zweck wurden computergestützte Analysemethoden entwickelt, die den sensitiven Nachweis und sogar die Klassifizierung der meisten tRNA-Modifikationen, basierend auf ihren Mapping-Profilen, ermöglichen. Dazu wurden sowohl RNA-seq Daten von verschiedenen Arten als auch chemisch behandelte Daten von *Bacillus subtilis* analysiert, die bestimmte tRNA-Modifikationen konvertieren, so dass diese im Mapping-Profil nachweisbar werden. Das Diskussionsthema konzentriert sich auf

Artefaktquellen, die das Detektieren von tRNA-Modifikationen erschweren und auf Strategien, um diese Hindernisse zu überwinden. Exemplarische Studien über das Modifikationsmuster verschiedener menschlicher Gewebe und über die Entwicklungsstadien von *Dictyostelium discoideum* wurden ergänzend durchgeführt. Diese suggerieren eine regulatorische Funktionen von tRNA-Modifikationen während der Entwicklung und Zelldifferenzierung.

Die wohl größte Herausforderung in der tRNA-seq besteht darin, dass stabile Sekundärstrukturen und chemische Modifikationen in tRNAs vorhanden sind. Verfügbare RNA-seq Methoden untersuchen nicht die Gesamtheit der tRNAs, verlieren kurze Fragmente oder zeigen keine tRNA-Spezifität. Innerhalb dieser Dissertation ergab das Benchmarking von LOTTE-seq, was die spezifische Auswahl von tRNAs für die Sequenzierung ermöglicht, dass diese Methode die experimentellen Herausforderungen löst. Ein Vergleich von LOTTE-seq mit anderen tRNA-seq Methoden demonstrierte, dass der Gehalt an reifen tRNAs in LOTTE-seq am höchsten ist und von 90% in Spinat bis 100% in *D. discoideum* reicht.

In einer zusätzlichen Studie addressiert diese Dissertation die Problematik, dass tRNAs in mehreren Genkopien vorliegen. Diese Genkopien unterliegen einer konzertierten Evolution, welche die Sequenzen von paralogen Genen effektiv identisch hält. Demnach ist es unmöglich, orthologe von paralogen Genen allein durch ihre Sequenzähnlichkeit zu unterscheiden. Für die Disambiguierung evolutionärer Beziehungen sind Syntänieinformationen, was die Aufrechterhaltung releativer genomischer Positionen beschreibt, hilfreich. Dementsprechend wurde ein Arbeitsablauf für die Identifizierung orthologer Beziehungen von tRNAs implementiert. Dieser basiert auf der Verwendung von präcomputierten genomweiten multiplen Seqeunzalignmentblöcken als Anker, um eine syntänische Konservierung von Sequenzintervallen zu ermöglichen. Syntänische Cluster von konzentierten tRNA-Genen werden anschliessend durch Co-Graph Editierung weiterverarbeitet, so dass ihre Duplikationshistorien rekonstruierbar sind. Um die Methodik vorzustellen, wurde die tRNA-Evolution in Primaten und Fruchtfliegen nachgebildet.

Darüber hinaus berichteten Studien über nukleär kodierte mitochondriale Transfer-RNAs (nm-tRNAs), deren Anwesenheit interessante Fragen über ihrer Funktionalität aufwirft. Im Rahmen dieser Dissertation wurde eine Annotationsstrategie entwickelt, die die Identifizierung von unentdeckten nm-tRNAs in Mensch (# 335) und Maus (# 43) ermöglichte. Interessanterweise legten weiterführende Analysen dar, dass viele nm-tRNAs eine Überrepräsentation von Bindungsstellen für Proteine aufweisen, die beim Spleißen interagieren.

Acknowledgment

Contents

CHAPTER 1

Motivation

Contents

Transfer RNAs (tRNAs) are among the most ancient ribonucleic acids (RNAs) in the world. Research on tRNAs can be traced back to the mid-1950s, when Francis Crick first hypothesized the existence of tRNAs in his so-called "adapter hypothesis" [1]. Crick proposed that each amino acid is first attached to its own specific adapter which mediates the translation of the RNA alphabet into the protein alphabet. The existence of these adapters was discovered by Hoagland and Zamenick in 1958. They observed in a cell-free rat liver system that a radio labeled amino acid (^{14}C-leucine) attached to an RNA acceptor was transferred to a microsomal protein [2, 3]. They concluded that this type of RNA functions as an intermediate carrier of amino acids in protein synthesis. These findings provided the basis for our later understanding of the role of tRNAs in protein biosynthesis.

1.1 tRNAs as Keyplayers in Protein Biosynthesis

A deoxyribonucleic acid (DNA) molecule is not just a nucleotide sequence of adenines (As), cytosines (Cs), guanines (Gs) and thymines (Ts). Instead, DNA is the genetic material of all organisms on Earth. The genetic information in a genome is mediated by genes. Genes are specific regions of the DNA sequence encoding for non-coding RNAs (ncRNAs) or proteins. Beside tRNAs, non-coding RNAs play a key role in regulating different cellular activities, e.g., microRNAs (miRNAs) in post-transcriptional gene silencing [4], small nuclear RNAs (snRNAs) in intron splicing [5] and Y RNAs in regulation of DNA replication and RNA processing [6].

During protein biosynthesis the genetic information flows from DNA to RNA and lastly from RNA to protein. This directional flow of information is known as the central dogma of molecular biology [8]. A gene that encodes a protein is expressed in two broad steps (see **Fig. 1**). In the first step, the so-called transcription, genes are transcribed by assembling a new sequence of single-stranded RNA using the coding region of the gene as template. The process of transcription takes place in the nucleus. In all species, transcription begins with the binding of RNA polymerase and other transcription factors to a conserved DNA sequence downstream of the gene referred to as promoter. Activation of the RNA polymerase complex facilitates transcription initiation followed by elongation of the transcript. During elongation, the DNA unwinds by disrupting hydrogen bonds between the bases of the opposite DNA strands by the enzyme helicase. Thus, the RNA polymerase can read a single template strand of the gene

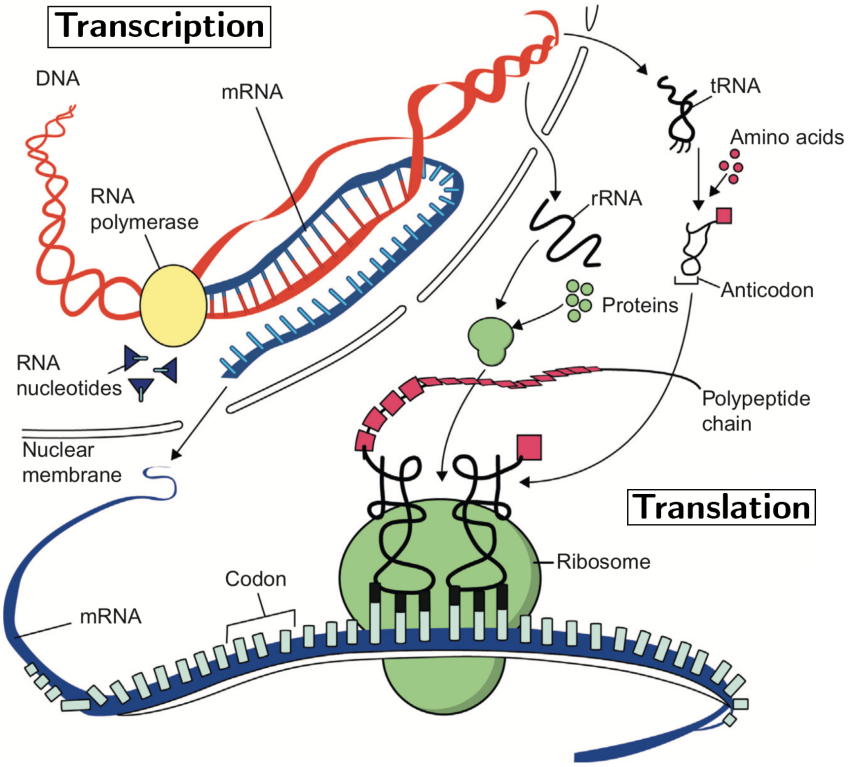

Figure 1: Scheme of protein biosynthesis. A protein-coding gene is expressed in two steps: transcription and translation. During transcription, genes are copied into a complementary, antiparallel messenger RNA (mRNA) strand by the enzyme RNA polymerase. In eukaryotes, transcription takes place in the nucleus. The mRNA is afterwards exported into the cytoplasm for translation. During the translational process mRNA binds to ribosomes, where the mRNA is decoded step-wise into the proteins polypeptide chain. Ribosomes consist of a small subunit, which reads the mRNA, and a large subunit, which attaches amino acids to synthesize the polypeptide chain. Both subunits consist of ribosomal RNAs (rRNAs) and ribosomal proteins. The genetic code of the mRNA is given by codons (nucleotide triplets). For each codon a tRNA charged with a codon specific amino acid binds to the codon through its complementary anticodon. Thus, the ribosome adds the amino acid of the tRNA to the polypeptide chain. This figure is modified after Lodish et al. [7].

and adds complementary nucleotides to the growing RNA chain in the 5'-to-3'-direction. The RNA transcript carries the same information as the non-template strand of the gene except that the nucleotide uracil replaces nucleotide thymine. In bacteria, the new synthesized RNA transcripts can act as messenger RNA (mRNA) right away. In contrast, the RNA transcript is called a precursor mRNA (pre-mRNA) and must undergo processing to become a mature

mRNA. Both ends of a pre-mRNA are modified by adding a cap and a poly-A-tail to the 5'- and 3'-end, respectively. The 5'-cap is a methylated guanine nucleotide that protects the transcript from degradation. It also promotes the binding of the mRNA to the ribosome which is necessary for the next expression step. The 3'-poly-A-tail consists of about 200 adenine nucleotides that stabilize the transcript and support the export of the mature mRNA through nuclear pores into the cytoplasm. Most higher eukaryotes undergo splicing of pre-mRNAs. In this process, segments of the pre-mRNA (introns) are cut out by the spliceosome and the remaining segments (exons) are joined together to build the mature mRNA. Splicing does allow for a process called alternative splicing, in which more than one mRNA can be made from the same gene. In alternative splicing, splicing positions may be altered for each pre-mRNA, e. g., introns can be retained or exons can by skipped or extended. This results in different mature mRNAs, each of which translates into a protein with a different structure [7]. Once the mature mRNA is exported and integrated into the cytoplasm, the mRNA is decoded to synthesize a protein in a process called translation. Here, the mRNA binds to ribosomes, where the mRNA sequence is step-wise translated into specific series of amino acids with the help of tRNAs. An mRNA can be translated by its genetic code (see **Fig. 2**). The genetic code is given by a series of three consecutive nucleotides (triplets) called codons. tRNAs contain an anticodon which is complementary to one of the possible mRNA codons and carries an amino acid specified by the codons. The tRNA is charged with an amino acid tRNA by an aminoacyl-tRNA synthetase. The complementary binding of codon and anticodon enables the ribosomes to connect the amino acids carried by the tRNA. After termination, the linear amino acid chain folds into a tertiary structure and undergoes processing to become a functional protein.

The genetic code is given by 64 possible codons, 61 of which are sense codons that collectively encode 20 amino acids (see **Fig. 2**). Translation is initiated by the sense codon methionine (AUG). The remaining 3 codons are nonsense codons (UAG, UGA, or UAA) which stop the translation. Not every codon can be matched by a tRNA with an exact complementary anticodon. This feature is possible because of the wobble rules [10]. The wobble hypothesis was proposed by Crick in 1966 which postulated that a G-U pair can be functional in the third position of the codon and inosine in the anticodon can recognize U, C and A. Thus, different minimal sets of chemically diverse tRNAs which are acetylated by the same amino acid (isoacceptor tRNAs) cumulatively decode all 61 sense codons. The minimum set of

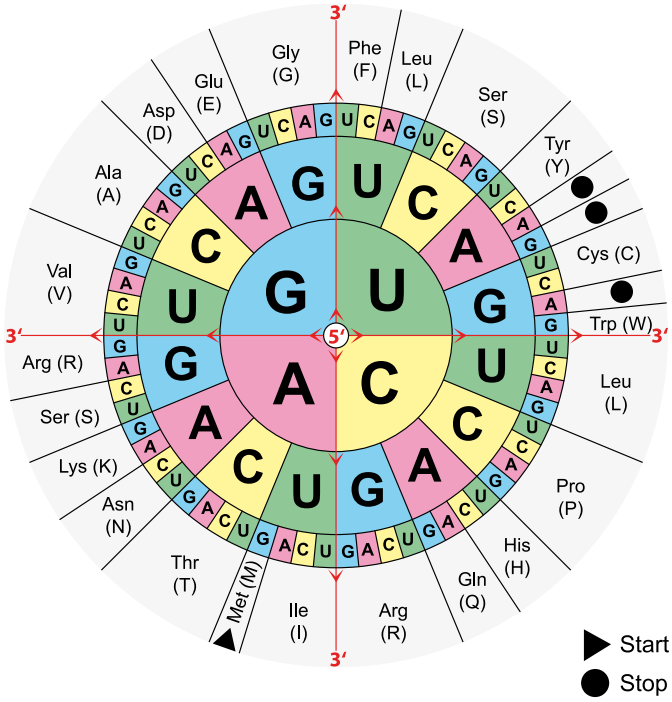

Figure 2: The genetic code illustrated as codon wheel. Starting from the four innermost letters and working to the outermost ring, this illustration shows which codon (three-letter base sequence) encodes which amino acid. The genetic code shows 20 amino acid to which the 64 possible codons corresponds. AUG is the start codon and UAG, UGA, or UAA are stop codons for translation during protein biosynthesis. The figure is taken from [9].

isoacceptor tRNAs is 32 [10], but only a few bacterial and archaeal organisms encode deviating numbers of tRNAs [11, 12]. Actual numbers of nuclear encoded tRNA genes vary greatly between organisms. For example, *S. pombe* contains 171 nuclear tRNAs, encoding for 41 distinct isoacceptor tRNAs. In contrast, the zebrafish *D. rerio* contain 12,794 tRNA genes encoding for 55 isoacceptors. The large number of tRNA genes arises because the isoacceptors are often encoded by an entire gene family. Thus, isoacceptors can be further subdivided into multiple isodecoder tRNAs that carry the same anticodon but differ in their sequence outside the anticodon [13].

1.2 Outline of the Thesis

This thesis is divided into four major parts. Part I underlines the biological background that inspired this thesis. Special peculiarities and structural features of different tRNA types as well as their biological relevance are elucidated in Chapter 2. Chapter 3 describes the individual processing steps of tRNA biogenesis. Special attention is given to possible chemical modifications that can alter tRNAs and their biological functions. Chapter 4 provides an overview of evolutionary events of homologous tRNAs.

Part II expounds on the technical background that is intended to provide a basic understanding of the tRNA analysis strategies developed for this thesis. Chapter 5 outlines the basic steps for the analysis of RNA sequencing (RNA-seq) data with respect to the underlying challenge of constructing a tRNA sequence library for RNA-seq and the difficulties of tRNA read mapping. Our developed theory on how to construct evolutionary events of homologous tRNAs in different species is explained in Chapter 6. This chapter is based on the publications Velandia-Huerto et al. [376] titled *Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies* and Berkemer et al. [377] titled *SMORE: Synteny Modulator of Repetitive Elements*.

In Part III, Chapter 7 describes our developed analysis methods in detail. This includes the tRNA and nm-tRNA annotation methods, a precise tRNA read mapping workflow, strategies to detect different types of tRNA modifications, and the creation of a synteny map for tRNA orthology identification.

Part IV shows the application of our developed analysis methods and presents the conclusion that can be drawn from their results. Our developed best-practice mapping strategy of tRNA reads is benchmarked in Chapter 8 based on the publication *Accurate mapping of tRNA reads* by A. Hoffmann et al. [397]. Chapter 9 based on L. Erber and A. Hoffmann et al. [406] with the title *LOTTE-seq (Long hairpin oligonucleotide-based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing*. In this chapter our newly developed tRNA-seq method is presented and evaluated on different species of all domains of life. Our analysis strategies to detect various types of tRNA modifications is applied to different RNA-seq data in Chapter 10. Included are the results of the following publications A. Hoffmann et al. [397] titled *Accurate mapping of tRNA reads*, L. Erber et al. [408] titled

Dictyostelium discoideum: Unusual occurrence of two active CCA-adding enzymes, L. Erber and A. Hoffmann et al. [406] with the title *LOTTE-seq (Long hairpin oligonucleotide-based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing*, A. Hoffmann and L. Erber et al. [409] titled *Changes of the tRNA modification pattern during the development of Dictyostelium discoideum*, and A. Hoffmann et al. [410] titled *Temperature Dependence of Bacterial tRNA Modifications*. In Chapter 11 the results of our synteny-based orthology identification workflow applied to primates and fruit flies are shown. This chapter is based on the publications Velandia-Huerto et al. [376] with the title *Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies* and Berkemer et al. [377] titled *SMORE: Synteny Modulator of Repetitive Elements*. Our analyses on nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) are given in Chapter 12. This chapter based on S. Hoser and A. Hoffmann et al. [411] with the title *Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing*.

The thesis is concluded in Chapter 13 with respect to their relevance for future research. The distinct appendices provide supplementary information of the respective studies for the interested reader, including further results and data overviews.

1.3 Author Contribution and Use of Personal Pronoun

In scientific writing the use of the personal pronoun "we" is common in order to take into account collective group work, regardless of whether certain parts were contributed by a single individual. Within this thesis the personal pronoun "we" is used as well, since it is based on scientific discussions and results of multiple collaborative projects. This does not invalidate my statement in the Declaration of Independence.

If a chapter is based on a specific publication in which I was involved, this is indicated at the beginning of the chapter. In this case, the content of the publication is not additionally quoted in the following chapter text, unless the part of the work was only done by one of my collaboration partners.

Part I

Biological Background

CHAPTER 2

Structure and Particularities of Different tRNA Types

Contents

The last two decades have seen a revolution in genome sequencing resulting in a high amount of completely or partially sequenced genomes. The number of sequenced organisms range from 13,378 prokaryotes, to 4,581 viruses and 4,132 eukaryotes [22]. Thus, a large diversity of coding and non-coding genes have recently been discovered in the three domains of life. Beside the annotation of *bona fide* tRNA genes, other tRNA types have been detected. It has been revealed, that tRNA structures deviating from the canonical tRNA structure are not a rarity and are ubiquitous in mitochondrial genomes [23]. Another kind of tRNAs are nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) which originated by genomic integration of mitochondrial DNA. nm-tRNAs are not only pseudogens, as originally assumed, but there are also indications of their biological relevance [24]. Even the cleaved fragments of tRNA genes, referred to as tRNA-derived small RNAs (tsRNAs), are dynamic regulators of biological processes and not simply random tRNA degradation products [25]. The following chapter highlights the peculiarities and main structural differences of the individual tRNA types mentioned.

2.1 The Highly Conserved Canonical tRNA Structure

Canonical tRNAs have a cloverleaf-like secondary structure which is highly conserved [28]. In contrast, the sequence can vary greatly between species and single tRNA types. The canonical cloverleaf structure mainly concerns cytosolic tRNAs and is composed of five domains: (i) the acceptor stem, (ii) the dihydrouridine arm (D-arm), (iii) the anticodon arm, (iv) the variable loop (V-loop) and (v) the TΨC-arm (T-arm) [27, 29], see **Fig. 3A**. The acceptor stem is a 7 base pairs (bp) stem made by base pairing the 5'-terminal nucleotides with those of the 3'-end. At the 3'-end of the acceptor stem there is an unpaired nucleotide (tRNA position 73), the so-called discriminator, followed by the CCA triplet. An exception are all histidine tRNAs which contain an additional nucleotide (G-1) at the 5'-end and form a mismatch base pair with the opposite discriminator base [30, 31]. In eukaryotes, G-1 is added post-transcriptionally [32] by a specific guanyl-transferase [33], whereas G-1 is encoded in bacterial species [30, 34]. The D-arm consists of a 4 to 6 bp long stem which ends in a loop of 8 to 11 nucleotides (nts) and takes its name from the modified base dihydrouridine (see Section 3.5) which it often contains. The anticodon arm is a 5 bp stem whose loop with the size of 7 nts contains the

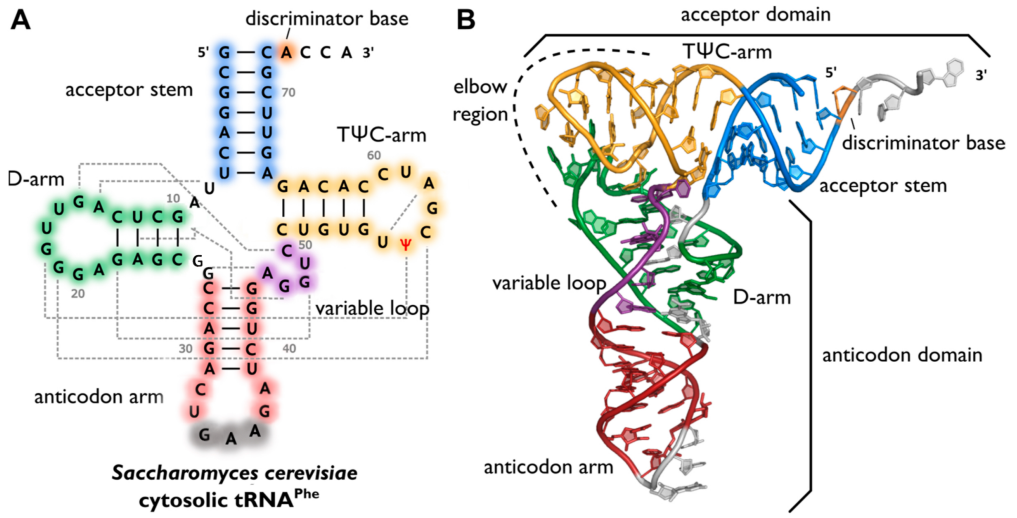

Figure 3: Cloverleaf secondary and L-shape tertiary tRNA structure. (A) Canonical cloverleaf secondary structure of cytosolic transfer RNAs (tRNAs) representing the acceptor stem (blue), anticodon (grey), anticodon arm (red), dihydrouridine-arm (D-arm; green), TΨC-arm (T-arm; yellow), and variable loop (purple). The discriminator base (orange) on the 3'-end is followed by the CCA triplet. Base numbering is according to the standard tRNA numbering system (see Section 7.1). Dashed lines mark tertiary interactions based on structural data [26]. **(B)** Canonical L-shaped tertiary structure of cytosolic tRNAs. The L-shape is characterized by the acceptor domain (stacked T-arm and acceptor stem) and the anticodon domain (stacked D- and anticodon arm). Based on tertiary interactions, both domains are connected by the elbow region. The color code is according to **(A)**. Both structures are based on the phenylalanine tRNAs of the reference model *S. cerevisiae*. The figure is modified after Lorenz et al. [27].

anticodon triplet. As the name suggests, the variable loop varies widely from 4 to 24 nts. The T-arm is a 4 to 5 bp stem and a 7 nt long loop containing the sequence TΨC where Ψ is pseudouridine (see Section 3.5), a modified uridine [29]. Together with the 3 nts of two connector domains (between acceptor stem and D-arm, between D- and anticodon arm), this results in short tRNA genes of a variable length of approximately 75 to 100 nts.

Because of tertiary interactions, tRNAs fold into an L-shaped three-dimensional structure [35, 36], see **Fig. 3B**. The universal structure is formed by two helical domains, the acceptor and anticodon domain. The T-arm and acceptor stem shape the acceptor domain while the anticodon domain is composed of the stacked D- and anticodon arm helical presenting the AC-triplet for codon recognition. Both domains interact by long-range tertiary interactions in the elbow region [27, 29]. With few exceptions, this architectural scheme is highly conserved

for tRNAs which tolerate variations in domain length and adaptations in the tertiary interaction networks [29]. The structural features and their flexibility are essential for recognition and interaction with other cellular components. Thus, the tRNA dynamics play an important role for the functionality of a variety of cellular processes, like translation (see Section 1.1) and tRNA maturation (see Section 3) [29, 37].

2.2 Bizarre mt-tRNAs are Ubiquitous

Beside cytosolic tRNAs, eukaryotic tRNA genes are also present in organelles like chloroplasts and mitochondria. Mitochondria encode a minimalist set of mitochondrial tRNAs (mt-tRNAs) to be used in their own mitochondrial translation. For instance, the mammalian mitochondrial genome (mt-genome) encodes for two ribosomal RNAs (rRNAs), 13 proteins which are essential subunits in the oxidative phosphorylation process and 22 mt-tRNAs [38, 39]. The 22 different tRNAs are sufficient to translate all 13 mitochondrial proteins which is possible by the wobble rules [40], see Section 1.1. In higher Metazoa only one mt-tRNA is conserved for each of the 18 amino acid and only leucine and serine occur twice [41]. There is not always a complete set of 22 mt-tRNA genes present in each species. In the contrary, the number varies strongly. In some Metazoa, like the opossum *Didelphis virginiana*, only one mt-tRNA gene is lost, whereas some Protozoa, like *Trypanosoma brucei*, can completely lack mt-tRNA genes. Thus, cytosolic tRNAs have to be imported to the organelles to allow mitochondrial translation of all proteins [42]. However, the loss of tRNA genes is not related to phylogeny and probably occurred during multiple independent events. For example, the genomes of the two fungi *Saccharomyces cerevisiae* and *Spizellomyces punctatu* encode for a set of 22 and 8 mt-tRNAs, respectively [43].

Surprisingly, not all mt-tRNAs share the canonical structural features. Early studies discovered that the D-arm loop is absent in serine mt-tRNAs extracted from bovine hearts [44]. Further analysis revealed other peculiarities in mt-tRNAs. A study of 300 different mammalian mt-genomes showed that mt-tRNAs mostly exhibit classical features, but also some particularities such as a lower GC content and a large variability in D- and T-loop size. This leads to a lack of classical tertiary interactions between both domains and to less post-transcriptionally modified bases (see Section 3.5). D-armless serine mt-tRNAs are also found in every of the

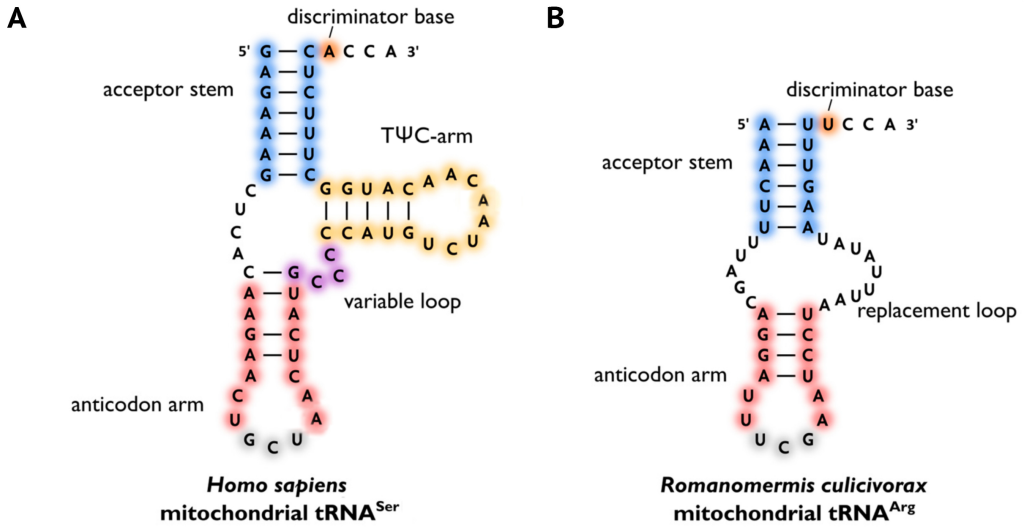

Figure 4: Example of "bizarre" mt-tRNA 2D structures. Examples of mitochondrial tRNAs (mt-tRNAs) that differ from the canonical transfer RNA (tRNA) shape. **(A)** Secondary structure of human serine mt-tRNAs lacking the dihydrouridine arm (D-arm). **(B)** Secondary structure of the arginine mt-tRNA of the nematode *R. culicivorax* which lacks the D-arm and shows a replacement loop instead of the TΨC-arm (T-arm). The acceptor stem (blue), anticodon arm (red), variable loop (purple), T-arm (yellow), and anticodon (grey) are highlighted. The discriminator base (orange) on the 3'-end is followed by the CCA triplet. The figure is modified after Lorenz et al. [27].

300 analyzed mammalian genomes [29, 45, 46]. An example of a mammalian D-armless serine mt-tRNAs is shown in **Fig. 4A**. Even more bizarre are nematodes [47, 48] and arthropods [49] mt-tRNAs lacking the D- or T-arm or both arms (**Fig. 4B**). Furthermore, a systematic analysis of 1,800 metazoan mt-genomes showed that less than 10% of the analyzed mt-tRNAs deviate from the classic canonical tRNA structure. In deuterostomes, including vertebrates, the D-domain is missing in all serine mt-tRNAs and infrequently absent in cysteine mt-tRNAs. Diversity hotspots are found throughout the Ecdysozoa, including Insecta and Nematoda, as tRNAs lacking one or even both arms seem to be the rule [23]. It is still unclear how "bizarre" tRNAs fold in a functional tertiary structure as required for translation. Hypothetically, the stem and arm connected areas gain the flexibility in folding, thus the CCA-end and anticodon are at the necessary distance [29].

2.3 nm-tRNAs are More than Molecular Poltergeists

According to the endosymbiotic theory, eukaryotic mitochondria originated from the progressive transfer of ancient α-proteobacteria DNA into the eukaryotic genome [50]. Thus, mitochondrial DNA (mt-DNA) of higher organisms are 100 to 300 fold smaller than bacterial genomes but still carry hallmarks of its bacterial ancestor [51]. The use of N-formylmethionyl (fMet) tRNA as initiator of protein synthesis is typical for bacteria and the tRNA is also found in mt-genomes [52]. The integration of fragments of mitochondrial DNA in the nuclear genome is an ongoing, frequent process. Thus, mammalian genomes harbor a large number of genomic regions designated as nuclear mitochondrial DNAs (NUMTs). Once integrated, NUMTs evolve largely without selective constraints since mitochondrial gene expression is so different from gene expression in the nucleus that NUMTs cannot be specifically expressed by the nuclear machinery. NUMTs can undergo duplication after genomic integration highlighted by the fact that old NUMTs are more abundant in the human genome than more recently integrated fragments [53]. NUMTs exhibit different degrees of homology to their original mitochondrial fragments as they can be highly rearranged and fragmented [54, 55]. The NUMT content varies dramatically across species, ranging from 400 kilo base pairs (kbp) in rice to 280 kbp in human and over 1 kbp in yeast to the NUMT-less mosquitoes *Anopheles gambiae* [53]. Interspecific variations in terms of germline stability, mitochondria number, frequency of chromosomal integration, and the dynamics of post-insertion processes can be a cause for the species-specific variability [55, 56].

Mitochondrial tRNA genes are also incorporated through the genomic mt-DNA integration, which are referred to as nm-tRNAs. Notably, nm-tRNAs often differ substantially from their mitochondrial counterparts. It has been reported that only eight nmt-RNAs are still identical in sequence to their primordial mt-RNAs in the human genome, while the remaining 489 nm-tRNA genes show up to 25 mismatches [57]. Despite the sequential mutations, the secondary structures are not strong altered by identified mutations [58]. At present, the biological function and relevance of nm-tRNAs is still unknown. Around 20% of known human nm-tRNAs are located in protein-coding or non-coding RNA transcripts [57]. In mitochondria, mt-tRNAs are cleaved out from a single polycistronic transcript [24]. It is speculated that a similar mechanism is used in nm-tRNAs. Nucleases and their interacting factors probably recognize

nm-tRNAs by their structure to cleave them out of the longer transcript. Since nm-tRNAs are presumably transcribed as part of other transcripts, they are not only pseudogenes, but more likely functional [24, 57].

2.4 tRFs are Not Randomly Degraded tRNAs

Beside microRNAs (miRNAs) and small interfering RNA (siRNA), another abundant class of small non-coding RNAs are derived from tRNAs, called tsRNAs. These molecules result from enzymatic cleavage of tRNAs and can be divided into tRNA-derived stress induced RNAs (tiRNAs) and tRNA-derived fragments (tRFs), see **Fig. 5**. Specific cleavage in or close to the anticodon loop of mature tRNAs results in 28–36 nts long 5'-tiRNA and 3'-tiRNA halves. In mammalian cells, angiogenin (ANG), a member of the Ribonuclease A superfamily, is responsible for the tRNA cleavage [59, 60]. Specifically, the tRNA cleavage is executed under stress conditions like heat shock [61], phosphate starvation [62], oxidative damage [63], and under growth conditions [64]. Such tRFs can also originate from pre-tRNAs and mature tRNAs (see Section 3). There are at least four different types of tRFs known which are classified by their site of origin: (i) tRF-1 fragments usually originate from the 3'-trailer sequences of pre-tRNAs by ribonuclease (RNase) Z cleavage and posses poly-U residues at their 3'-end [59, 65]. (ii) 2-tRFs are typically derived from the internal region of mature tRNAs, including the anticodon triplet. The length of 2-tRFs varies between 14–33 nts [66]. The responsible ribonuclease is still unknown. (iii) 3-tRFs are about 18-22 nts long and are cleavage products of Dicer, ANG or other members of the Ribonuclease A superfamily. Due to the T-loop cleavage site of mature tRNAs, 3-tRFs contain the CCA terminus [59, 67]. (iv) 5-tRFs are generated from cleavage in the D-loop of mature tRNAs by Dicer. Depending on their length, 5-tRFs can be further divided into 5a-tRFs (~15 nts), 5b-tRFs (~22 nts), and 5c-tRFs (~30 nts) [59, 65]. However, tRFs are found in different organisms ranging from bacteria to humans, and vary widely between sex, tissue, and disease status [68–72].

Structural features and post-transcriptional modifications of tRNAs could effect the tRF's biogenesis. Especially tiRNAs are cleaved from wrongly modified or folded tRNAs. It has been revealed that oxidative stress conditions disrupt the tRNA folding which in turn favored the tRNA fragmentation as early pathogenic mechanism [73]. In addition, a reduced 5-

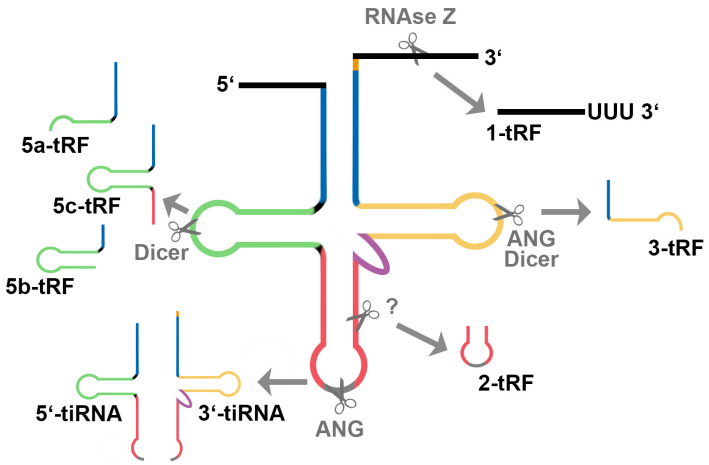

Figure 5: Different structural types of tsRNAs. tRNA-derived small RNAs (tsRNAs) can be classified into the longer tRNA-derived stress induced RNAs (tiRNAs) and tRNA-derived fragments (tRFs). Angiogenin (ANG) cleavage in or near the anticodon loop leaving the 5'-tiRNA and 3'-tiRNA halves. tRFs can be subdivided into tRF-1, 2-tRFs, 3-tRFs and 5-tRFs. tRF-1 fragments originate from the 3'-trailer precursor tRNAs by ribonuclease (RNase) Z cleavage, containing poly-U residues at the 3'-end. 2-tRFs are typically derived from the internal region of transfer RNAs (tRNAs). The responsible ribonuclease is still unknown. 3-tRFs are generated from Dicer or ANG cleavage at the TΨC-arm (T-arm). 5-tRFs can in turn be subdivided into 5a-tRFs, 5b-tRFs, and 5c-tRFs. 5-tRFs are cleavage products of the ribonuclease Dicer and originate from the dihydrouridine arm (D-arm). Except for 1-tRFs, all tsRNAs are fragments of mature tRNAs. The acceptor stem (blue), anticodon arm (red), variable loop (purple), T-arm (yellow), anticodon (grey), and discriminator base (orange) are highlighted.

methylcytidine (m^5C) modification (see Section 3.5) causes aberrant cleavage of tsRNAs into tiRNAs, resulting in a repressed protein translation and an activation of stress pathways [74].

tsRNAs are dynamic regulators of biological processes with diverse functions. For example, tRFs regulate mRNA stability, similar to miRNAs, by directly binding to protein factors of mRNAs [25]. These molecules are also able to inhibit translation initiation and elongation [75], are regulators of ribosome biogenesis [76], alter transcriptional cascades in intergenerational inheritance as paternal epigenetic factor [77], and interact with cytochrome c promoting cell survival [78]. Further, tRFs and tiRNAs are probably causal factors for human diseases, such as cancer [79] and infectious disorders [80]. Thus, the widespread occurrence of tsRNAs and their regulatory significance of biological processes indicate that tsRNAs are not simply randomly degraded tRNA products [59].

CHAPTER 3

From Initial Transcript to Functional tRNAs

Contents

The life of a tRNA molecule starts with its transcription as precursor tRNA (pre-tRNA) followed by a species-specific series of maturation events to fulfill its biological function. The post-transcriptional maturation steps vary in their sequential order from case to case. In general, 5'-leader and 3'-trailer sequences are removed by a set of endo- and exonucleases. Afterwards, a CCA sequence is enzymatically added at the trimmed 3'-end which then represents the site of aminoacylation (see Section 1.1). The CCA addition is necessary only in species where the CCA triplet is not genomically encoded. Introns, which are present in a subset of pre-tRNAs transcribed from intron-containing genes, are removed by tRNA-splicing endonucleases. Additionally, nucleosides are modified in different tRNA processing steps, altering nucleotide properties in different ways. However, tRNA biogenesis occurs at several distinct subcellular locations which are specific for each processing step [81].

In the following chapter, the individual processing steps of tRNA biogenesis are described in more detail with focus on tRNA modifications. An overview of all tRNA processing steps including their particularities and subcellular locations is illustrated in **Fig. 6**.

3.1 tRNA Transcription

Numerous mammalian organisms have over 400 tRNAs which makes it metabolically meaningful to coordinate their transcription [12]. The three-dimensional structure of a tRNA enables their clustering within the nucleolus, although their genomic loci are dispersed in the genome [82]. Transcription of tRNA genes requires a type 2 internal promotor consisting of an A- and B-box (see **Fig. 6**). These intragenic boxes encode parts of the dihydrouridine- and TΨC-stems and loops respectively [11], see Section 2.1. The promotor is modulated by upstream sequence motifs frequently including a TATA element [83]. Initiation of transcription is caused by a concerted action of transcription factors binding to the tRNA genes which recruit the RNA polymerase III (Pol III). At first, the transcription factor for polymerase III C (TFIIIC) binds the promotor, followed by recruitment of the transcription factor for polymerase III B (TFIIIB) to a ∼50 bp upstream region of the transcription start site. The binding of TFIIIB recruits Pol III and promotes transcription initiation. A key component of TFIIIB is the TATA-box binding protein (TBP) that interacts with upstream DNA. TFIIB-related factor 1 (BRF1) and TFIIB double prime 1 (BDP1) form the last two subunits of TFIIIB [83, 84].

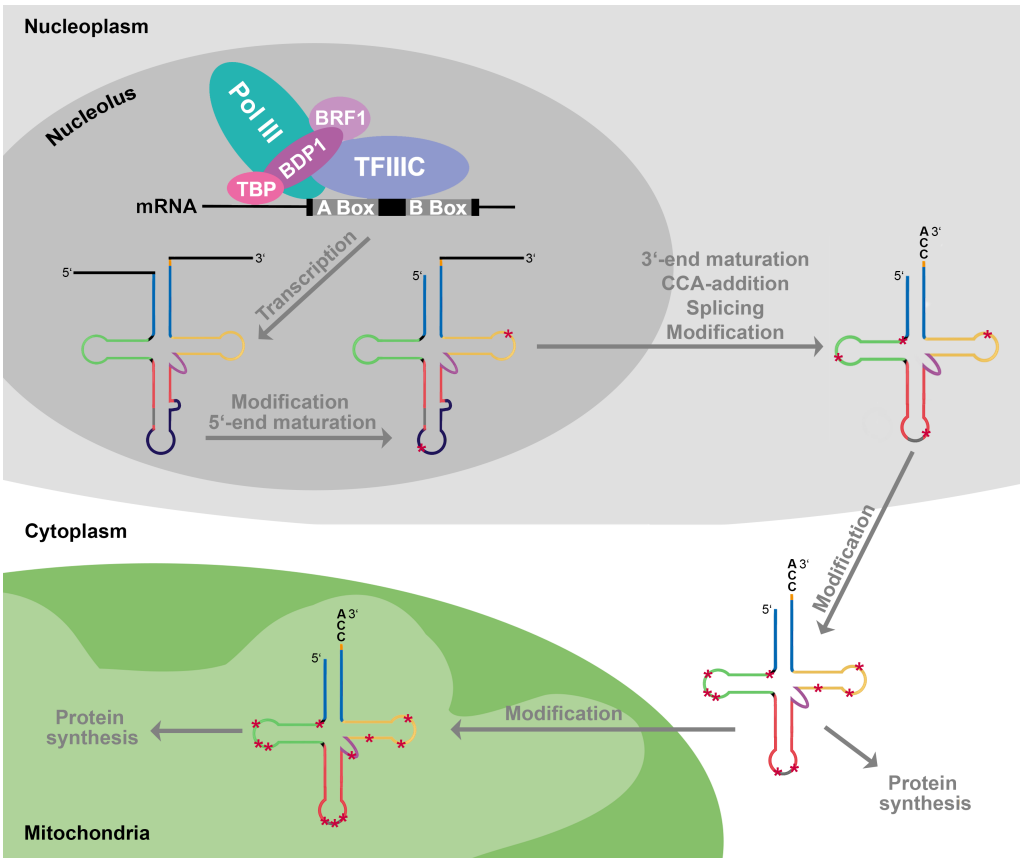

Figure 6: Cell biology of eukaryotic tRNA biosynthesis. Transcription and maturation of transfer RNAs (tRNAs) occur at several distinct subcellular locations, whereas tRNAs are dynamically in- and exported between nucleus, cytoplasm and mitochondria. Transcription of precursor tRNA (pre-tRNA) genes caused by a concerted action of binding of transcription factors. Transcription factor for polymerase III C (TFIIIC) binds the A- and B-box promotor which encode parts of the dihydrouridine- and TΨC-arm, respectively. The binding recruits the transcription factor polymerase III B (TFIIIB) which interact with the ∼50 bp upstream region of the transcription start site. TFIIIB consists of the three subunits TFIIB-related factor 1 (BRF1), the TFIIB double prime 1 (BDP1) and the TATA-box binding protein (TBP). The binding of TFIIIB recruits RNA polymerase III (Pol III) for transcription initiation. Pre-tRNA transcription, as well as the following 5'-leader sequence removal and first modifications of a few nucleotides take place in the nucleolus. After pre-tRNA export to the nucleoplasm, the 3'-trailer sequence is removed and replaced by the CCA termini. Existing introns are then spliced and nucleotides modified as required. Pre-tRNAs are then exported to the cytoplasm of the cell where additional base modifications may originate. Now the pre-tRNA is maturated and can fulfill its biological function. Beside the dynamic pre-tRNA trafficking between nucleus and cytoplasm, a small fraction of cytosolic tRNAs can be imported into the mitochondria to compensated incomplete or redundant mitochondrial tRNA sets.

The 5'-flanking regions of animal tRNA genes are highly divers and lack a common conserved sequence motif [85]. Thus, a strong and regulated Pol III binding is required [86]. Strikingly, tRNA species which encode for the same anticodon share conserved motifs in their upstream regions that might reflect coordinated regulation [11, 85]. Pol III interactions are based on differential expression of tRNA isoacceptors also between tissues within an organism which can vary up to tenfold [11, 87].

The strong interaction of the transcription factor TFIIIB with the 5'-regions of single subgroups may modulate the tRNA tissue-specific expression, since various animal tissues differentially express multiple isoforms of BDP1 and BRF1. These multiple forms explain the lack of a uniform signature motif in the 5'-upstream regions of animal tRNA genes. In contrast, a highly conserved TATA motif followed by a CAA motif in the tRNA upstream regions was found in plant genomes [85].

Mitochondrial DNA (mt-DNA) is transcribed into single polycystronic RNA, where mitochondrial tRNAs (mt-tRNAs, see Section 2.2) are processed by punctuated endonucleolytic cleavage [88] by mitochondrial RNA polymerase (mtRPOL) [89]. It is still unknown how nuclear-encoded mitochondrial-derived tRNA (nm-tRNAs, see Section 2.3) are transcribed. Due to the identification of a sequence motif which strongly resembles the consensus sequence of B-boxes, the transcription mechanism of nm-tRNAs is speculatively similar to that of nuclear tRNAs [24].

3.2 5'- and 3'-End Maturation

At both 5'- and 3'-ends, pre-tRNAs contain terminal extensions, the 5'-leader and 3'-trailer sequence, respectively. Across all kingdoms, the 5'-leader is removed by RNase P, producing a monophosphate at the 5'-end and a terminal 2'-3'-cis glycol [90, 91]. Only few exceptions to the requirement for RNase P activity are known for the archaea *Nanoarchaeum equitans*, *Pyrobaculum aerophilum* and for the bacterium *Aquifex aeolicus* [92]. Randau et al. [93] provide experimental evidence that the accurate promotor replacement in *N. equitans* ensures that the tRNA transcription starts at the first nucleotide of the mature tRNA, producing leaderless tRNA transcripts.

While tRNA 5'-processing is almost ubiquitous in all species, tRNA 3'-end cleavage varies in the three domains of live. The temporal order of the cleavage reactions of the terminal tRNA extensions has not yet been investigated. Bacteria use a multi-step pathway for the 3'-trailer cleavage (see **Fig. 7A**). At first, an endonucleolytic cutting process starts at the tRNA 3'-end and extends downstream [94]. Responsible for the endonucleolytic reaction is a combination of the two enzymes RNase E and RNase III. RNase III has been found to recognize double-stranded RNA as substrates, while RNase E cleaves single-stranded AU-rich sequences [95, 96]. In a second step, the 3'-end is further shortened by an exonucleolytic reaction before RNase P processes the 5'-end. Finally, a second exonucleolytic trimming degrades the remaining nucleotides [94, 97]. Depending on the 3'-trailer requirements, the exonucleolytic cleavage can be catalyzed by a variety of six known exonucleases, namely the RNases II, BN, D, PH, T, and polynucleotide phosphorylase (PNPase) [97, 98]. CCA terminus is encoded in most bacterial tRNA 3'-trailers, followed by a series of appended nucleotides. RNase D specifically cuts right up to the CCA terminus without disrupting it, resulting in a mature tRNA 3'-end without any further processing steps [99]. RNase BN is specific for trimming incomplete 3'-trailers or trailers with missing CCA tail [100, 101]. Since the described processes are based on *Escherichia coli* data, which is currently the only well-studied bacterial organism in this context, the 5'- and 3'-end maturation mechanism may differ in other bacteria. In contrast, 3'-trailers in eukaryal as well as archaeal organism are primarily removed by tRNA 3'-endonuclease (RNase Z), see **Fig. 7B**. The cleavage results in a tRNA ending with a 3'-hydroxy group of the discriminator (see Section 2.1) nucleotide [102–104].

3.3 tRNA 3'-Terminal CCA Addition

In species that do not genomically encode the terminal CCA sequence (Eucarya, Archaea, and some Bacteria), the CCA triplet is added post-transcriptional to the cleaved tRNA 3'-end by tRNA nucleotidyl transferase catalysis [105]. Beside that, the tRNA nucleotidyltransferase restores CCA termini which were degraded by exonucleolytic activities [106, 107]. If the tRNA nucleotidyl transferases are located in the nucleoplasm, they take part in the 3'-end processing. If the enzymes are located in the mitochondria or cytoplasm, they take part in the 3'-CCA tail repair [108]. Terminal CCA addition is essential for the functionality of the mature

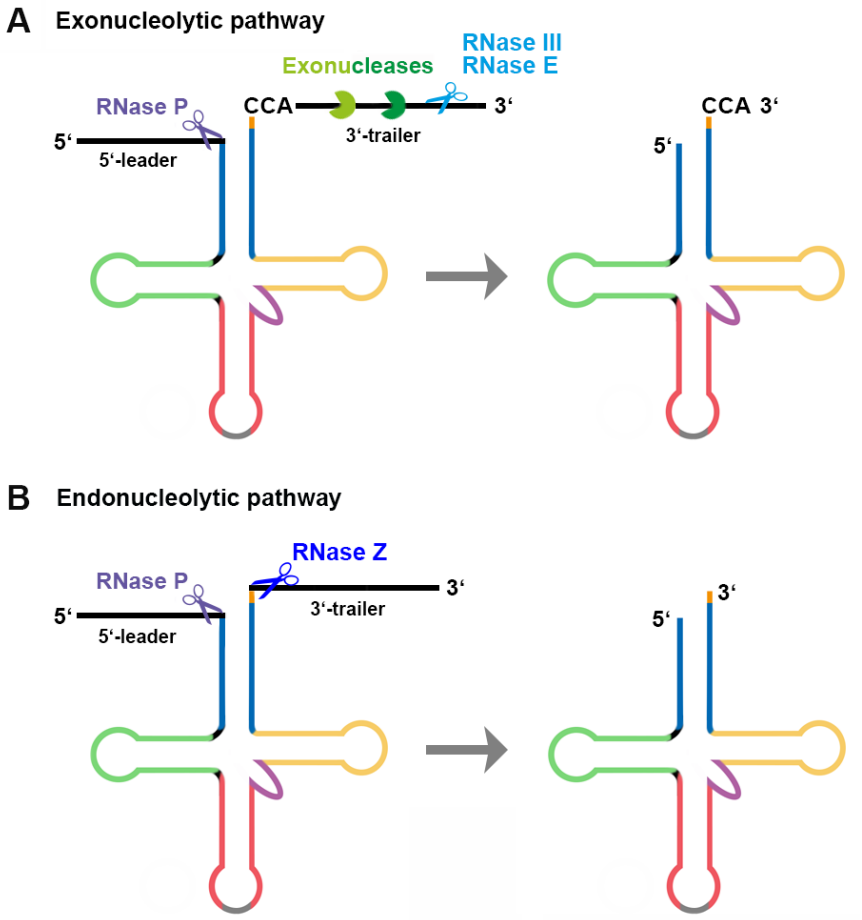

Figure 7: 5'- and 3'-end tRNA maturation pathways. Two different pathways exist for 5'-leader and 3'-trailer processing. **(A)** In the exonucleolytic pathway, the 5'-leader sequence is cleaved off by ribonuclease (RNase) P (purple). The 3'-end maturation consists of multiple steps starting with an endonucleolytic cut (light blue) at the 3'-end of the transfer RNA (tRNA) followed by exonucleolytic trimming (light and dark green) reactions. The endonucleolytic cut based on the combined activity of the enzymes RNase E and RNase III, which are specific for single-stranded AU-rich and double-stranded RNAs, respectively. Depending on the 3'-trailer requirements, the exonucleolytic cleavage can be catalyzed by a variety of different enzymes. This pathway occurs in bacterial species where the CCA terminus is encoded in 3'-trailers of tRNAs. **(B)** Within the endonucleolytic pathway, the 3'-trailers of the primary transcripts are removed by the tRNA 3'-endonuclease (RNase Z; dark blue). The 5'-end maturation is similar to the exonucleolytic pathway. The pathway is conserved among eukaryotes and archaea. The structural features of the tRNAs are color-coded as follow: acceptor stem (blue), anticodon (grey), anticodon arm (red), dihydrouridine-arm (D-arm; green), discriminator base (orange), TΨC-arm (T-arm; yellow), and variable loop (purple).

tRNA during protein biosynthesis. The amino acid loaded onto the tRNA by aminoacyl tRNA synthetases, to form aminoacyl-tRNA, is covalently bonded to the 3'-hydroxyl group on the CCA tail [109], see Section 1.1.

3.4 Pre-tRNA Splicing

A minority of pre-tRNAs are encoded by intron-containing genes. Introns are spliced to get functional mature tRNAs [81]. In tRNAs of Eukarya and Archaea, introns are mainly located 3' to the anticodon and do not interrupt the overall tRNA structure [110]. Unusual types of tRNA introns are found at 14 positions within the tRNA genes [111]. Introns interrupting the dihydrouridine (D)-stem (see Section 2.1), for example, were found particularly in archaeal genomes. Except for tyrosine tRNAs, which generally contain introns, the phylogenic distribution of introns within particular tRNA genes is not conserved [112]. It could be shown that tRNA introns generate substrates for particular modification enzymes, e.g., yeast tyrosine tRNAs produce a pseudouridine (see Section 3.5) in the anticodon loop [113].

In all organisms, tRNA-splicing taken place in different enzymatic steps, starting with the removal of the intron itself. This step is catalyzed by the tRNA-splicing endonuclease, leaving a 5'-tRNA half-molecule ending in a 2'-3'-cyclic phosphate, and a 3'-tRNA half-molecule beginning with a 5'-hydroxyl group [114]. The bulge-helix-bulge (BHB) motif is essential for splice cite recognition by the tRNA-splicing endonuclease in archaea [111, 115], while eukaryotic introns do not have clear splicing motifs. Eukaryal enzymes use a measuring mechanism to determine the positions of the splice sites relative to the conserved pre-tRNA domain (for details see H. Li et al. [116]). Archaea and most eukaryotes join the tRNA halve-molecules directly by the 3'-5'-activity of the tRNA ligase [117]. Yeast and plants require two steps, starting with the 5'-3'-ligation which leaves a 2'-phosphate group at the spliced junction [114]. Finally, a 2'-phosphotransferase catalyzes the binding of the 2'-phosphate group to its cofactor, nicotinamide adenine dinucleotide (NAD). However, in bacteria, tRNA introns are very rare self-splicing group I introns which means they accurately and efficiently excise themselves from the pre-RNA and ligate the flanking exon sequences in the complete absence of splicing enzymes [118].

3.5 Nucleotide Modifications

The highest level of post-transcriptional modifications is found in tRNAs thinking of their small sizes and low chemical diversity they show. So far, 93 different tRNA modifications have been identified [119], where mitochondrial tRNA (mt-tRNA) modifications are mostly of bacterial origin [120]. During tRNA processing, both nuclear and mitochondrial tRNAs are modified by specific tRNA-modifying enzymes. The nature of nucleoside modifications in tRNAs varies and can be classified into light or complex alterations. If single tRNA-modifying enzymes methylate the base or ribose ring of tRNA molecules, these alterations are referred to as simple modifications. Simple modifications occur most frequently and can be mainly found at the elbow region (see Section 2.1) of the L-shaped tRNA [27]. In contrast, complex altered bases, termed hypermodifications, are characterized by radical structural changes and can involve multiple enzymatic steps [121]. Hypermodifications are mainly located in the anticodon loop [27]. A collection of almost all annotated tRNA modification including their symbols and common names are listed in **Suppl. Tab. B1**.

3.5.1 tRNA Modifications are Highly Distributed Across Kingdoms

Generally, the frequency and composition of tRNA modifications vary across the three domains of life or even between isoaccepting tRNAs (see Section 1.1) within an organism [122]. In comparison to archaeal and bacterial species, the highest frequency of modified tRNA molecules is found in eukaryotes. With up to 23.7% modified residues, green plants show the highest level of post-transcriptional tRNA modifications, whereas lower modification rates are found in single-celled eukaryotes ($\sim$16.6%) [27]. As an example, 25 different modifications at 36 positions are found in *S. cerevisiae*, with an average of 12.6% modifications per tRNA species [123]. The lowest modification rate is found in Gram-positive bacteria ($\sim$6.6%) [27].

A set of 17 different kinds of modifications ($\sim$18% of all known tRNA modifications) are spread across all domains of life, representing universal core modifications (see **Fig. 8**). These include the well studied modifications 1-methyladenosine (m^1A), inosine (I), D, and pseudouridine (Ψ) [119, 122]. Despite core modifications, $\sim$62% of all known tRNA modifications are specific to one domain of life, whereas the remaining 20% overlap between two domains. Hypermodifications could be isoacceptor specific and are mostly domain specific [27].

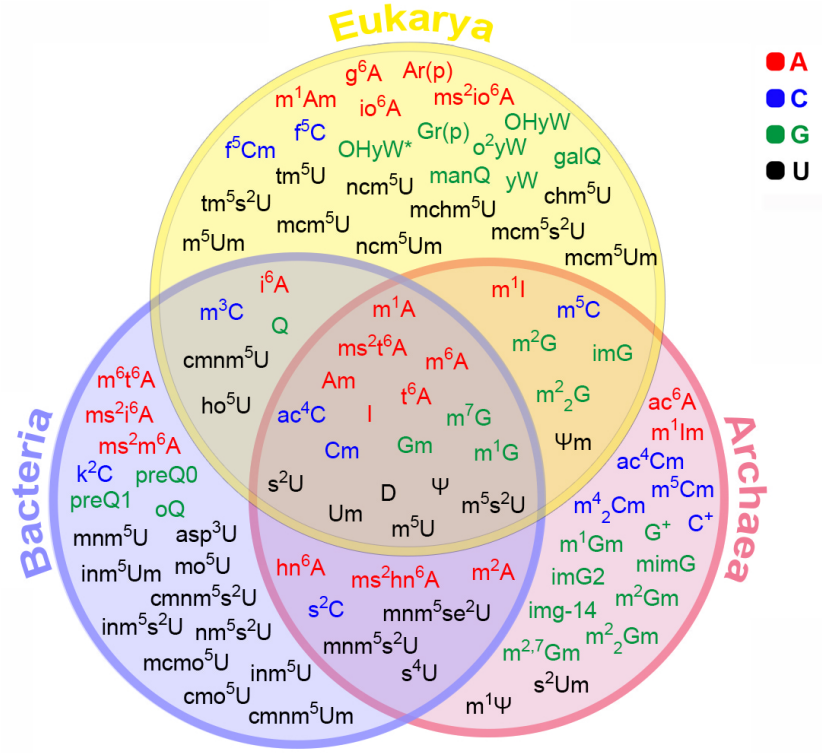

Figure 8: Distribution of tRNA modification across the domains of life. A total of 93 known transfer RNA (tRNA) modifications are assigned to the three domains of life (Eukarya, Bacteria, Archaea). The data were collected from the RNA modification database [119]. Some tRNA modifications overlap between all domains (18%), two domains (20%) or are domain-specific (62%). Modifications are color-coded referring to the nucleotide they modify: adenine (red), cytosine (blue), guanine (green), uracil (black). The conventional abbreviations are used for the modifications: ac – acetyl, acp – aminocarboxypropyl, chm – carboxyhydroxymethyl, cmo – oxyacetic acid, cmnm – carboxymethylaminomethyl, f – formyl, g – glycinyl, gal – galactosyl, hn – hydroxynorvalylcarbamoyl, ho – hydroxy, i – isopentenyl, inm – isopentenylaminomethyl, io – cis-hydroxyisopentenyl, m – methyl, man – mannosyl, mchm – carboxyhydroxymethyl methyl ester, mcm – methoxycarbonylmethyl, mcmo – oxyacetic acid methyl ester, mnm – methylaminomethyl, mo – methoxy, ncm – carbamoylmethyl, nm – aminomethyl, r(p) – 5-O-phosphono-b-d-ribofuranosyl, s – thio, se – seleno, t – threonylcarbamoyl, tm – taurinomethyl.

3.5.2 Nucleotide Modifications are Functionally Diverse

Mainly, tRNA modifications act as checkpoints for tRNA integrity, regulate the protein translation, modulate the structural stability, and ensure the correct folding of the linear tRNA molecule. They are further involved in the structural fine-tuning of local elements and are able

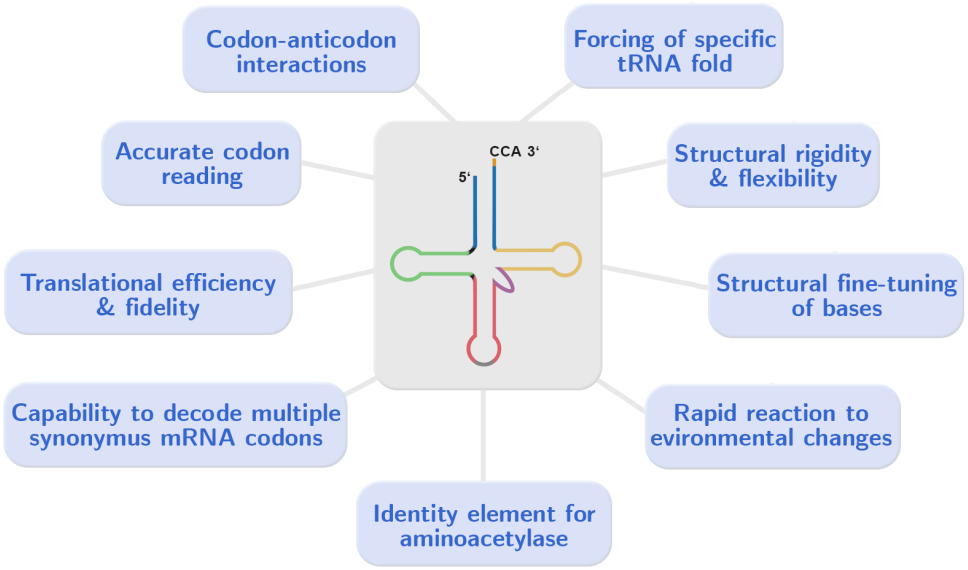

Figure 9: Primary functions of tRNA modifications. An overview of the main functions of tRNA modifications is illustrated. In summary, transfer tRNA (tRNA) modifications are essential for structural adaptations, for folding of the linear tRNA molecule, and are involved in the regulation of protein biosynthesis.

to rapidly adapt tRNAs to environmental changes, such as stress and temperature [27, 124]. An overview of the most important functions of tRNA modifications is depicted in **Fig. 9**. The absence of altered bases can cause dramatically pathological consequences and severe diseases, ranging from respiratory and metabolic defects over myopathies, mitochondrial disorders and to X-linked intellectual disability (reviewed in Duechler et al. [124]). Post-transcriptional modifications of tRNAs are crucial to maintain their functionality in the cell.

tRNA Modifications are Crucial for tRNA Structure and Folding

In general, the nucleotides adenine (A), cytosine (C), guanine (G) and uracil (U) allow for the formation of different base pairs, which can lead to deviations of the typical cloverleaf or L-shaped formation of tRNAs (see Section 2.1). Such alternative structures are often non-functional and show a similar level of free energy. They can be triggered, e.g., by a pronounced nucleotide bias as found in GC-rich thermophilic organisms and some AT-rich mitochondrial

genomes [125]. Specific tRNA modifications force the tRNA to adopt a certain functional structure. As an example, in GC-rich thermophilic archaea, the N^2,N^2-dimethylguanosine ($m^2{}_2G$) modification suppresses the G-C pairing by removal of the hydrogen-bond donor in the Watson/Crick edge, forcing the modified G residues to built a G-U wobble pair and additional non-Watson/Crick base pairings [126]. Similarly, unmodified human lysine mt-tRNAs (see Section 2.2) fold a non-functional rod-like tRNA structure. The presence of m^1A at position 9 prohibits a base pairing between A9 and U64, which disrupts the non-functional structure and forces the tRNA to adopt the cloverleaf formation [127].

Besides these massive structural rearrangements, tRNA modifications also act as identity elements [128]. In the nematode *Ascaris suum*, 90% of the mt-tRNAs lack the entire T-arm carrying the m^1A modification at position 9. In particular, the methylation within the bizarre methionine and phenylalanine mt-tRNAs (see Section 2.2) leads to a different folding pattern in the D-arm and the loop region that replaces the T-arm. Here, the distance between CCA-end and anticodon is affected which in turn binds efficient to the corresponding aminoacyl tRNA synthetase or interacts with the mitochondrial elongation factor EF-Tu [129].

Modifications, mainly located in the D- and T-loop, are crucial for the stability and flexibility of the tRNA structure [130]. Especially pseudouridine stabilizes the overall tRNA structure and increases its rigidity by contributing water-mediated bridging interactions between modified bases and the RNA backbone [131]. On the other hand, dihydrouridine, which is frequently found in the D-arm, gives tRNAs local and functionally important flexibility. Here, a saturation of the C5-C6 bond in the pyrimidine ring of uridine promotes the C2'-endo conformation of the ribose. This sugar pucker is also relayed to the 5'-adjacent ribose, increasing the local structural rearrangement [27, 132].

In thermophilic and psychrophilic organisms the structural adjustment of macromolecules is interpreted as a frequent strategy for thermal adaption. Many of these strategies are not biomolecule-specific and can be found in tRNA as well [27, 133]. For example, Ψ increases thermostability in order to protect RNAs from denaturation and degradation [134]. Nuclear magnetic resonance spectroscopy analysis demonstrated, that the structural rigidity of tRNAs introduced by Ψ39 modifications result in a up to 5 °C higher melting temperature [135]. Additionally, Ψ at position 55 in the T-loop is contributing to thermal stress tolerance in *E. coli*. Another thermostabilizing modification is 5-methyluridine (m^5U), which is often further

modified to 2-thiouridine (s_2U). These modifications promote C3'-endo sugar conformation and tertiary interaction with A58 and by this increase the melting temperature of a tRNA by 3 °C [136, 137]. It could be demonstrated, that the equilibrium between C2'- and C3'-endo conformation of dihydrouridine and the upstream located base is strongly temperature depended. These structural flexibility introduced by dihydrouridine allows the shift to the C2'-endo sugar conformation as a strategy for cold adaptation [27, 138].

tRNA Modifications Regulate Protein Synthesis

Modifications involved in the regulation of protein biosynthesis (see Section 1.1) are primarily located at or near the anticodon loop (see Section 2.1) of tRNAs. Wobble base position 34 and the anticodon adjacent position 37 are prominent regulatory sites. Both sites show a high abundance of hypermodifications to regulate the correct and efficient codon-anticodon base pairing [139, 140]. Modifications at position 34 are necessary for tRNA integrity. They increase the capability of tRNAs to decode multiple synonymous mRNA codons according to the wobble rules [141]. Meier et al. [142] have reported that histidine $tRNA^{His}_{G34}$ clearly prefers the codon CAC to the codon CAU, whereas a histidine tRNA with a hypermodified nucleoside queuosine (Q) in the wobble position has slight preference for the codon CAU. Additional, modified residues at the tRNA position 34 improve accurate codon reading. For instance, isoleucine tRNAs with the anticodon UAU or CAU can bind the mRNA corresponding codons AUA and AUG, which encode for isoleucine and methionine, respectively. Usually, isoleucine tRNAs have a strong preference for their cognate AUA codon. An increased misreading of methionine may occur if the isoleucine tRNA is carrying the anticodon CAU. To avoid the misreading, C34 is modified to lysidine (k^2C) restricting the codon recognition to only AUA. Thus, the amino acid identity is changed from methionine to isoleucine (reviewed in [27]).

The coordination of the concentration of a particularly modified tRNA with the frequency of its cognate codon in the mRNA introduces an additional level of regulation to fine-tune translation and influences the translational efficiency for particular proteins. An example is the 5-methoxycarbonylmethyl-2-thiouridine (mcm^5s^2U) modification that occurs in the wobble position of arginine and glutamine tRNAs. The more mcm^5s^2U modifications occur within the cell, the stronger is the synthesis activity of DNA damage response proteins [124, 143].

Post-transcriptional modifications at base 37 help to stabilize codon-anticodon interactions that contribute to reading frame maintenance [144]. They stabilize tRNA-mRNA interactions by improving intrastrand stacking within the anticodon loop and interstrand stacking between codon and anticodon base [145]. In detail, the canonical anticodon is a purine-rich loop that consists of seven unpaired nucleotides refereed to as U-turn. This unpaired formation is necessary for stable codon-anticodon interaction with the ribosomal A residue and therefore for increased translation efficiency and fidelity. Since the purine-rich sequence is predicted to have strong favorable base stacking energy, tRNA modifications are necessary to preserve the U-turn [27, 146]. For instance, the 1-methylguanosine (m^1G) modification, which contains around 75% of bacterial tRNAs and is present in over 95% in all known sequences of proline tRNAs [147]. The absence of this modification results in a deformation on the opposite side of the anticodon loop at nucleotide U32, leading to the disruption of interactions with A38. The lack of U32-A38 inhibits the direct ribosome contact, while the destabilized stem structures are not able to be recognized by the elongation factor EF-G during translocation [146]. Consequently, the ribosome decodes four rather than three nucleotides, resulting in a $+1$ frameshifting. Thus, the m^1G37 modification helps proline tRNAs to stabilize the tRNAs anticodon to prevent $+1$ ribosomal frameshift errors [148]. Similar observations were made for bacterial phenylalanine tRNAs. The unmodified anticodon stem-loop from phenylalanine in *E. coli* forms a trinucleotide loop in solution, due to the base pairing of U32-A38 and U33-A37. The N^6-isopentenyladenosine (i^6A) modification at position 37, which is further modified to 2-methylthio-N^6-isopentenyladenosine (ms^2i^6A), disrupts the non-canonical loop conformation [145, 149].

Some tRNA nucleoside modifications are changed dynamically by environmental factors such as stress or nutrition [124]. The modification rearrangements prime the protein synthesis capacity for the currently most needed proteins. In mammals, oxidative stress increased the 5-methoxycarbonylmethyl-2'-O-methyluridine (mcm^5Um) modification to promote the expression of selenocysteine containing proteins. Such proteins contribute to detoxification of reactive oxygen species [150].

3.5.3 Base Modifications are Not Restricted to tRNAs

Complex and diverse nucleoside modifications are not restricted to tRNAs. Chemical base modifications in RNAs are a wide-spread phenomenon that affects all four nucleotides at different positions and occurs in all domains of life. Target-specific methylation enzymes extend to all classes of both coding and non-coding RNAs [151, 152]. The most prevalent methylations in eukaryotic mRNAs are m^1A [153], N^6-methyladenosine (m^6A) [154, 155], 5-methylcytidine (m^5C) [156–158], and 5-hydroxymethylcytidine (hm^5C) [159]. In mammalian mRNAs, m^1A is low abundant, with a frequency of 0.015-0.16% of all adenosines, and was found around start codons to enhances translational efficiency [153]. The m^6A methylations are highly abundant and enriched near stop codons and in 3'- untranslated regionss (UTRs). They support tissue-specific regulation and markedly increases throughout brain development [160]. Several other functions have been described for the reversible m^6A modification. For example, they increases translational efficiency [160], influences the circadian rhythm [161], controls translation rates in heat shock response [162], and decreases the codon-anticodon binding during translation [163]. The m^5C methylation is primary found in untranslated regions of mRNAs [164]. Since its occurrence in the 3'-UTR correlates with Argonout binding proteins, m^5C is likely involved in miRNA targeting [124]. Further, m^5C can be oxidized to hm^5C which can enhance mRNA translation [159].

Base methylations are also present in other RNAs. A significant fraction of precursor microRNAs (miRNAs) contain m^6A modifications that control the processing of their steady-state level [165]. Methylation m^5C is also present in small amounts in non-coding RNAs. They are crucial for the processing of vault RNAs (vtRNAs) into microRNA-like small RNAs [74].

A high density of Ψ is found in mRNAs [166, 167]. Site-specific pseudouridylation of eukaryotic mRNAs naturally occurs in multiple transcripts providing important regulatory functions [168]. Pseudouridine modifications in human and yeast mRNAs allow a rapid response to environmental changes [169]. Mostly, they are highly conserved and found in mRNAs harboring other evolutionary conserved, ancient RNA modifications. Thus, pseudouridine can be used as an epitranscriptomic marker [166]. In non-coding RNAs, Ψ is one of the most widespread modifications and is highly conserved among species. These modified residues are often located in functionally important regions of the major spliceosomal small nuclear

RNAs (snRNAs) that participate in the intermolecular RNA–RNA or RNA–protein interactions involved in the function of the spliceosome. Ψ stabilize also the structure of rRNA [131].

A-to-I editing targets several thousands of mRNAs [170]. Here, adenosine deaminase acting on RNA (ADAR) enzymes converts adenosine in double-stranded regions of transcripts into inosine [171]. A-to-I editing leads to a unique pattern of amino acid changes characterized by enriched stop-to-tryptophan, positive-to-neutral, and neutral-to-positive charge changes which strongly influence the protein function [172]. In miRNAs, A-to-I editing leads to re-targeting of the mature miRNA [173].

3.6 The Long Cellular Way of tRNA Biogenesis

Transcription and maturation of tRNAs take place at distinct subcellular locations necessitating the dynamic in- and exported of tRNAs between nucleolus, nucleoplasm, inner nuclear membrane (INM), cytoplasm, and cytoplasmic surface of mitochondria [81], see **Fig. 6**. The subcellular tRNA traffic differs from the much more localized processing events of other RNAs. In case of mRNA, most of the processing factors are recruited to the transcription site or are co-transcriptionally functional [174].

Transcription of tRNA genes is located in the nucleolus. The clustering of tRNAs within the nucleolus is based on the chromosome-condensing complex condensin. Haeusler et al. [175] has demonstrated that tRNAs are physically associated wit condensin and cells with conditionally defective condensin subunits fail to cluster tRNA genes in the nucleolus. Since the endonuclease RNase P is localized in the nucleolus, the 5'-end processing of tRNAs takes place in the nucleolus as well [176]. Unlike 5'-end maturation, the 3'-processing occurs in the nucleoplasm. Before the tRNAs are exported to cytoplasm the CCA tail is added [81].

The intracellular location of splicing enzymes differs among species, despite their conservation in eukaryotes. Vertebrates splicing enzymes are mainly concentrated in the nucleoplasm [177, 178]. In contrast, pre-tRNA splicing in yeast and speculatively also in plants proceed in the cytoplasma of cells. In yeast, tRNA splicing endonuclease was found to associate with the mitochondrial surface [179]. In plants, tRNA splicing is not entirely understood yet. It is hypothesized that eukaryotic cells can tolerate drastic changes in the site of tRNA splicing [180].

Subcellular organization of nuclear modification enzymes is followed by an ordered process for tRNA modification. Both nuclear-encoded and mitochondrial-encoded tRNAs are modified in the nucleus, cytoplasm, and mitochondria. Modification enzymes in the nucleus have distinct subnuclear distributions, and can be located in the nucleolus, in the nucleoplasm, or at the INM. Enzymes which are responsible for m^1A58 methylation in yeast occurring on some tRNA initial transcripts, are located in the nucleus [181]. Intron containing tRNAs are also modified in the nucleus [182] or mitochondria [183]. Intronless tRNAs are catalyzed by enzymes that reside in the nucleus and/or cytoplasm [182]. Modification enzymes that are responsible for modifications in the anticodon loop are located in the cytoplasm [184]. Beside the dynamic tRNA trafficking between nucleus and cytoplasm, a small fraction of cytosolic tRNAs can be imported into mitochondria. In some species, this allows the compensation of an incomplete or redundant set of mt-tRNAs [185].

CHAPTER 4

Evolutionary Events of tRNA Genes

Contents

The origin of tRNAs developed before the separation of the three domains of life. Furthermore, there is clear evidence that all tRNA genes are homologs, i.e., derived from an ancestral "proto-tRNA" [186] which in turn may have emerged from even smaller components [187]. These are indispensable in all organisms. Since tRNA genes occur in multiple copies throughout the genome they belong to repetitive elements. Beyond *bona fide* tRNAs, there is a rich universe of tRNA-derived repetitive short interspersed nuclear elements (SINEs) [188] and small RNAs that either directly derive from tRNAs [189, 190] or arose indirectly as exapted SINEs [191]. Thus, tRNAs and many other classes of small RNA genes, e.g., small nucleolar RNAs (snoRNAs) [192], behave like mobile genetic elements. As a consequence, the tRNA repertoire can change rapidly even between closely related genomes [193, 194]. In particular, duplications of tRNAs, known as gain events, lead to paralogous genes. Gene duplication can occur in two ways. First, tRNA sequences build so-called tandem duplications by being copied once or more times into the genome region around the original tRNA. Such tandem duplications lead to the formation of tRNA clusters. Secondly, the tRNA copy is inserted between other tRNA genes further away from its original location. In the case of tRNAs this typically leads to pseudogenization resulting in a non-functional tRNA. Therefore, a rapid net turnover of tRNA genes at individual loci can be observed sometimes [194–197]. Further mutations in pseudogenes do not cause any detectable similarity with any other tRNA sequence. Thus, the mutated pseudogene cannot be found in the genomic background any more which is referred to as loss events. Turnover can be estimated quite accurately by simply comparing gene copy numbers between species when gain and loss events are rare as in the case of miRNAs [198]. However, the fraction of conserved tRNA loci quickly decreases with phylogenetic distance, so that similar tRNA numbers among different mammalian families are the consequence of compensation between large numbers of gain and loss events [194]. An overview of evolutionary sequence homology phenomena is depicted in **Fig. 10**.

In addition to gain and loss of entire tRNA genes, mutations in the anticodon loops may change the identity of the tRNA. This process is known as *tRNA remolding* [199]. The modified anticodon usually corresponds to the same amino acid (isoacceptor remolding), however, in mitochondrial genomes also alloacceptor remoldings, i.e., a change in the addressed amino acid, is observed with surprising frequency [200, 201]. In contrast, the nuclear tRNAs of eukaryotes are largely restricted to isoacceptor remoldings [196, 202], presumably because proper loading

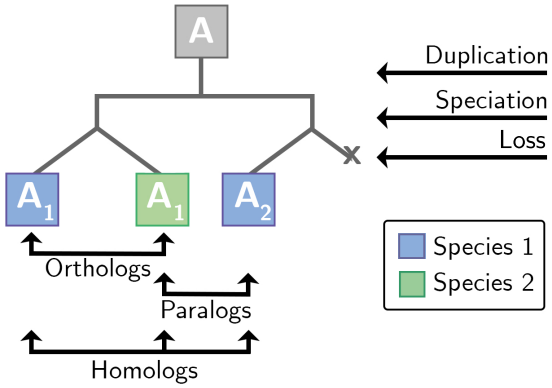

Figure 10: Overview of evolutionary events. Ancestral gene duplication of gene A resulting in two paralogous genes A_1 and A_2. A speciation event produces the orthologous genes A_1 in species 1 and 2, while the orthologous gene A_2 of species 2 to species 1 has been lost. All genes A1 and A2 are homologs since they originate from the same ancestral gene.

of a tRNA depends on a complex system of identifying elements that may even be disjoint from the anticodon sequence [128]. Surprisingly, even isoacceptor remoldings are rare in Archaea and Eubacteria [202]. Like the estimation of tRNA gain and loss, a quantitative investigation of tRNA remolding events also hinges on the correct identification of orthology. Homologous tRNAs are orthologous if they are inferred to be descended from the same ancestral tRNA gene separated by a speciation event: when a species diverges into two separate species, the copies of a single tRNA gene in the two resulting species are said to be orthologs [203].

4.1 tRNA Genes Undergo Concerted Evolution

A larger sequence similarity of members of a repetitive family occur within a species than between species. This suggests that members of a repetitive family do not evolve independent of each other. This effect leads to homogenization of repetitive elements which is known as concerted evolution [204]. In case of tRNAs, paralogous genes with the same anti-codon maintain up to nearly identical sequences over long evolutionary time-scales. Duplicated genes will not remain subject to concerted evolution forever, but will escape with a roughly exponentially distributed waiting time and will start to accumulate mutations [205]. High sequence similarity between paralogs may be maintained by homologous recombination events

which leads to intergenic conversion. It has been shown already in the 1980s that intergenic conversion is an important contributing factor [206]. Ectopic gene conversion involves the unidirectional copy of genetic material from a "donor" sequence to a homologous "acceptor" [205]. Due to the extremely low rates of sequence evolution in tRNAs, gene conversion events are frequent enough for the information transfer to be effectively bidirectional. Hence, the entire set of nearly identical paralogs is kept coherent throughout evolution. Gene conversion is also responsible for preventing the divergence of individual copies of the ribosomal RNA cistron [207] and histone genes [208]. In many cases genes evolving under concerted evolution are arranged in genomically localized clusters.

Part II

Technical Background

CHAPTER 5

Basic Workflow for RNA Sequencing Focused on tRNAs

Contents

A transcriptome is defined as a complete set of RNA transcripts in a cell. The quantity of transcript is specific to a particular stage of development or physiological state [209]. Understanding the transcriptome is essential for the interpretation of the functional elements of the genome and the elucidation of the molecular components of cells and tissues. Thus, the analysis of the transcriptome is necessary to understand cell development and the pathogenesis of diseases. While researchers try to understand how the transcriptome shapes biology, various technologies have been developed for its analysis. Such technologies make it possible to characterize the fine architecture of the transcriptome, which includes multiple isoforms of non-coding RNAs, gene fusions, and single nucleotide variants. Quantification of gene expression changes during developmental stages or under different conditions can be additionally investigated without prior knowledge [210]. In general, gene expression is the process by which the genetic blueprint of a gene is translated into functional, biologically active products. Usually, these gene products are proteins, however, in non-protein-coding genes, e.g., transfer RNAs (tRNAs), microRNAs (miRNAs) or small nuclear RNAs (snRNAs), the products are functional RNAs.

Previous transcriptomics studies mainly relied on hybridization-based methods, e.g., northern blotting, real-time reverse transcription- polymerase chain reaction (PCR) and microarry analysis technologies. In contrast to northern blotting and real-time reverse transcription-PCR, microarrays can be employed for genomewide profiling. Hybridization-based methods have several limitations: they require a high amount of RNA material, previous knowledge of the genome sequence is prerequisite, and cross hybridisation as well as varying background noise may occur. Cross hybridization occurs when a homologous transcript is hybridized instead of the queried gene. Comparison of gene expression across different experiments is often difficult and may require elaborate normalization methods [209, 210].

The development of high-throughput next-generation sequencing (NGS) technologies revolutionized RNA analysis. RNA sequencing (RNA-seq) eliminated several challenges posed by hybridization-based technologies. However, RNA-seq experiments are not only able to capture the entire transcriptome, but also allow an quantitative measurement of individual gene expression and the discovery of novel transcribed regions in an unbiased manner [211].

The basic workflow for performing RNA-seq can be divided into library construction, sequence data generation, and final data analysis. Depending on the RNA species to be inves-

tigated, a specific experimental design is necessary to obtain biologically relevant information. This section describes a basic workflow of creating RNA-seq data and their bioinformatic analysis strategy focused on tRNAs.

5.1 Library Construction

Construction of sample libraries requires different steps (see **Fig. 11**) starting with the total RNA isolation from the cell or tissue, and elimination of ribosomal RNA (rRNA). Since over 90% RNA present in human cells is of rRNA, its removal is necessary to obtain information about the true diversity of the transcriptome present in the remaining RNA pool. Alternatively, polyadenylated (poly-A) transcripts (see Section 1.1) can be filtered directly to obtain only messenger RNAs (mRNAs), miRNAs and small nucleolar RNAs (snoRNAs). Poly-A transcripts can be separated by hybridization of total RNA with oligo (dT) primers covalently attached to a substrate, typically magnetic beads. The poly-A transcripts can then be isolated by magnetic separation technology. This sequencing method is commonly known as poly-A-selected RNA-sequencing (mRNA-seq). To obtain other types of RNAs, such as tRNAs and other non-poly-A transcripts, ribo-minus RNA sequencing (rmRNA-seq) is used, where highly abundant RNAs are depleted through hybridization capture followed by magnetic bead separation (for details see W. Zhao et al. [212] and Cui et al. [213]).

Larger RNA molecules must be fragmented into smaller pieces (200–500 bp) due to the size limitations of most common sequencing platforms. In the next step, a reverse transcriptase is added to generate complementary DNA (cDNA) of the desired RNA transcripts. For this reaction a short primer that is complementary to the 3'-end of the RNA is required to direct the cDNA synthesis. Subsequent to cDNA synthesis, adapters are ligated to the cDNA, allowing the interaction with the specific sequencing platform. The cDNA for each experiment can optionally be indexed with a barcode so that these experiments can be pooled into a single lane for multiplexed sequencing. Finally, the RNA library is amplified using PCR. Loss of strand information during RNA amplification can be avoided by chemical labeling or single molecule sequencing. Another bias that may occur is that many identical sequences can be retrieved from amplified cDNA libraries. These could be PCR artifacts or a true reflection of abundant RNA species [195, 211, 214].

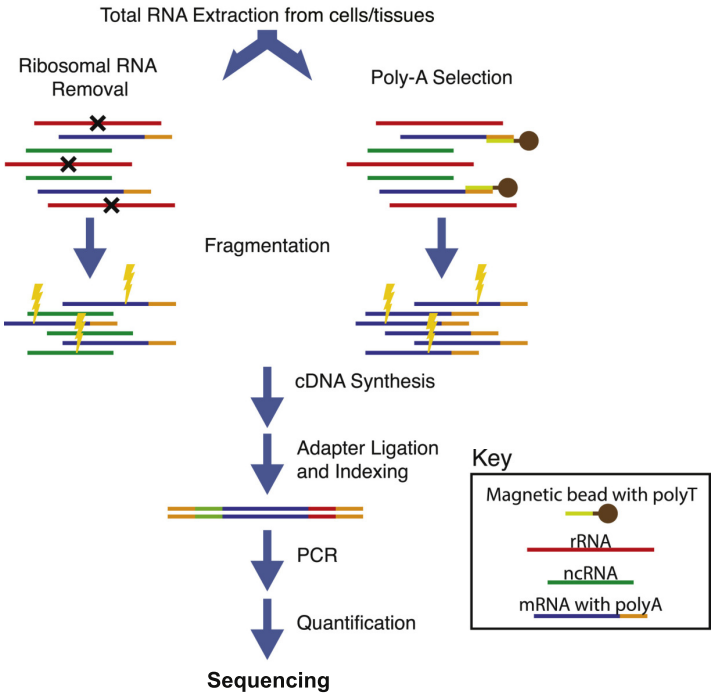

Figure 11: RNA-seq library preparation workflow. First, total RNA is extracted from the cell or tissue. Ribosomal RNA (rRNA) is then either removed (left) to enrich other RNAs, e.g., non-coding RNAs (ncRNAs), or polyadenylated (poly-A)-tailed messenger RNA (mRNA) is isolated (right). Larger RNA molecules are fragmented into smaller pieces, followed by complementary DNA (cDNA) synthesis. Adapters are added based on the sequencing platform and each experiment is indexed for sample identification. After subsequent RNA enrichment by polymerase chain reaction (PCR), the library can be used for sequencing. The figure is modified after Chaitankar et al. [214].

5.2 Challenges in Library Construction for tRNA-seq

Different methods for tRNA quantification have been developed, highlighting the various challenges arising for high-throughput analysis of the tRNA pool within a cell. Since mature tRNAs have a short, single-stranded 3'-end and a double-stranded 5'-end (see Section 3), essential steps in high-throughput tRNA sequencing such as adapter ligation and reverse transcription are difficult to perform. Although tRNAs comprise a very small proportion of total RNA in the cell [215], most approaches require a tRNA-specific enrichment step to filter out the tremendous background of rRNA in the ligation reaction. Shigematsu et al. [216] developed the YAMAT-seq method to circumvent these problems. YAMAT-seq allows for

a selective tRNA amplification without prior enrichment [216] using a specific adapter that hybridizes to the 3'-CCA end and is then ligated to the 5'- and 3'-ends. However, as adapter oligonucleotides are simultaneously fused to the tRNA 5'- and 3'-ends, only tRNAs that are fully reverse transcribed into cDNA are amplified, while prematurely terminated cDNAs, e.g., due to modifications and/or structural obstacles, are lost. Hence, no information concerning tRNA fragments or tRNAs carrying a substantial amount of nucleoside modifications (see Section 3.5) are retrieved in YAMAT-seq. Pang et al. [217] overcame this problem by introducing a two-step adapter ligation, allowing to recover such cDNA fragments. This approach, however, requires an extensive purification of tRNA fractions out of a total RNA reparation, and the authors apply five consecutive high performance liquid chromatography preparation steps of the reaction intermediates, likely resulting in a considerable loss of valuable tRNA material. Thus, an urgent need remains for methods that are easy to handle and combine adapter ligation without prior tRNA isolation. For a comprehensive investigation of the tRNA pool, a method should also include tRNA-derived cDNA fragments, as they contain valuable information concerning expression of tRNAs and specific modification positions.

5.3 RNA Sequencing Using Next Generation Sequencing

After library construction, cDNA fragments are sequenced with NGS to obtain short sequences from one end (single-end) or from both ends (paired-end). Although, paired-end sequences are preferable for *de novo* transcript discovery or isoform expression analysis [218, 219]. Most common used NGS platforms for performing RNA-seq are, e.g., Illumina, Life Technologies, and Helicos BioSciences [220]. Analysis in this work were performed exclusively on Illumina platforms, which will be explained in detail. Illumina NGS platform uses a sequence-by-synthesis system which is able to sequence millions of cDNA fragments in parallel [214].

With Illumina sequencing technology, library fragments are clustered before the actual sequencing takes place. Cluster generation is performed by passing denatured library fragments through a flow cell that randomly hybridize on a lawn of complementary Illumina adapter oligonucleotides (see **Fig. 12A**). The extension of the flow cell oligonucleotides with the hybridized library fragment as a template results in a newly synthesized strand. A complementary copy of the newly synthesized strand is then generated through bridge amplification. During

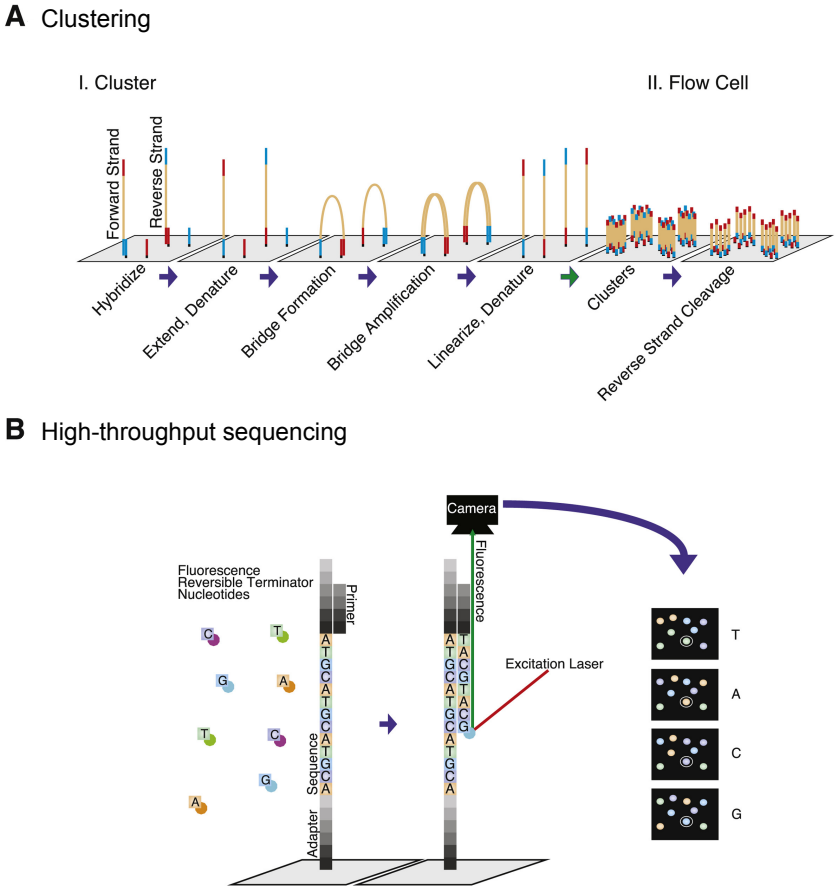

Figure 12: Illumina sequencing workflow. (A) Cluster generation is performed by random hybridization of library fragments to the complementary adapter oligonucleotides of the flow cell. Complementary adapters are then extended, amplified by bridge amplification polymerase chain reaction, linearized and denaturated. Bridge amplification is repeated several times to produce clusters of identical copies of the library fragments. After cleavage of the reverse strands, only the forward strands are retained. **(B)** Fragments are primed so that reversible terminator nucleotides can incorporate. Nucleotides are fluorescently labeled allowing imaging through laser excitation. The label is cleaved off after imaging so that the next nucleotide can incorporate. This process is continued until the predefined sequence length is displayed. The figure is modified from Chaitankar et al. [214].

bridge amplification, the strand bends to hybridize to a complementary oligonucleotide of the flow cell so that the polymerase can extend the complementary strand. Bridge amplification is repeated several times to produce clusters of complementary copies of the original fragments. The reverse orientation strands are then cleaved so that only the forward strands are primed. In

the subsequent high-throughput sequencing, cluster synthesis is performed by primer extensions with fluorescently labeled nucleotides (see **Fig. 12B**). Fluorescently labeled nucleotides are reversible terminators allowing only one nucleotide base to be incorporated. The clusters on the flow cell surface are imaged by laser excitation which returns a single color corresponding to the incorporated nucleotide. After imaging, the fluoresce label is cleaved off and the next nucleotide can be incorporated. This process is repeated until the sequence of each fragment can be determined [214]. The size of the final fragments (typically 50–200 bp) corresponds directly to the used sequencing reagents, i.e., more chemistry cycles generate longer fragments. Library fragments that are shorter than the pre-defined length contain adapter sequences in the final synthesized fragments.

5.4 Data Filtering

Each sequencing technology produces inferred sequences of bases corresponding to the library fragments, referred to as sequencing reads. In addition, a quality score is associated to each base estimating the accuracy of the base call. Most sequencing technologies use a so called Phred quality score that is calculated as follows:

$$Q = -10 \cdot \log_{10} p$$

were p is the probability that the called base is incorrect. The Illumina quality score is asymptotically identical to the Phred score for low error probabilities, but it is smaller for higher propabilities [221], given by:

$$Q = -10 \cdot \log_{10} \frac{p}{1-p}.$$

It happens that reads contain parts of the adapter sequences that were synthesized to the RNA fragments during library construction. Since low-quality and adapter containing reads can affect downstream analysis, they must be discarded and trimmed. Trimming tools such as `Cutadapt`, `FASTX-Toolkit`, `SolexaQA-BWA` and `Trimmomatic` process the reads without a noticeable loss of the covered reference genome [222]. Additional quality criteria such as the analysis of GC content, overrepresented k-mers, and duplicated reads are calculated in order to detect PCR artifacts or contaminations. The values are experiment- and organism-specific, but they should be homogeneous for replicated samples [223].

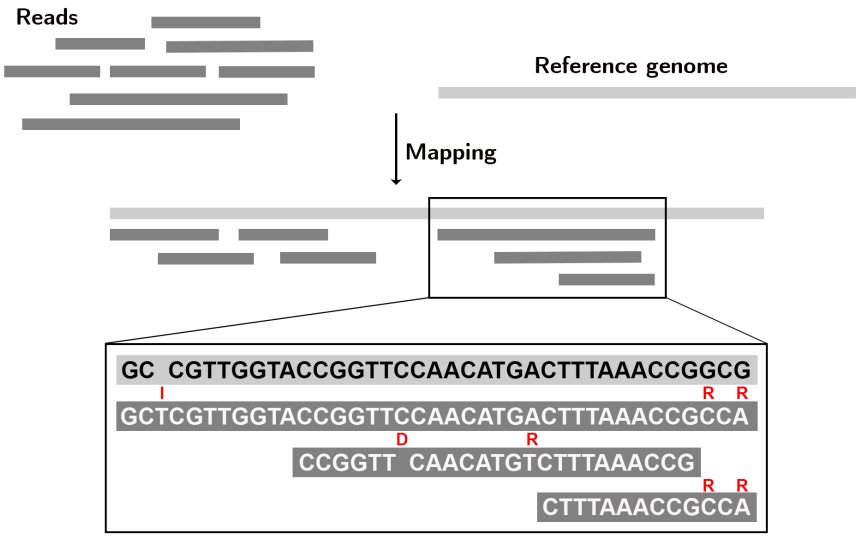

Figure 13: Scheme of the read mapping process. Read mapping is a process to align a set of reads on a reference genome (top). Thus, the location of the reads on the reference genome (middle) can be investigated. During read mapping, the following edit operations are performed: base matches, mismatches (R), insertions (I), or deletions (D).

5.5 Read Mapping to the Reference Genome

Each sequencing run produces millions of individual sequence reads that alone provide no information about the biological context. To get biological meaning from the data, the read must be aligned to a reference genome to find their approximate origin (see **Fig. 13**). In case of NGS data, this sequence alignment is termed as "mapping". The remainder of this section explains the fundamentals of sequence alignments. On this basis, the topic of read mapping is further addressed in more detail.

5.5.1 Overview of Sequence Alignment Methods

A sequence alignment is a rectangular arrangement of two (pairwise) or more (multiple) sequences so that similar features are ordered in one column to reflect their evolutionary relationship. The intention is to maximize the sum of similarities by inserting gaps in the sequences in order to align homologous positions. A specific combination of edit operations are necessary for the arrangement. When comparing two sequences a and b at position n, edit

operations can be defined as follows:

match (M) a and b share the same nucleotide,

mismatch (R) a and b have different nucleotides,

insertion (I) b has a extra nucleotide (gap in a), and

deletion (D) a has a extra nucleotide (gap in b).

The minimum number of non-matching operations that are necessary to align two sequences define the Levenshtein distance, which quantifies the dissimilarity of two sequences [224]. For example, the Levenshtein distance of two sequences a and b is $d_{edit}(a, b) = 6$:

```
edit transcript =  M M D M I M M R R D M R
              a =  A C C G - A U A G C C A
              b =  A C - G U A U U A - C G
```

In principle, a sequence alignment enables the annotation of an unknown sequence using a known sequence as a template, showing the differences between the aligned sequences. Alignments can be used to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences [225]. If aligned sequences share a common ancestor, sequence mismatches can be interpreted as nucleotide substitutions. Gaps can be assigned as insertion or deletions (indels) introduced in one or both sequences since the divergence of their most recent common ancestor. Highly conserved sequences or motifs indicate a similar structural or functional importance as their common ancestor.

Pairwise Sequence Alignments

Based on the length of the compared region of two alignments are either global, local or semi-global. Global alignments attempt to align every residue in every sequence (see **Fig. 14A**). All possible alignments are scored based on the needed changes and an optimal set of alignments is retained. A popular algorithm to compute optimal global alignments is the Needleman–Wunsch algorithm [226]. The algorithm is based on dynamic programming that relates the optimal

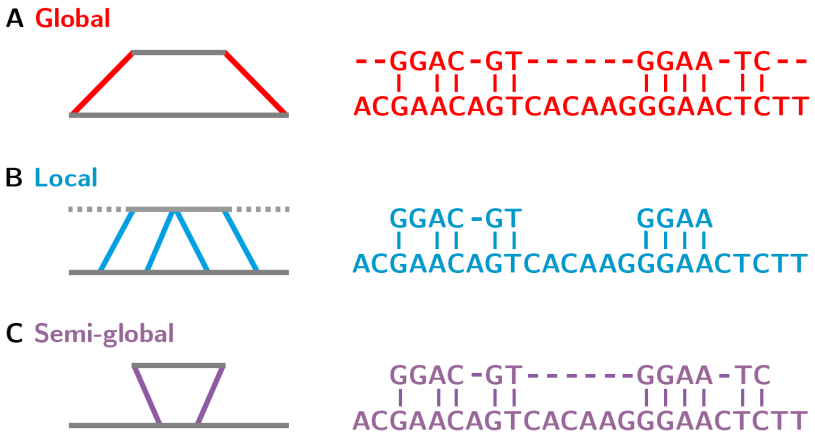

Figure 14: Overview of different sequence alignment methods. Pairwise **(A)** global, **(B)** local or **(C)** semi-global alignments are shown. **(A)** Global alignments attempt to align every residue in every sequence (end to end alignment) and may end up with multiple gaps if the sequences differ in length. A global alignment contains all letters from both sequences. **(B)** Local alignments find regions with high similarity between the two sequences. Thus, only substrings of the sequences are aligned, which helps to identify regions of similarity within long sequences that are widely divergent. **(C)** A Semi-global alignment matches a complete sequence to region of the second sequence. Semi-global alignments are useful, e.g., when one sequence is much shorter than the other. In that case, the short sequence should be globally aligned but only a local alignment is desired for the longer sequence.

alignment of two sequences a and b to the optimal alignment of subalignments in a repeated manner. Subalignments are stored in a matrix D as following:

$$D_{i,j} := \max \begin{cases} D_{i-1,j-1} + \delta(a_i, b_j) & \text{(Mis)match,} \\ D_{i-1,j} \quad + \delta(a_i, \gamma) & \text{Insertion,} \\ D_{i,j-1} \quad + \delta(\gamma, b_j) & \text{Deletion,} \end{cases}$$

where an entry $D_{i,j}$ represents the best score for aligning the prefixes $a_{1..i}$ and $b_{1..j}$. The initialization score for the empty alignment is given by $D_{0,0} = 0$. The cost function δ for edit operations is typically $\delta(a, b) < 0$ for $a \neq b$, $\delta(a, b) > 0$ for $a = b$ and $\delta(a_i, \gamma) = \delta(\gamma, b_j) < 0$ and which can be determined specifically based on the study. Global alignments are meaningful if two sequences are similar in their entirety, but fail to discover local similarities.

A local alignment is a matching of similar regions in two sequences (see **Fig. 14B**). For its calculation an adaption of the Needleman–Wunsch algorithm has been first proposed by

Smith et al. [227]. The dynamic programming Smith–Waterman algorithm identifies the local subsequences that are most similar based on the similarity score. The extension to the Needleman-Wunsch algorithm is given by the additional case "0". If the score of an alignment during dynamic programming is < 0 then $D_{i,j} = 0$ and the alignment start from the beginning.

A tool specially designed for fast local alignments of biological sequences is `BLAST`. `BLAST` uses a heuristic approach that approximates the Smith–Waterman algorithm. It is used to compare unknown sequences to a large database of annotated nucleic acid or protein sequences to determine potential homologous sequence a the query sequence [228]. The `BLAST` algorithm divides the query sequence q in short words s of length W that are stored in a list l. Each word is then locally aligned against the database d to detect their alignment hits t. The resulting segment pairs (s, t) (seeds) are extended with neighborhood words of l as long as the score $\delta(s, t)$ exceeds the threshold T. T displays the best score for shorter extensions. A seed is called a high-scoring segment pair (HSP) if it is locally maximal and $\delta(s, t)$ is greater or equal to a given minimum score threshold C. The score of the HSPs cannot be improved by shortening or extending the seed. Finally, all HSPs are reported. `BLAST` cannot guarantee optimal alignments of query and database sequences as Smith–Waterman algorithm does. However, `BLAST` is much more efficient (up to 50 times faster) than the Smith-Waterman algorithm through the seeding [229].

Hybrid methods of local and global alignments are useful to align short sequences to larger ones. The length of the reference genome surpasses the NGS reads several orders of magnitude. A semiglobal alignment is able to map the full read to a local position in the genome (see **Fig. 14C**). To this end, the Needleman–Wunsch algorithm is modified as following: $D_{\gamma,\gamma} = 0$, $D_{\gamma,j-1} = 0$ and $D_{i-1,\gamma} = 0 = \delta(a_i, \gamma) \cdot |i - 1|$. This simple modification of the initialization of the algorithm allows to shift a along the larger sequence b.

Multiple Sequence Alignments

Alignments of more than two biological sequences are referred to as multiple sequence alignments (MSAs). They are better suited than pairwise alignments to answer evolutionary questions, as the likelihood of random similarities decreases as the number of aligned sequences increases. MSAs are more computationally expensive than pairwise alignments and therefore

require more sophisticated methods than global optimization. To built up MSAs from pairwise sequence alignments, heuristic approaches have been developed which utilize basic global or local dynamic programming algorithms. Commonly, these heuristics are based on progressive alignment techniques [230], e.g., `ClustalW` [231] and `T-coffee` [232]. Progressive alignments are not guaranteed to be globally optimal. If mistakes are made at any stage of the MSA's growth, these mistakes are passed on to the end result. Thus, alignment quality can be difficult to evaluate and their true biological significance can be unclear [233].

In a first step, progressive alignments built a so-called guide tree in which the relationships between the sequences are represented as a tree. MSAs are built in a second step, where the sequences were sequentially added to the growing MSA according to the guide tree. Calculating the order of the sequences in the guide tree based on an estimated substitution matrix is the most essential step of the algorithms [234]. Entries in a substitution matrix describe a relative rate at which one amino acid mutates into another in the course of evolution. The scoring model of the substitution matrix assigns gap penalties and substitution costs. Scoring functions can be rather simple by giving a single penalty score per gaps, or more sophisticated by introducing higher penalties for gap opening than for gap extensions, or by penalizing certain substitutions based upon biological probabilities [235]. Frequently used substitution matrices are the blocks substitution matrix (BLOSUM) [236] and the point accepted mutation (PAM) matrix [237].

5.5.2 Read Mapping

Once the cDNA sequence reads have been filtered to remove aberrant reads, they are mapped to the reference genome. Read mapping is computationally the most challenging and expensive step in the RNA-seq data analysis. The enormous amounts of data generated by NGS requires specifically designed and fast read mapping tools. Basic alignment tools are not suitable for read mapping of large data. For example, `BLAST` would take 43 hours to map 10 million 32 bp reads to the reference genome [238]. In contrast, suitable mapping tools need less than 10 minutes for the same amount of data [211].

In case of tRNAs and other repetitive sequences, mapping tools must also deal with reads which map to more than one location in the genome. The amount of repetitive sequences in

higher organisms such as human and mouse could be up to 50% of the genome [211]. This multiple mapping problem may also occur for short and/or erroneous reads [221]. Typically, mapping protocols accept only reads with map to one position in the genome (unique best match) and thus almost completely disregard with the multi-mapping ability of tRNAs reads. The post-transcriptional maturation of tRNAs (see Section 3) also makes mapping challenging. While the post-transcriptionally spliced-out introns are encoded in the genome, the CCA-ends are not. The CCA-tail implies up to three mismatches between query and target within only 76 nucleotides, which often exceeds the thresholds for mapping accuracy. The same happens for most base modifications which affect reverse transcription during cDNA synthesis leading in an incorrect base inclusion. Some tRNA modifications also cause a stop of the reverse transcription activity, which may result in an increase of very short reads and thus also of multiple mapped reads. As a consequence, specialized mapping strategies are required to analyze tRNAs with respect to both their expression levels and the patterns of chemical modifications.

With growing interest in detecting tRNA modifications by high-throughput sequencing, these issues have been addressed by using alterations of the reference sequences. Mainly, only a consolidated tRNA-transcriptome is applied as reference [239–241]. Other mapping strategies use the complete native genome extended with mature tRNA sequences [242]. Multiple mapped reads are typically filtered out resulting in a set of only uniquely mapped reads [239]. Another strategy is to apply a "any-best" mapping strategy where only one position of a multiple mapped read is retained using a mapping tool like `Bowtie` with option -k 1 [243].

Choosing the right mapping tool is crucial. The mapping tool should be able to handle the large rates of differences between reads and reference genome, the different read sizes, and it should report every multiple mapped read. There is a large number of mapping tools, but not all fulfill the requirements of tRNA read mapping. For example the popular tools `BWA` [244] and `Bowtie2` [243] typically allow only a few errors per read. `Bowtie2` is further not suitable because it is designed to handle reads longer than 50 bp [221]. Some tools, e.g., `TopHat` [245], only report one hit per read, regardless of whether several equally good hits were achieved. In contrast, `segemehl` [246] allows higher error rates, is able to handle different read lengths and report all multiple mapped reads of a comparable good score. In a benchmark study of various mapping tools, including `Bowtie2`, `segemehl`, `BWA` and others, `segemehl` had the highest

sensitivity and the lowest number of false-positive hits [247]. Apart from the high memory requirements, `segemehl` is suitable for tRNA read mapping.

The mapping tool `segemehl` uses enhanced suffix arrays (ESAs) [248] to find the best local alignments of a read. ESAs are used as index structure for the reference sequence such that occurrences of read alignments can be found efficiently without scanning the complete reference. Suffix arrays [249] store suffixes obtained from suffix trees of the reference sequence in a lexicographical order. A suffix tree [250] is a rooted tree containing all the suffixes of the given sequence as keys and their positions in the sequence as values. Suffix trees contain $\leq n$ paths from the root to their leaves where n is the number of characters of the sequence and each leaf holds one suffix. ESAs are suffix arrays with additional tables that reproduce the full functionality of suffix trees preserving the same time complexity. The space requirement in large scale applications, e.g., whole genome analysis, can drastically be reduced by using enhanced suffix arrays instead of suffix trees [251]. After indexing the reference sequence, the algorithm searches for all exact and inexact matches of all suffixes of a read. All matching suffixes are quantified as so-called seeds when a score-based maximum e-value criterion and a maximum occurrence threshold are reached. Reads are then mapped to all of the corresponding seed loci in the reference sequence using Myers' semi-global bit-vector algorithm [252]. Only mappings with an accuracy above a minimum threshold are reported. Because exact and inexact matching positions of the seeds are considered, all multiple mapping loci of a read are returned [246, 253].

5.6 Annotation of tRNAs

In order to assign the mapped reads to tRNAs, the tRNA genes must be annotated to maintain their genomic position. The most commonly tool used for predicting tRNA genes is `tRNAscan-SE` which is based on heuristic search algorithms [254, 255]. In the initial first-pass scan, `tRNAscan-SE` uses the tool `Infernal` [256] to search for tRNA-like structure and sequence similarities in the reference sequence. `Infernal` implements a special case of profile stochastic context-free grammars called covariance models (CMs) [257, 258]. A CM is similar to a sequence profile, but combines an RNA secondary structure consensus in addition to the sequence consensus. Typically, CMs contain a consensus based on sequence and structural

alignments (or single sequences and structures) generated from a variety of RNAs with specific characteristics in order to obtain high sensitivity. They capture position-specific information about how conserved each column of the alignment is, and which residues are likely. In CMs basepaired positions are considered in relation to each other allowing `Infernal` together with a position-specific scoring system to identify also RNA homologous where the secondary structure is more conserved then their primary sequence. Candidates detected in the first-pass scan are then rescanned in a second-pass scan of the `tRNAscan-SE` algorithm. The second-pass scan also uses `Infernal`, but with a higher score threshold than the first-pass, to increase selectivity and alignment accuracy. In the second-pass scan the CMs must only analyse a small fraction of the total sequence, greatly improving the search speed [254]. `tRNAscan-SE` uses heuristics to try to distinguish pseudogenes from true tRNAs, primarily on lack of tRNA-like secondary structure features and a relatively weak overall score [255]. This is important especially for many mammalian genomes that are known to have a variety of non-functional tRNAs, like short interspersed nuclear elements (SINEs) where the internal regions originate from tRNA and remain highly conserved [188].

Initially, `tRNAscan-SE` was designed to identify *bona fide* tRNAs with nearly perfect accuracy in the genomes of eukaryotes and prokaryotes [255]. This is possible due to the strong sequence conservation and preservation of a common structural layout in canonical tRNAs (see Section 2.1). As an additional feature, the tool enables maximal search sensitivity for low-scoring canonical-like tRNA sequences, but at the expense of runtime. This search strategy can be applied to bizarre mitochondrial tRNAs (mt-tRNAs) (see Section 2.2), since structural deviations sometimes lead to the complete absence of entire stem-loops in comparison to the canonical cloverleaf-like secondary structure.

Other tools such as `AWEN` [259] and `MitFi` [23] were explicitely developed to solve the computational challenging problem to annotate bizarre mt-tRNAs. `ARWEN` searches for well conserved anticodon stem structures and subsequent evaluates possible flanking dihydrouridine (D)- and TΨC (T)-stem structures considering base pairing interactions. The evaluation also provides possible inferences about the existence of an anticodon stem. `ARWEN` is limited to the fact that at least three out of four stems of the canonical structure must be present. In addition, ARWEN buys its increased sensitivity at the expense of a high false detection rate [259]. `MitFi` invokes `Infernal` to search for mt-tRNAs in the mitochondrial genome

using specialized covariance models for each of the 22 mt-tRNAs and for some of the bizarre mt-tRNA structures. For all `Infernal` hits, `MitFi` attempts to predict an anticodon. The number and length of interior stems and loops is then evaluated. An advantage compared to `AWEN` is that `MitFi` can recognize structures in which more than one stem is missing. `MitFi` is more sensitive than `AWEN` reaches the same precision as `tRNAscan-SE`.

Currently, no comprehensive annotation of nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs, see Section 2.3) is available. Only a single nm-tRNA annotation strategy was published [24, 57]. This strategy is based on a `BLAST` search (see Section 5.5.1) of the known nuclear and mitochondrial tRNA sequence against the nuclear genome. Hits that match to interspersed repeats or annotated tRNAs are removed. Consequently, only nm-tRNAs with high sequence conservation to mt-tRNAs can be annotated this way. Since structural conservation is not included, nm-tRNAs that have diverged at the sequence level but may have retained tRNA-like structures are not annotated. Further analysis strategies are necessary to obtain an almost complete set of nm-tRNAs.

5.7 Detection of tRNA Modifications in RNA-seq Data

Nucleotide modifications (see Section 3.5) may affect reverse transcription and thus become visible in RNA-seq data sets in different ways (see **Fig. 15**). While several modifications do not cause any changes in the cDNA, other modifications result in a position-specific increase in the rate of sequencing errors. It is also possible that a modification blocks the complementary base-pairing interaction during cDNA synthesis which is visible as an accumulation of read terminations (RTs) at the position before the modified base [260, 261]. By stopping the reverse transcriptase it can also lead to a faulty nucleotide incorporation at the read stop [262]. These events constitute the so-called reverse transcription signature which varies with the modification type and with the used reverse transcriptase.

The 1-methyladenosine (m^1A) modification is the most prominent one which is directly visible as conspicuous accumulation of mismatches in RNA-seq. It has been reported that this particular modification of adenine (A) is typically interpreted by the sequencer as an A-to-T (thymine) transversion (purine is changed for a pyrimidine) or an A-to-G (guanine) transition (purine is changed for another purine) [263].

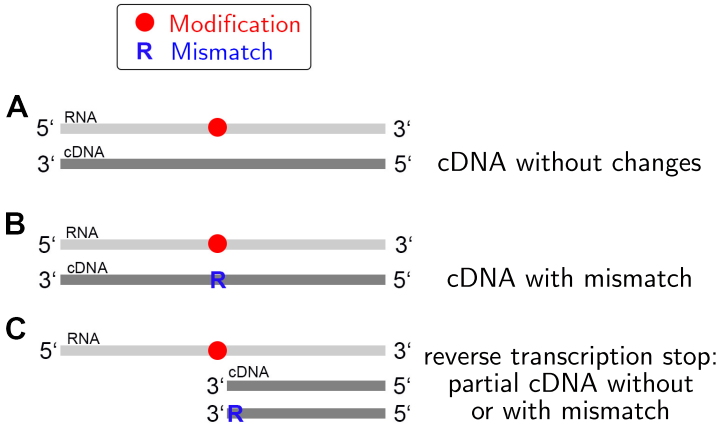

Figure 15: Different types of reverse transcription signatures. Different types of reverse transcription signatures are shown that arise from specific transfer RNA (tRNA) modifications (red dot) affecting reverse transcriptase during complementary DNA (cDNA) synthesis. While several tRNA modifications do not cause any changes in the cDNA **(A)**, others become visible after read mapping as position-specific increase of base misincorporations (R) **(B)**. Other tRNA modifications block reverse transcription activity and are detectable as an accumulation of apparent read terminations (RTs) **(C)**. RTs occur one position before the modified base. In some cases an additional mismatch is added to the RTs.

A-to-I (inosine) editing is also directly visible in RNA-seq data. While a non-edited A pairs with a T during reverse transcription, I pairs with C (cytosine) so that it can be detected as an apparent A-to-G mismatch by comparing RNA and DNA sequence [264]. Other modification such as 1-methylguanosine (m^1G), N^2-methylguanosine (m^2G), N^2,N^2-dimethylguanosine ($m^2{}_2G$) also show increased error rates in RNA-seq data. It was also reported that modifications, e.g., m^1G, m^2G, $m^2{}_2G$, 3-methylcytidine (m^3C) and m^1A lead to premature RTs [242, 265]. An overview of common tRNA modifications and their reverse transcription signatures is given in **Suppl. Tab. B2**.

A general problem in using RNA-seq to detect modifications is the need to distinguish modifications from other sources of disagreement between read and reference sequence. Usability of reverse transcription-based methods to detect modified nucleotides depends on several parameters. Pyrimidine and A bonds (C–A and U–A) are very sensitive to nuclease cleavage and the characteristics of the RNA secondary structure could also lead to RTs. The presence of RTs does not necessarily indicate a modified residue [262]. Other errors further

complicate the identification of modifications, e.g., polymerase chain reaction (PCR) errors accumulate during amplification and incorrect nucleotides are called during the sequencing step. After read mapping to the reference sequence, these errors are visible as mismatches or indels (insertions or deletions). Additional mismatches and indels may occur during read mapping due to alignment errors and the genome sequence itself. Since these errors cannot be fully controlled when preparing the library, sequencing and read mapping, noise must be separated from true signals during the modification calling step [266].

To call only true signals and filter noise, general purpose variation callers have been developed, e.g., `bcftools` of the `samtools` suite [267, 268] and `GATK` [269]. Both tools are designed to call for genetic variants from next-generation sequencing (NGS) data and apply probabilistic variant calling methods. However, these tools can also be employed to identify genomic positions with increased levels of sequence variations. In brief, a probabilistic algorithm is a combination of a Bayesian model [270] and a Maximum Likelihood approach to calculate posterior and error probabilities of the variant. Bayes' formula is applied to calculate the posterior probability of the genotype at a particular site based on the read data. Usually, the genotype with the highest posterior probability is chosen. This probability is used as a measure of confidence. In order to separate true variants from errors, likelihoods of genotypes can be calculated from the quality values (Phred quality score) of the reads provided by the sequencing platform, taking into account the expected error rate for each individual read at a location [271]. Quality scores of all reads at the particular site are rescaled and the genotype likelihood is calculated directly by the product of the probabilities of the genotype of each read. Another technique that has successfully been incorporated into error models is per-base quality recalibration using empirical data [272]. During quality recalibration error rates are calculated based on prior knowledge about error patterns for each possible nucleotide substitution [273].

A tool specifically developed for the identification of RNA modifications by reverse transcription misincorporation sites is `HAMR` [261]. `HAMR` tests for high confidence (quality score >30, error probability $<1/1000$) mismatches and for significance by ruling out that the changes are merely sequencing errors and by excluding genetic variants or editing sites. In addition, the tool characterizes the identified modification sites based on the entries in the tRNAmodviz database [119].

While some modifications directly affect reverse transcription activity, other modifications require a preliminary chemical treatment to become visible (see **Suppl. Tab. B2**). A chemical treatment of RNAs converts an originally modified nucleotide to a different form that can be detected via sequencing. Mainly, the treatments can lead to a conversion of the modified nucleotides such that they can be processed by the reverse transcriptase. Other treatments also induce a block of the primer extension by the reverse transcriptase. As result the read-out of the treated nucleotide is different to that untreated sample. Thus, modifications can be detected by comparing the treated to the untreated samples. Nevertheless, it is often difficult to distinguish modifications, which are directly visible via mismatches, and experimental noise. This is the case when the frequency of modifications is low or the sequencing coverage is limited. To overcome this issue specific chemically treated RNA-seq data can also be used.

Various RNA-seq protocols which depends on chemical treatment have been developed to detect RNA modifications. For example bisulfite sequencing uses a bisulfite treatment of RNA or DNA before routine sequencing resulting in a specific read-out [274–277]. The bisulfite treatment converts all cytosine (C) residues to uracil (U), but leaves 5-methylcytidine (m^5C) residues unaffected (see **Fig. 16A**). Thus, only m^5C residues are still readable as C after sequencing.

Apart from nucleotide transformation, various chemical treatments for modification-specific generation of RTs are developed. For example, base-pairing properties of pseudouridine (Ψ) do not differ from standard uridine which can only be recognized by introducing a chemical modification. Carbodiimides like 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene (CMCT) are used to acylate one of the nitrogen positions of the base which results in Ψ-CMC adducts that block reverse transcription [166, 169], see **Fig. 16B**. Extended alkaline treatment hydrolyses uracil-CMC adducts, while Ψ-CMC remains intact allowing the specific detection of Ψ. Further, hydrazine-mediated cleavage facilitates reliable identification of m^3C residues [264] depicted in **Fig. 16C**. The treatment induces a breakage of the RNA backbone by cleavage of both U and m^3C, while unmodified nucleotides do not break. Finally, an accumulation of apparent RTs can be found for U and m^3C. Another example is a sodium borohydride ($NaBH_4$) treatment which reduces the stability of the saturated pyrimidine ring found in D. The product of D reduction is an open pyrimidine ring that is no longer able to base pair with any other nucleotide. The ring cleavage is followed by a breakage of the RNA chain

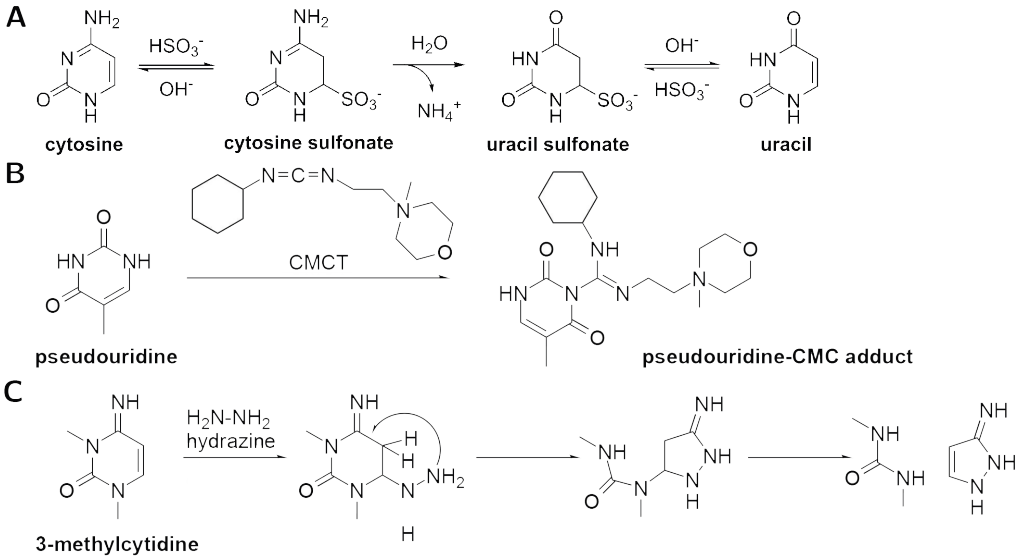

Figure 16: Specific chemical treatments for modification detection. (A) Detection of 5-methylcytidine (m^5C) by bisulfite treatment. Addition of bisulfite (HSO_3^-) to RNA leads to deamination of cytosine to uridine. Based on the resistance of m^5C to deamination reactions, sequencing after bisulfite treatment will reveal m^5C residues as cytosine signals. **(B)** Identification of pseudouridine (Ψ) by 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate (CMCT) treatment. When RNA is treated with CMCT, the N3 of Ψ is acetylated, resulting in a Ψ-CMC adduct. Adducts block reverse transcription and become visible in RNA sequencing as read terminations (RTs). **(C)** Chemical reaction for 3-methylcytidine (m^3C) detection. Hydrazine-mediated treatment induces a breakage of the RNA backbone by cleavage of both uracils and m^3Cs. An accumulation of apparent RTs is finally found for m^3C and uracil.

(see **Fig. 17A**) which facilitates position determination by reverse transcription [262]. Detection of 7-methyl-guanosine (m^7G) can be additionally achieved by using $NaBH_4$ treatment. m^7G reduction by $NaBH_4$ leads to the formation of a basic site in RNA followed by the cleavage of the RNA chain by β-elimination as depicted in **Fig. 17B**.

Instead of inducing read stops, one can also generate the production of complete cDNAs of modification which shows (low) RTs even in untreated samples. In AlkB-facilitated RNA methylation sequencing (ARM-seq) [242] and Demethylase-thermostable group II intron RT tRNA sequencing (DM-tRNA-seq) [265, 278], a treatment by the dealkylating enzyme *E. coli* AlkB demethylates m^1A, m^1G and m^3C. Demethylating results in the production of full-length cDNAs, while the untreated sample shows RTs.

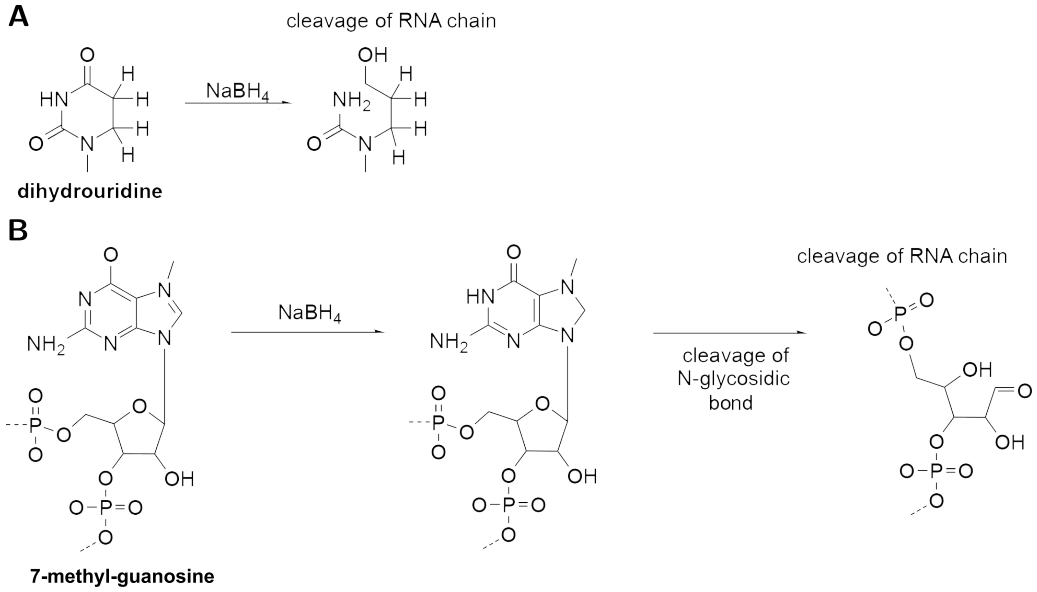

Figure 17: Sodium borohydride treatments for modification detection. (A) Sodium borohydride ($NaBH_4$) treatment for dihydrouridine detection. Cleavage of the dihydrouridine ring upon reduction by $NaBH_4$ is followed by a breakage of the RNA chain which facilitates position determination of dihydrouridine by reverse transcription. **(B)** 7-methyl-guanosine modifications are frequently present in the variable region of tRNAs. Its detection can be achieved by aniline-induced cleavage of the RNA chain by β-elimination after its reduction by $NaBH_4$.

CHAPTER 6

The Theory Behind Synteny-Based Orthology Identification

Contents

The reconstruction of detailed evolutionary histories of gene families is a prerequisite for dating and understanding innovations, for example see Capra et al. [279] and Holland [280]. It plays an important role in the emerging field of forward genomics [281]. Of particular importance is the distinction between orthologs and paralogs [203], see Chapter 4. Orthology detection is usually based on evolutionary distances that are estimated from sequence similarities, and proceeds either directly using a "reciprocal best match" approach [282] or indirectly by computing a gene phylogeny and its reconciliation with the species tree (see Kristensen et al. [283], Dalquen et al. [284], and Altenhoff et al. [285] for reviews). Both approaches make the assumption that distinct genes evolve essentially independently, so that their evolutionary distance is strongly correlated and thus can be inferred from sequence similarity. In the case of concerted evolution, however, this is not possible because the sequences of the family members within each species are essentially identical [286]. Even paralogs that have escaped concerted evolution carry no informative signal about the time before their escape.

Consequently, a completely different approach is required. The most reliable alternative source of information is syntenic conservation, i.e., preservation of relative positions within the genomic DNA sequence. It was exploited in Rogers et al. [196, 197] to devise a strategy in which the query tRNA is embedded in intervals of flanking sequences whose size is increased until a unique `BLAST` match (see Section 5.5.1) in the target genome is found. In this manner, an approximation to orthology is obtained. By the time the uniqueness condition is satisfied, intervals may extend across entire transfer RNA (tRNA) clusters, calling for methods to further refine the orthology assignments.

In the following section our developed concept for synteny-based orthology identification is described, which is more systematic than previous studies as explained above. Our concept is based on Velandia-Huerto et al. [376] titled by *Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies*. Their implementation and applications are described in Section 7.4 and Chapter 11, respectively. The biological background about synteny-based tRNA events is given in Chapter 4.

6.1 From Synteny to Candidate Orthologs

We consider a set Σ of species or genomes. Each genome $a \in \Sigma$ comprises a discrete set of loci. Genomic coordinates establish an order relation $\prec$ among loci. Since genetic elements have an intrinsic reading direction the order $\prec$ is either the same or the inverse of the coordinate system. We write $\bar{u}^a$ for the reverse complement of locus u^a on genome a. Note that $u^a \prec v^a$ is equivalent to $\bar{v}^a \prec \bar{u}^a$. Since the reverse complement of a locus is also a valid locus we arbitrarily choose the orientation.

For a subset of loci we assume that they evolve independently by vertical inheritance and are not subject to duplication (see Chapter 4) in the set of species under consideration. We say that two tRNAs t^a and t^b in genomes a and b, respectively, are $1:1$ orthologs, if t^a is the only ortholog in genome a of t^b in genome b, and *vice versa*. Therefore we can compute $1:1$ orthologs of p^a in a set of species $\Sigma_p \subseteq \Sigma$. We will refer to such a set of orthologous loci $p = \{p^a | a \in \Sigma_p\}$ as an *anchor*. An anchor p may connect all or only a subset $\Sigma_p \subseteq \Sigma$. The orthologs within an anchor are defined to be oriented in the same reading direction. Therefore, if p and q are anchors with $p^a \prec q^a$ then $p^b \prec q^b$ for all $a, b \in \Sigma_p \cap \Sigma_q$. That is, we assume that anchors preserve synteny including relative reading direction in the set of genomes of interest. We can therefore write $p \prec q$.

Now we consider a set T of loci of interest; in our case tRNAs. None of the $t^a \in T$ gives rise to an anchor, i.e., we assume that the multiple, nearly identical sequences are present in the genome. We make two basic, simplifying assumptions:

(S1) There are anchors p and q such that $p^a \prec t^a \prec q^a$.

(S2) A pair of anchors can be chosen such that the relative order of homologous loci is preserved between p and q.

Both assumptions are approximations to reality. Condition (S1) stipulates that the locus t^a of interest is not too close to the end of a contig, scaffold, or chromosome. It will be violated essentially by incomplete data and flaws in genome assemblies. Condition (S2) is a more severe restriction. It allows only unduplicated vertical inheritance and tandem duplications of individual loci. It explicitly rules out genome arrangement between anchors sufficiently close to the locus of interest and also neglects tandem duplications affecting more than a single gene.

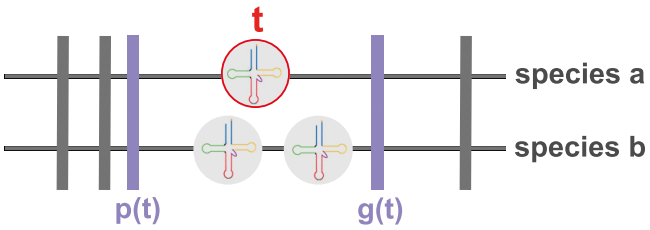

Figure 18: Scheme of tight anchors for the loci of interest. Possible anchors (grey) and tight anchors (purple) for t^a (red bordered circle) into species b are indicated in boxes. The tight anchors are the anchors closest to t that connect species a and b. By synteny, the only possible orthologs of t are the two tRNA loci indicated by the white bordered circles.

In essence it forces us to treat a multi-locus tandem duplication as if it was a combination of unduplicated vertical inheritance combined with the insertion of the second copy of pair.

6.1.1 Determining of Tight Anchors

Condition (S1) allows us to obtain initial candidates for orthology assignments. We assume that they have a set of homologous elements T_a for each genome $a \in \Sigma$. If p and q are anchors with $p^a \prec t^a \prec q^a$ then any $t' \in T_b$ with $t' \prec p^b$ or $q_b \prec t'$ cannot be co-ortholog of t^a in genome b.

Practical difficulty is that in general we might not have anchors that cover all species of interest but only a subset of them. For any "query" locus t^a and any species $b \in \Sigma$, $b \neq a$ we therefore define a *pair of tight anchors for* t^a *into* b as a pair of anchors $p_b(t^a) := \{p^a, p^b\}$ and $q_b(t^a) = \{q^a, q^b\}$ such that (i) $p^a \prec t^a \prec q^a$ and (ii) the pair $(p_b(t^a), q_b(t^a))$ is minimal in the sense that there is no further anchor $u = (u^a, u^b)$ with $p^a \prec u^a \prec t^a$ or $t^a \prec u^a \prec q^a$, see **Fig. 18**.

Under our assumption (S1), there is a unique pair of tight anchors of t^a into b for every $b \in \Sigma$. In practice, however, there may be exceptions: in the case of genome arrangement or a fragmented genome assembly the anchor points $p_b(t^a)$ and $q_b(t^a)$ may be located on very far apart or even on different chromosomes, contigs, or scaffolds. Condition (S2), or even a much weaker locality assumption, implies that only the homologs $t' \in T_b$ enclosed by the pair of tight anchors for t^a are possible co-orthologs of t^a in genome b.

The exact nature of the anchors is irrelevant. In a very conservative approach, known sets of orthologous protein-coding mRNAs can be used. If a more fine-grained resolution is desired, one can use e.g., blocks of genome-wide multiple sequence alignment (MSA). Since our method relies on the use of multiple whole-genome alignment blocks, the following section describes the basic procedure to create such alignments blocks using the MULTIZ software.

Multiple Whole-genome Alignments as Tight Anchors

The task of calculating MSAs of entire genomes is associated with a multitude of new challenges for alignment methods (see Section 5.5.1) due to extreme long genome sequences which are highly heterogenous in function and conservation rate. In addition, possible duplication (repetition of a sequence segment), inversion (reversed sequence segment), and translocation (sequence segments have been exchanged between distant parts) events have to be considered.

A popular tool for multiple whole-genome alignments is MULTIZ [287]. Before the MULTIZ software can be used, the whole genome alignment problem have to be splits into a set individual distinct local alignment blocks. A block is an optimal alignment between two or more genome sequences. A designated "reference" sequence, is present in each block of a set referred to as ref-blockset. Each position of the reference sequence appears exactly once throughout the ref-blockset, averting overlapping regions between blocks. To overcome this issue, the threaded blockset aligner (TBA) software can be applied. TBA computes the alignment blocks under the assumption that the matching regions occur in the same order and orientation in the given sequences. The MULTIZ program dynamically performs the alignment for three or more sequences (ref-blocks), based on pairwise alignments generated by BLASTZ [278]. One of the biggest differences to a conventional alignment program is that MULTIZ is able to merge two existing MSAs (sets of blocks) into one larger MSA (see **Fig. 19**) instead of just aligning individual sequences. The software treated the MSAs as sequences for which a pairwise alignment is generated algorithmically similar to the progressive methods described in Section 5.5.1. For a detailed description of the algorithm, see Blanchette et al. [287].

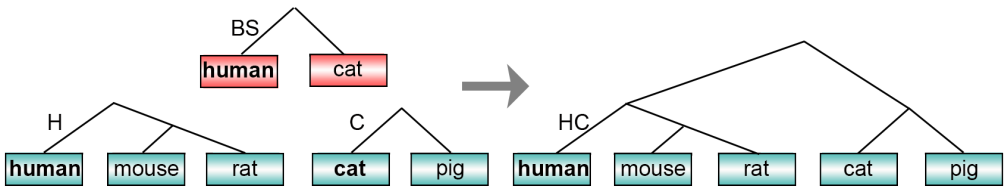

Figure 19: Merge of aligment ref-blocks by MULTIZ. A human ref-blockset BS based on a pairwise alignment generated by BLASTZ guided the merge of the human ref-block H and the cat ref-block C into a new multiple sequence alignment HC. HC contains all sequences from both ref-blocks H and C. Reference species are written in bold letters.

6.1.2 Candidate Graph Construction

From the sets of homologous loci T^a and a collection of anchors on Σ the *candidate graph* Γ_c can be construct as follows. The vertices Γ_c are the annotated homologs, i.e., $T = \bigcup_{a\in\Sigma} T_a$. An edge between $t^a \in T_a$ and $t^b \in T_b$ is inserted if $p_b(t^a) \prec t^b \prec q_b(t^a)$, i.e., if t^b is located between the pair of tight anchors from t^a into b. In order to accommodate some local inversions and/or assembly errors one might want to relax this definition and to draw an edge between t^a and every locus $t \in T_b$ so that $p_b(t^a) \prec t \prec q_b(t^a)$ or $p_b(t^a) \prec \bar{t} \prec q_b(t^a)$. By construction, the true orthology relation is a sub-graph of Γ_c, see **Fig. 20A**. Its nodes are the tRNAs and there is an edge between two tRNAs if they are possibly orthologuous, thus if they are flanked by the same tight anchors and belong to distinct species.

The graph Γ_c is not sufficient to completely solve the orthology problem because in general two tRNA loci t_i^a and t_j^a will not be separated by anchors. The available anchors in fact may enclose entire tRNA clusters (see **Fig. 18**). For tRNAs, however, we can clearly distinguish subgroups by sequence similarity. In particular, tRNAs of different isoacceptor families (see Section 1.1) and within these, most subgroups with distinct anticodons, exhibit clearly separate sequences. We therefore can prune the edge set of Γ_c by removing all edges that connect tRNA loci with clearly distinct sequences. We therefore require that the genetic distance satisfies $d_G(t_i^a, t_j^b) < \varepsilon$ for all edges of the pruned candidate graph, which we denote by Γ_a, see **Fig. 20B**. The threshold ε can be chosen as an upper bound on divergence of genes in phylogenetic range of interest.

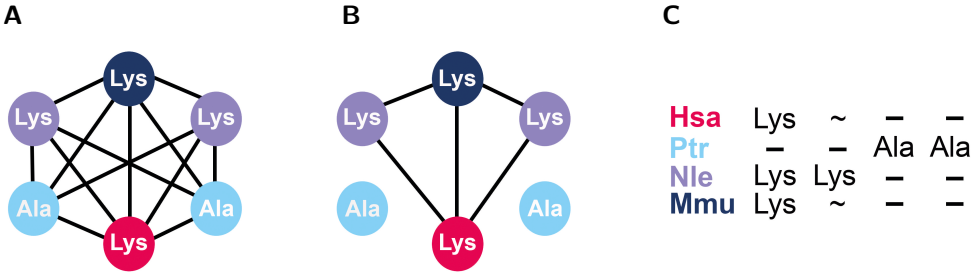

Figure 20: Step-wise refinement of the candidate graph Γ_c. **(A)** The graph Γ_c represent the possible orthology assignments among tRNA loci derived from the synteny anchors. Only genes from different species can be orthologs, hence no edges connect loci in the same species. **(B)** Based on sequence similarity edges are removed between tRNAs from different isoacceptor families. **(C)** A modified Needleman-Wunsch alignment algorithm is used to identify order-preserving subgroups. This step admits local tandem duplications but not duplications of larger subclusters. Species abbreviations: human, *Homo sapiens*: Hsa; rhesus macaque, *Macaca mulatta*: Mmu; gibbon, *Nomascus leucogenys*: Nle; chimapanzee, *Pan troglodytes*: Ptr. tRNA isoacceptor classes abbreviations: alanine: Ala; lysine: Lys.

6.2 Order Preservation within Clusters

Assumption (S2), stipulates that co-orthologous loci preserve relative order. In the context of tRNA clusters, this amounts to the assumption that tRNAs within a gene cluster proliferate by means of single gene tandem duplications or by retroposition-like insertions.

The relationship between clustered tRNAs in two species corresponds to a generalized version of an alignment problem. In order to see this, it is necessary to consider each tRNA cluster as an ordered list of tRNAs and tRNA pseudogenes t_i^a and t_j^b in the two genomes a and b. For the sake of the argument, it makes sense to first neglect gene duplications and consider insertion, deletion, and remolding only. In this case the correspondences between orthologous loci form an order-preserving matching in the induced subgraph of Γ_a restricted to every pair of species. This amounts to an alignment of the tRNA loci in T_a with those in T_b with alignment edges allowed only between loci that are connected by an edge in Γ_a.

6.2.1 Modified Needleman-Wunsch Alignment to Account for Duplications

In order to account for local, i.e., order-preserving duplications an alignment model (see Section 5.5.1) can be simply extended. In the usual setting of matchings, one locus t_i^a can match

at most a single locus t_j^b. Otherwise one of t_i^a and t_j^b is deleted. This is called a $1:1$ alignment. In the simplest extension also $1:2$ and $2:1$ matches are allowed, i.e, two positions (t_i^a, t_{i+1}^a) may collectively match a single position t_j^b, or *vice versa*. More generally, $p:q$ matches may be considered. Such extensions to one-to-many or many-to-many matches lead to a quite simple modification of the Needleman-Wunsch [226] algorithm.

As stated above, condition (S2) is a restrictive approximation that rules out tandem duplications of subclusters larger than a single locus as well as any local genome rearrangements. More inclusive assumptions could be made instead. The full duplication-loss alignment problem that allows copying of subclusters of arbitrary size is approximable hard [288], but a practicable dynamic programming heuristic is available [289]. Recently, it was extended further in `OrthoAlign` to include also genome rearrangements [290]. In principle these approaches could be substituted into our workflow.

A simpler model can be used (see **Fig. 20C**) since it avoids the problem of estimating weight parameters for complex duplications and rearrangements operations. As an alternative to alignment-like approaches for disentangling the history of individual loci it may also be fruitful to consider generalizations of gene order methods.

6.2.2 Estimation of Orthology Graph

The alignment edges predicted by the pairwise generalized alignment algorithm serve our best estimates for the orthology relation. For $1:2$ duplications an edge is inserted from the "original" to both "copies"; in the more general case of $p:q$ duplications, we accept all edges of the complete bipartite graph corresponding to the $p:q$ duplication. Superimposing all pairwise alignments yields the *estimated orthology graph* Γ_o, conceptually shown in the bottom row of **Fig. 21**. It contains only edges between tRNAs that can be orthologs according to their sequence similarity, and all connected components of Γ_o are order preserving since their edges result from the order-preserving alignment step. By construction Γ_o is a spanning subgraph of Γ_a, which in turn is a spanning subgraph of the initial candidate graph Γ_c. In general, Γ_o will consist of many small connected components, each comprising members of a single tRNA family that locally has expanded and contracted by duplication and loss events.

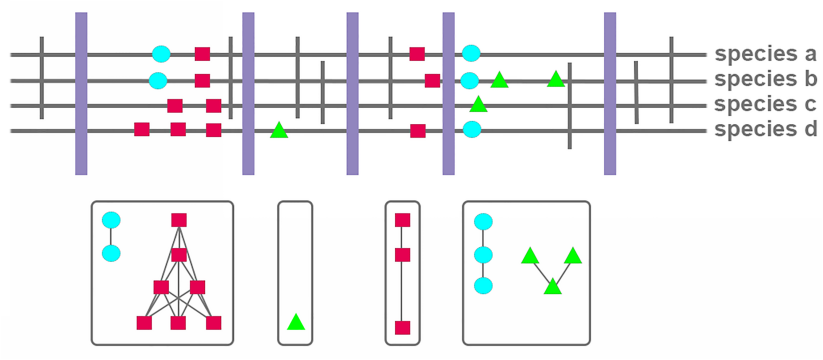

Figure 21: Scheme of step-wise orthology identification. Top: Genomic organization of tRNAs (colored symbols). Possible anchors (grey) and tight anchors (purple) are shown as boxes. Anchors are defined based on unique sequence alignment blocks. These anchors subdivide the genome into syntenic clusters forming the connected components of the graph of candidates Γ_c, here shown for blocks of a genome-wide alignment as delimiters. Each cluster forms a connected component of Γ_c. Bottom: Pairwise generalized list alignments lead to an estimate of the co-orthology relation for each group of homologous tRNAs. Each of these estimated graphs is then corrected to the nearest cograph.

6.3 Cographs and Orthology

Recent results in phylogenetic combinatorics [291–296] show that orthology relations are cographs. There are many equivalent characterizations for this well-studied class. In particular, G is a cograph if it does not contain a P_4, a path on 4 vertices, as an induced subgraph [297]. In particular, complete graphs are cographs. A cograph is associated with a unique cotree, which corresponds to the (not necessarily fully resolved) gene tree with labels at the interior vertices that identify speciation and duplication events, respectively [291, 292].

We expect that Γ_o is already a very good approximation to tRNA orthology. Various sources of noise, however, will introduce violations of the cograph structure. Therefore, the orthology estimates can be improved further by editing Γ_o to the nearest cograph. This amounts to inserting and deleting the minimal number of edges so that all P_4s are destroyed. Although the cograph editing problem is nondeterministic polynomial time (NP) hard [298], this is not a practical problem here. It is not difficult to see that the connected components C_i of Γ_o can be edited independently of each other [292]. Empirically, we observe that most connected components of Γ_o are complete graphs and this already correct cographs.

From the final, corrected orthology estimates $\hat{C}_i$ it is now straightforward to infer the evolutionary events. The cographs $\hat{C}_i$ themselves provide direct information on the tandem duplication events. To this end it suffices to convert the $\hat{C}_i$ into its equivalent cotree [297], from which the duplication events can be directly read off. Deletion events as well as gain events in which a particular locus was settled can be obtained by assigning each of the $\hat{C}_i$ to the species tree. A Dollo parsimony (see Farris [299] for algorithmic details) approach to derive the numbers of gain, losses, and duplications from the co-ortholog groups can be applied. Here, duplication events identified from the cographs $\hat{C}_i$ can be counted separately from gains.

Part III

Methodology

CHAPTER 7

Bioinformatic Analysis

Contents

7.1 Annotation of tRNAs

Cytosolic transfer RNAs (tRNAs, see Section 2.1) were annotated with `tRNAscan-SE` v2.0 [255] (see Section 5.6) using the default model for eukaryotes or bacteria. For the mitochondrial tRNA (mt-tRNA, see Section 2.2) annotation the `-M` option was applied additionally.

The secondary structure predicted from `tRNAscan-SE` depict only tRNA stem and loop structures in dot-bracket annotation with missing information about the exact nucleotide position in the partially present regions. For the unambiguous assignment of each tRNA position, especially those that are only partially present, nucleotides must be fitted to the standard tRNA model. The standard tRNA numbering system has been developed by Sprinzl et al. [300] and is shown in **Fig. 22**. Apart from a few known exceptions, the standard tRNA is numbered from 1–76 and depicts nucleotides present in each tRNA structure. Aside from the partially present positions located at the 5'-end (position 0), in the dihydrouridine (D)-loop (positions 17a, 20a and 20b), and in the variable loop (V-loop, position 47), a variable arm is located between nucleotides 45 and 46 obeying the base-pairing rules. The numbering of the nucleotides in the variable arm positions begins with the letter "e" followed by the numbers 1 to 5 in the loop region. To indicate base-pair formation "e" is followed by 11 to 17 at the 5'-branch and 27 to 21, in the reverse order at the 3'-branch [300].

In order to fit tRNAs to the standard tRNA model, the annotated tRNA sequences were aligned against the database entries of the `tRNAdb` database [301] containing the missing secondary structure notation. `BLAST` v2.4.0 [228] was used for sequence alignment and only those tRNA database entries were selected that showed the closest evolutionary similarity to the annotated tRNAs (see Section 5.5.1). Due to the high evolutionary distance of some tRNAs to the database entries, the notation of some tRNAs had to be adjusted manually based on the base pair rules.

7.2 Mapping of tRNA Reads

7.2.1 Data Pre-Processing

To trim adapter sequences and low quality portions of raw reads, different read trimming tools were applied depending on the underlying next-generation sequencing (NGS) method (see Sec-

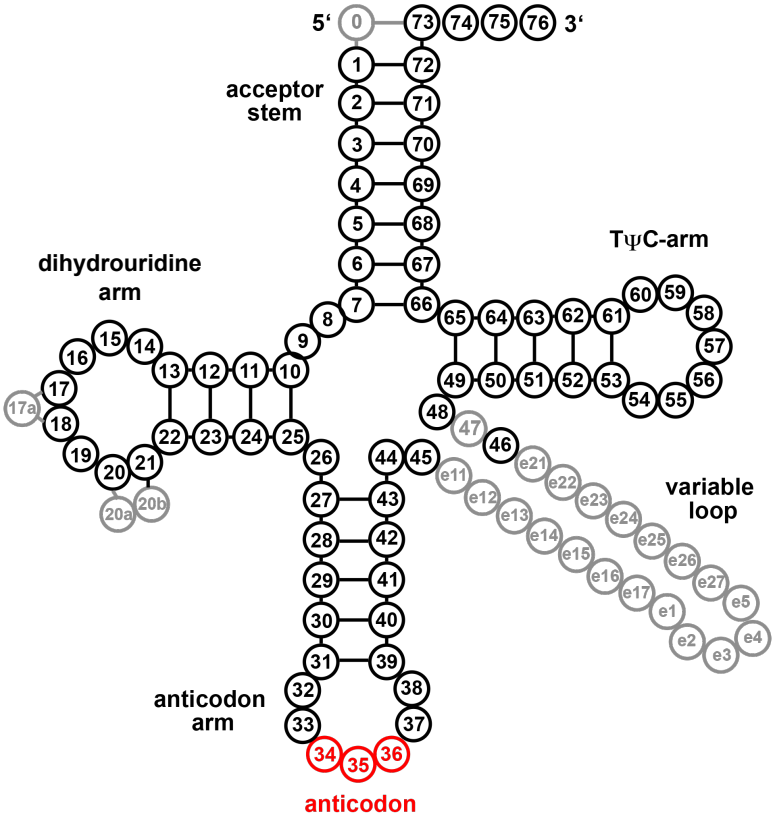

Figure 22: Standard tRNA numbering system. Canonical transfer RNA (tRNA) cloverleaf secondary structure with the standard nucleotide numbering system. Black circles are always present nucleotides numbered from 1 to 76. Gray circles are not present in each tRNA structure which are at position 0 (5'-end), position 17a, 20a and 20b (dihydrouridine arm), position 47, and a variable arm located between nucleotides 45 and 46 (variable loop). The numbering of the variable arm starts with the letter "e" followed by the numbering 1 to 5 in the loop region, by 11 to 17 at the 5'-branch, and 27 to 21, in the reverse order at the 3'-branch. The second digit identifies the base-pair of the variable stem. Red circles represent the anticodon.

tion 5). In our analyses based on open source ribo-minus RNA sequencing (rmRNA-seq) data, we used `BBDuk` from the `BBMap` toolkit v36.14 [302] with a k-mer size of ten allowing to use shorter 8-mers at the end of the read and a Hamming distance of one. To pass the determined quality filter, read quality needed to surpass a Phred score of 25 and achieve a minimal length of 50 nt and a maximum length of 100 nt after trimming of adapter and low quality bases.

Using our own transfer RNA (tRNA)-enriched RNA sequencing (RNA-seq) data (see Chapter 9 and Section 10.3), `BBDuk` resulted in insufficiently trimmed reads which probably

occurred due to our custom primer structure. Thus, after testing of different trimming tools, the straightforward tool `Cutadapt` v1.16 [303] produced comparatively good results. Here, a quality cutoff of 25 is also defined and a maximum error rate of 0.15 allowed. As the used reverse transcriptase reacts sensitively to tRNA modifications, the reads can be very short. To capture these short reads as well, reads of 8 to 95 nts length after trimming of adapter and low quality bases were selected. `FASTQC` v0.11.4 [304] was applied for standard pre- and post-trimming quality control for all samples.

7.2.2 tRNA Library Preparation and Genome Pre-Processing

Two different tRNA libraries were generated from the annotated tRNA genes, with the omission of pseudo-tRNA genes, in order to create precursor tRNAs (pre-tRNAs) and mature tRNA sequences (see Chapter 3). For the pre-tRNA library, 3'- and 5'-genomic flanking regions were extracted from the genome with `BEDTools` v2.25.0 [305] and added to the corresponding tRNA sequence in order to simulate the elongated 5'-leader and 3'-trailer sequences. The mature tRNA library was created by appending 3'-CCA tails to the tRNA genes. Intronic sequences were removed from both libraries.

Based on the multiple copies of tRNA genes with many identical and nearly identical tRNA genes (see Section 4), tRNAs of the same type were clustered according to identity thresholds of 97%, 98%, and 100%. A consensus sequence for each cluster was constructed. We applied `usearch`, v9.2.64, [306] which is a centroid-based greedy algorithm (for algorithmic details see Edgar [306]), for the tRNA clustering.

All annotated tRNA genes, including pseudo-genes and sequence regions identical to annotated tRNAs, were masked in the native reference genome applying `BEDTools` v2.25.0. The pre-tRNA sequences were appended as additional "chromosomes" to the tRNA-masked genome which is now referred to as *artificial genome*.

7.2.3 Read Mapping and Filtering

To permit error-tolerant mapping and keeping track of "all best" alignments, reads were aligned to the artificial genome using `segemehl` v0.2.0-418 (see Section 5.5.2) requesting a minimal accuracy of 80% (allowing up to 20 mismatches for a 100 nt long read). Anticipating a high

density of modification-induced mismatches and short reads due to the nature of tRNA or read terminations induced from the reverse transcriptase, we opted for a reduced mapping sensitivity at the expense of longer computation time: we allowed a maximum of 3 mismatches in the seed regions, increased the e-value cut-off to 500 for seed extension, and considered at most 1000 mappings per seed. Reads that do not map the concatenated pre-tRNA chromosomes of the artificial genome or reads which map to the remaining tRNA masked genome were filtered out, respectively. Reads of possible pre-tRNA origin were selected by identifying reads which partly align to the flanking regions of the pre-tRNA chromosomes, tolerating the CCA overhang of mature tRNA reads.

The remaining reads were mapped in a subsequent step against all mature tRNA sequences. In another variation not all tRNAs were added, but only clusters of more or less similar (97%, 98%, and 100%) sequences were used as reference sequences. Those two methods were called *unclustered* and *clustered*, respectively. `segemehl` was again used for the mapping with the same custom parameter settings as in the first mapping step, except for 85% mapping accuracy. These settings lead to best results and accurately map reads while preserving modifications. In the first mapping step, which acts as a filter to remove reads that do not map within the defined boundaries of mature tRNAs, the reduced accuracy is required to keep the false negative rate low. Finally, reads with mismatches in the CCA tail were filtered out, since there can also be possible pre-tRNA sequences. An additional filter step was applied for RNA-seq data which exhibits specific enrichment and selection of tRNA sequences (see Chapter 9 and Section 10.3). Here, depending on the adapter ligation protocol, the tRNA sequences are sequenced beginning either from the 5'- or, in our case, from the 3'-end. Thus, reads which do not map to the 3'-end or the CCA tail are discarded to further minimize false hits. Finally, to analyze the true origin of the mapped reads, we applied three different read filter strategies:

all were all score-optimal alignments of a read were considered,

phased a middle-ground strategy that allows also multiple mapping reads only if they show exactly the same misincorporation pattern for all alignment positions and

unique only uniquely mapped reads were retained, i.e., those that have a unique score-optimal alignment to the reference genome.

7.3 Detection of Modification Sites in tRNAs

7.3.1 Identification of Significant Base Misincorporation Sites

For the detection of tRNA modification (see Section 3.5) which are directly visible as accumulation of nucleotide mismatches in mapped reads it is important to distinguish random sequencing errors and mapping artifacts from true misincorporation sites (see Section 5.7). For this purpose we applied three different modification site calling approaches after read mapping (see Section 7.2):

(i) our *ad hoc* `Pfropfen` variant caller [397],

(ii) `GATK`'s `UnifiedGenotyper` v3.6-0-g89b7209 [269],

(iii) `bcftools` with the `mpileup` and `call` option v1.8 [268].

We tested our custom `Perl` implementation `Pfropfen` because we expected that random incorporated bases would produce sequencing patterns that are systematically different from those produced by single nucleotide polymorphisms in large cohorts of individuals. There is no reason to assume that variation calling algorithms expecting polymorphism data as input would perform particularly well with sequencing errors introduced by chemical modifications. Therefore we aimed to call modification sites exhibiting more errors than expected for several substitution events. To this end, we independently tested for each site whether for a given substitution it occurred more often than expected by chance. The background distribution of the misincorporations is assumed to resemble a binomial distribution. The p-values obtained in this manner are merged over all replicates, using Fisher's method, resulting in a merged p-value for each site which was corrected for multiple testing [307]. The implementation of `Pfropfen` can be found at `https://github.com/fabou-uobaf/Helferlein/blob/master/Pfropfen`. `Pfropfen` was applied with the following parameters: *-delta* 0.5 *-cov* 4 *-qual* 20 *-pval* 0.01 *-noterm -indel -windsor* 1 . These settings translate as follows:

-delta only sites with an overall substitution rate below 0.5 are considered for the background rate determination,

-cov only sites covered by at least 4 reads are considered for modification calling,

-qual only bases with quality score above 20 are considered,

-noterm premature read termination events are not considered,

-indel inserts and deletions are not considered,

-windsor the highest and lowest p-values are removed before Fisher's method is applied,

-pval only sites with a multiple testing corrected p-value below 0.01 are reported.

To apply `GATK`'s `UnifiedGenotyper` variant caller, we realigned all mapped reads with `GATK`'s `IndelRealigner` which minimizes the number of mismatching bases, especially around indels, across all reads. We adjusted the minimum Phred-scaled confidence threshold at which variants should be called to 50, in order to reduce false positive calls.

Applying `bcftools` we used at first the `mpileup` command to generate genotype likelihoods and read coverage at each genomic position. In a second step we used the `call` command with the `-m` option for rare-variant calling. Variants under the Phred-scaled confidence threshold of 20 were filtered out. We used different threshold values for both variant callers, as these need to be adapted to tool specifics. For each tool, we considered only called modification sites with a coverage of more than 10 reads.

7.3.2 Detection of Modification Sites by Read Terminations

Identification of chemical tRNA modifications which are not visible by reverse transcriptase signatures requires the use of RNA-seq data based on protocols that include chemical treatments leading to specific modifications (see Section 5.7). These RNA-seq protocol make use of chemically generated read terminations (RTs) and misincorporation signatures produced by reverse transcriptases so that they yield a specific read-out in the subsequent sequencing in comparison to an untreated library. Here, the technical implementation of the post-mapping (see Section 7.2 for the mapping procedure) analysis is explained addressing the profiling of modified sites in treated RNA-seq data.

Library Normalization

For direct comparisons of the control and treated libraries, the raw data are scaled library- and replica-wise. Library-wise normalization was performed by scaling the number of mapped reads for each tRNA position to the number of reads of the whole sample. We weighted each multiply mapped read by division through the number of loci it maps to. Thus, multiply mapped reads are not weighted stronger than uniquely mapped reads. For replica-wise normalization, the mean of the replicas for each tRNA and position was calculated.

Determining Significant Modification Sites

To differentially quantify effect sizes in read termination coverage between treatment and control, we used the fold change (FC) as measure. FC is defined as the ratio between the two conditions. For each tRNA and position n the FC is calculated by the equation:

$$FC_n = \frac{RT^+_{n+1}}{RT^-_{n+1}\gamma + \alpha},$$

were RT^+ and RT^- representing the number of read terminations of the treated and untreated library, respectively. In case of modified tRNAs the reverse transcription terminates one position before the modified nucleotide. Thus, the number of RTs was counted at tRNA position $n+1$, since the reverse transcription is 3'- to 5'-directed based on the 3'-adapter ligation of the used library preparation protocols. For inter-sample normalization, the untreated library is scaled according to the library size of the treated sample with the total number of mapped reads per tRNA's cluster as scaling factor given by $\gamma = \sum Reads^+_{clust} \ / \ \sum Reads^-_{clust}$. Some tRNA positions do not show read terminations within one or both conditions. Zero values in both conditions result in an infinite logarithmic FC, although no effect is present. In the other case, a zero value in the control leads to the undefined division by zero. To solve those zero-frequency problems, a pseudocount of $\alpha = 1$ is added to make all counts strictly positive.

In another measurement we applied the Poisson distribution to determine statistically significant differences in treatment compared to the negative control [308] as follows:

$$P(x \geq k) = P_\lambda(k) = 1 - \left(\frac{\lambda^k}{k!}e^{-\lambda}\right),$$

where e is the Euler's number, $k = RT^{+}_{n+1} + \alpha$, and $\lambda = RT^{-}_{n+1}\gamma$. We implemented a separate maximum likelihood estimator of λ from the mean of the cluster's RTs of the untreated sample. This tRNA cluster-wise estimation of λ reduces the influence of outliers. Outliers may be also caused by strong or weakly expressed modifications. Thus, we considered that the RT expression is tRNA cluster-specific. Since we expect an enrichment of RTs only in the treated sample, we used the RT expression of the negative control for λ.

To decrease the false discovery rate (FDR; see Section 7.6) we adjust the p-values received by the Poisson distribution by applying the Benjamini-Hochberg procedure [309]. To this purpose, all p-values were sorted in ascending order and ranked, such that the smallest p-value had rank one. Each individual p-value was classified if it was smaller than the Benjamini-Hochberg critical value, calculated by:

$$p_i \leq \frac{i}{m}Q,$$

where i is the rank, m is the total number of test events, and Q is the FDR of 0.05. All tRNA positions showing a p-value lower than 0.01 were accepted as significantly enriched RT sites in the treated sample.

7.4 Creation of a Synteny Map for tRNA Orthology Identification

A key step in the reconstruction of the evolutionary history of tRNA genes is the creation of a synteny map. The synteny map harbors information about syntenic tRNA gene clusters which are subdivided by genomic anchors (see Chapter 6 for the technical background). To this end, we annotate tRNA genes (see Section 7.1) and used multiple sequence alignments (MSAs, see Section 5.5.1) created by the `MULTIZ` pipeline [287] (see Section 6.1.1) to define tight anchors following the approach described in Section 6.1. We emphasize that MSAs in general do not correctly align multi-copy genes since well conserved multi-copy elements are often used for the generation of anchors for the MSA itself. This creates artifacts because the initial alignment step by construction cannot distinguish between the individual copies of a family of loci that is subject to concerted evolution. The `MULTIZ` pipeline allows the same sequence to appear in more than one alignment block. This is the case in particular for duplicated genome regions. In order to remove all such ambiguities, we filtered the set of

alignment blocks in the following manner: alignment blocks were first converted to a sorted BED format describing position of each alignment block within a corresponding genome. For each annotated tRNA, the 5'- and 3'-adjacent alignment blocks without overlaps with any tRNA gene or other alignment block were identified.

Although the construction of synteny map is rather conceptually simple, practical issues arise from less than perfect genome assemblies. tRNA genes that could not be placed in an unambiguous genomic context because no anchor or only a one-sided anchor was available were excluded from the analysis of tRNA clusters. These tRNAs were included in detecting remolding events since the analysis was mainly based on alignments.

7.5 Analyses Concerning nm-tRNAs

7.5.1 Search for Genomic Loci of nm-tRNA Genes

We applied two different annotation tools for the detection of nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) located in nuclear genomes (see Section 2.3). First, we used the tRNA annotation tool `tRNAscan-SE` v2.0 (see Section 5.6) in a modified manner, applying the integrated mitochondrial tRNAs (mt-tRNAs) search mode (`-M` option) not to mitochondrial genomes, but to nuclear sequences. Regardless of whether the default (20 bits) or a very low (0–20 bits) cutoff score was used for filtering hits, same results were returned. The second search strategy was to apply the `Infernal` v1.1.2 [256] software as search engine with specific covariance models (CMs) for each of the 22 mt-tRNA families taken from `MiTFi` [23], see Section 5.6. All `Infernal` hits were retained to help to find nm-tRNAs which are not well conserved. For each nm-tRNA annotation strategy, we used nuclear mitochondrial DNA (NUMT) sequences obtained from Telonis et al. [24] (*NUMT-based* approach) or the entire nuclear genome (*genome-based* approach) as reference. Since we ran `Infernal` separately for each of the 22 CMs, we sometimes found the same nm-tRNA for different CMs with comparable scores. To define the primordial mt-tRNA for each nm-tRNA, we used synteny information (see Section 4) provided by NUMTs, since the exact mitochondrial origin of each NUMT is known. For nm-tRNAs annotated outside of NUMTs, the nm-tRNA hit with the highest score was retained.

To determine the transcriptional origin, e.g., protein-coding, non-coding, pseudogenic, and exonic, we assigned transcript annotations to the nm-tRNAs. We defined nm-tRNAs as intergenic if they could not be assigned to an annotated transcript. nm-tRNAs that are located in introns of any kind of transcripts are also called nuclear-encoded intronic mitochondrial-derived tRNAs (nim-tRNAs).

7.5.2 Measurement of Evolutionary Conservation

We compared the NUMTs to the extant human mitochondrial genome sequence to test for evolutionary conservation. The observed sequence divergence is in this case a sum of two independent effects: (i) the evolution of the NUMT since its insertion and (ii) the evolution of the mitochondrial genome (mt-genome) since the insertion event. We expect that the selection pressure on the mt-genome has remained neutral over time t_0 because its functionality has been preserved. Since tRNAs are among the most stringently conserved genetic elements, the mitochondrial substitution rate of mt-tRNAs is smaller than the substitution rate of the mitochondrial proteins. We, therefore, expect that the evolutionary distance d_t between nm-tRNA and mt-tRNA is $d_t = (s_n + s_t)t_0$, while for the NUMTs we have $d_p = (s_n + s_p)t_0$, where s_n is the neutral substitution rate in the mt-genome. The substitution rates for nm-tRNAs and NUMTs are given by s_t and s_p, respectively. Outliers of this linear regression with unexpectedly large values $d_p - d_t$ are then identified as the nm-tRNAs that have evolved slower than expected, i.e., those that have become subject to stabilizing selection after their insertion into the nuclear genome. Thus, the difference $d_p - d_t$ is expected to be a linear function of t_0. An schematic overview of the model is illustrated in **Fig. 23A**.

Since we are not able to calculate substitution rates and t_0, we linearly transformed the model with $s_n + s_p$. The linear transformation leads to a model (see **Fig. 23B**) enabling nm-tRNAs to be obtained as outliers that are subject to a stronger selection pressure relative to NUMTs. Therefore, we can use the sequence divergence as measurement for the evolutionary sequence conservation. We computed the sequence divergences (Hamming distance) d_t and d_p by dividing its edit distance (see Section 5.5.1) to the primordial mitochondrial sequence by its length. The edit distances were obtained by mapping (see Section 5.5.2) the sequences to the mitochondrial genome. For this purpose, we used `segemehl` v0.2.0-418 [246] with a

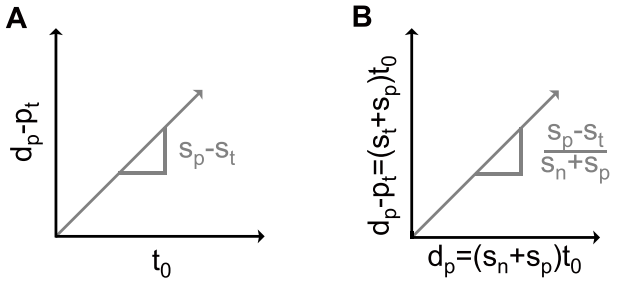

Figure 23: Model for evolutionary conservation measurement. (A) Linear model of $d_p - d_t$ of t_0. The evolutionary distance between nuclear mitochondrial DNAs (NUMTs) and mitochondrial proteins d_p can be calculated by $d_p = (s_n + s_p)t_0$, where s_n is the neutral substitution rate in the mitochondrial genome and s_p is the substitution rate of the NUMTs. Accordingly, the evolutionary distance between nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) and mitochondrial transfer RNAs (mt-tRNAs) is $d_t = (s_n + s_t)t_0$, where s_t is the substitution rate of nm-tRNAs. A linear transformation of this model with $s_t + s_p$ results in **(B)**, a liner model of $d_p - d_t$ of d_p. Outliers in the linear regression indicate nm-tRNAs which are subject to stronger or lower selective pressure in relation to the remaining NUMT sequences.

low accuracy of 50% and searched for seeds with two differences, to allow mapping of strongly degraded sequences. We calculated the Cook's distance [310] for the outlier test which was performed in R v3.6.0 using the `stats` package [311]. In general, Cook's distance shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier. Each element in the Cook's distance C is the normalized change of the fitted response values due to the deletion of an observation. The Cook's distance of observation i is:

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE},$$

where $\hat{y}_j$ is the jth fitted response value, $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i, MSE is the mean squared error, and p is the number of coefficients in the regression model. We have only considered NUMT sequences which are longer than 50 nucleotides (nts) to avoid overestimating shorter sequences.

In another measurement we assigned phylogenetic p-value (PhyloP) scores to each sequence which has been predicted from multiple genome alignments of mammals. PhyloP scores are available from UCSC [312] and can be used to detect nucleotide substitution rates that are faster or slower than expected under neutral drift in genomic sequences of different species.

7.5.3 Determining Protein Binding Sites of nim-tRNAs

To investigate the potential regulatory role of nim-tRNAs by interaction with RNA-binding proteins (RBPs), we intersected their genomic loci with a list of experimentally validated RBP binding sites. The latter is available at the GENCODE project, which hosts a repository for BED files containing binding sites of a large set of RBPs derived from eCLIP experiments. By applying the `BEDtools` suite v2.29.0 [305] we intersected genomic coordinates of nim-tRNAs with RBP binding sites on the same strand to derive a list of overlaps. RBPs that bound to each type of nim-tRNA were then annotated according to their biological function with information derived from the GeneCards database [313]. We calculated the expected coverage of RBP per nucleotide intron from intersections of the eCLIP dataset with intron annotations (ENSEMBL biomart, hg38, version 98 [314]) for each RBP in the collection. By comparing this to the RBP coverage of binding sites in nim-tRNAs we calculated the relative enrichment of RBP binding events in nim-tRNAs over background.

7.6 Performance Evaluation of Different Analysis

To determine the sensitivity and specificity of different analysis strategies we counted all true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs). All performances are expressed as:

(i) true positive rate (TPR), also called sensitivity, with $TPR = \frac{TP}{TP + FN}$,

(ii) false negative rate (FNR) with $FNR = \frac{FN}{FN + TP}$,

(iii) true negative rate (TNR), also called specificity, with $TNR = \frac{TN}{TN + FP}$, and

(iv) false discovery rate (FDR) with $FDR = \frac{FP}{FP + TP}$.

Mapping of tRNA Reads

To develop a best-practice analysis strategy to map tRNA reads to the reference genome (see Chapter 5) we used simulated data to test different strategies and handle the associated difficulties which may arise. Therefore, we simulated three replicas of human single-end

50 bp RNA-seq reads using the `RNASeqReadSimulator` [315]. The reads for each tRNA were generated with a similar expression strength and contain random sequencing errors with an overall error rate of 0.5%. Five percent of the simulated reads come from the pre-tRNA sequences and 95% from mature tRNAs. Modification sites were randomly chosen with a rate of 5% of all nucleotides from the mature tRNA library. Subsequently, the bases at chosen positions were altered within the simulated reads, following a random substitution matrix which was determined by the immediate neighboring nucleotide [316].

In addition to determining the modification pattern randomly for each genomic tRNA locus, we prepared a second test in which tRNAs with identical sequences also have identical modified positions. This set is of course easier to handle in the computational analysis. It is not clear at present whether the biological reality is closer to the *random modifications* scenario, where mature tRNAs with identical sequence are treated differently by the enzymatic modification machinery depending on their genomic origin, or to the *identical modifications* scenario, in which modification patterns depend on the mature sequence only. A careful analysis of modification patterns should be able to shed light on this question. In total our simulations consider 2,324 modified sites in the *random modification* scenario and 3,001 modified sites in the *identical modification* model. To determine the sensitivity and specificity of the different analysis steps of the simulated data we compared the predicted variation with our simulated modification sites.

Specific Enrichment and Selection of tRNA Sequences

To compare the specificity of different RNA-seq methods with our newly developed long hairpin oligonucleotide-based tRNA high-throughput sequencing (LOTTE-seq) method, the number of uniquely mapped tRNA reads showing a 3'-CCA, -CC, -C , or no 3'-CCA-end was counted. Multiply mapped reads were counted as fraction of their number of hits or filtered to obtain uniquely mapped counts.

Comparison of Modification Callers

When comparing different tools for the identification of candidate modification sites, we visually examined all identified sites. We distinguished between hits due to mapping artifacts or

misinterpreted reverse transcriptase signatures (counted as FPs) and the remaining potentially true sites (counted as TPs). In order to assign candidate sites to known modifications we used the tRNA modification information stored in the tRNAmodviz database [119]. Since the database does not contain all human tRNAs, we classified the candidate sites according to its overlap with the known modified positions of the contained human tRNA subset.

Annotation of nm-tRNAs in Nuclear Genomes

For each NUMT, the mitochondrial origin is traceable, so we are able to reconstruct the number, types, and order of mt-tRNA copies within each NUMT. Thus, we used this synteny information (see Chapter 4) to validate our obtained hits from the different analysis strategies. We count each hit as TP if the hit is located inside a NUMT following the occurrence and order of the given synteny information (see Section 4). Some NUMTs were copied from mitochondrial sequences which lack mt-tRNAs. Thus, we counted hits within such NUMTs as FPs. Hits obtained outside from NUMTs were also counted as FPs, because we have to assume that the underlying NUMT annotation is complete.

Part IV

Applications

CHAPTER 8

Accurate Mapping of tRNA Reads

Contents

To the best of our knowledge the consequences of different mapping strategies for the detection of transfer RNA (tRNA) modifications have not been investigated systematically (see Section 5.5). To close this gap, in this contribution we aim to evaluate the performance of different mapping strategies with the help of simulated RNA sequencing (RNA-seq) reads. These observations performed a best-practice workflow that uses modified genomic reference sequences to accommodate CCA tails (see section 3.3) and a reduced set of tRNA sequences that represent groups of very similar paralogs (see Section 4).

In the following chapter the development and benchmark of the best-practice mapping strategy of tRNA reads is shown. The methodical implementation is described in Sections 7.1 and 7.2 which include tRNA annotation and read mapping, respectively. The method used to call variations is specified in the Section 7.3.1 and the performance evaluation using simulated reads is described in Section 7.6. The technical background is given in Chapter 5. This chapter is based on A. Hoffmann et al. [397] titled *Accurate Mapping of tRNA Reads*.

8.1 Best-Practice Mapping Strategy

The detection of modified RNA nucleotides from RNA-seq data by means of patterns of base misincorporation requires that each next-generation sequencing (NGS) read is precisely assigned to its true genomic origin. This is of course a non-trivial task for NGS applications in general. This problem is even more difficult for tRNA modification calling due to the large number of modifications, and thus misincorporation sites in tRNAs and their multi-copy nature with many identical and nearly identical tRNA genes. This makes it virtually impossible to determine with certainty the exact genomic origin of any particular tRNA read. We therefore resorted to simulated RNA-seq data to establish a best-practice mapping strategy because the known ground truth allows us to evaluate the effect of different analysis steps on the sensitivity and specificity of RNA modification site calling. We identified three critical problems for the successful RNA modification site detection:

(i) The backend of the pipeline needs to discriminate between sites with a significant base misincorporation pattern indicative of a modification site and sites with spurious differences between the NGS read and the reference genome due to sequencing or mapping errors. To solve this problem, we use the `GATK`

framework. In addition, we devised a more naïve approach to check for sites with a significantly higher misincorporation rate for all alternative bases compared to a transcriptome-wide binomial background error model.

(ii) The most difficult challenge are the ambiguities in determining the true origin of many NGS reads. To address this issue, we used the three different filter strategies in the read alignment step: *all*, *unique*, and *phased*.

(iii) The unusual processing of tRNAs with added CCA tails and the coexistence of tRNA genes with and without introns producing the same mature product requires adjustments to the reference against which the RNA-seq data are mapped. To address this issue, we evaluated different more or less modified reference genomes and tested all combinations with above described strategies for read filtering and modification site calling and evaluated its performance.

Our baseline, the starting point of our workflow development, was the most straightforward approach: reads were mapped against the unaltered human reference genome using the *all*, *phased*, or *unique* filtering rule. Modifications were called as statistically significant misincorporation sites without further processing (see **Fig. 24**). All those approaches resulted in a reduced true positive rate (TPR) (*all*: 0.82, *phased*: 0.53, *unique*: 0.25), in an increased false negative rate (FNR) (*all*: 0.18, *phased*: 0.47, *unique*: 0.75) and in a very high false discovery rate (FDR) (*all*: 0.44, *phased*: 0.42, *unique*: 0.43) for the simulated reads containing *identical modifications*.

To get a handle on the complexity of the transcriptome in general and the tRNA transcriptome in particular, we masked all tRNA loci in the human reference genome and subsequently appended customized tRNA sequences as extra "chromosomes". Most importantly, we attempted to distinguish NGS reads derived from immature tRNA precursors and those that are produced from mature tRNAs (see Section 3). To this end we attached the reference sequences of the tRNA precursor (with flanking regions but without CCA tails and without introns) to the masked genome and parceled out reads mapping at least partially to the flanking sequences or introns. Detailed investigations of mapped reads from human esophagus muscularis mucosae tissue with the `IGV` [317, 318] confirms that the exclusion of pre-tRNA reads after the first

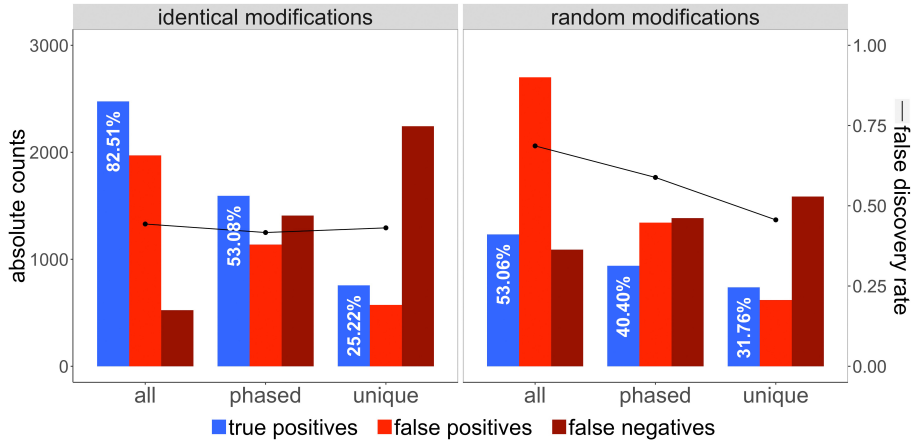

Figure 24: Evaluation of the straightforward approach. Absolute numbers of true positive (blue), false positive (red), and false negative (dark red) modification calls obtained for simulated reads containing *identical modification* (l.h.s.) and *random modification* sites (r.h.s.) in transfer RNAs (tRNAs) with the same sequence arising from distinct loci. In the most straightforward approach, the simulated reads were mapped against the native human reference genome. Additional significant misincorporation sites on *all*, only *phased*, and on only *unique*ly mapped reads were called using `GATK`'s `UnifiedGenotyper`. The *all* mapped reads option shows the highest sensitivity, but also the highest false discovery rate. The best balance between true positive calls and errors is shown for the *unique*ly mapped reads filtering method for both simulated data sets.

mapping step is helpful to reduce false positive (FP) hits originating from modified pre-tRNA reads (see **Suppl. Fig. A1**). This pre-tRNA cleaning step is more efficient than softclipping of fragments on the read ends, due to pre-tRNA reads spanning the whole tRNA. Soft clipping, on the other hand, could lead to the retention of the pre-tRNA read that is mapped to the reference and only the overhanging sequence is being cut off. The outfiltered pre-tRNA reads can be used for modification calling of tRNA precursors or be discarded. The remaining reads were mapped in a subsequent step against all mature tRNA sequences. In another variation not all tRNAs were added, but only clusters of more or less similar tRNA sequences were used as reference sequences. Those two methods ere called *unclustered* and *clustered*, respectively.

The differences between *clustered* and *unclustered* reference tRNA sequences are only a minor factor when multiply mapped reads are allowed. The *clustered* tRNA reference genome in which tRNA sequences are clustered with a 100% sequence identity performs much better in the case of *unique*ly mapped reads. For simulated reads with *identical modifications* using only reads that *unique*ly map to tRNA *clusters*, a TPR of 0.85 is achieved. Using reads mapping

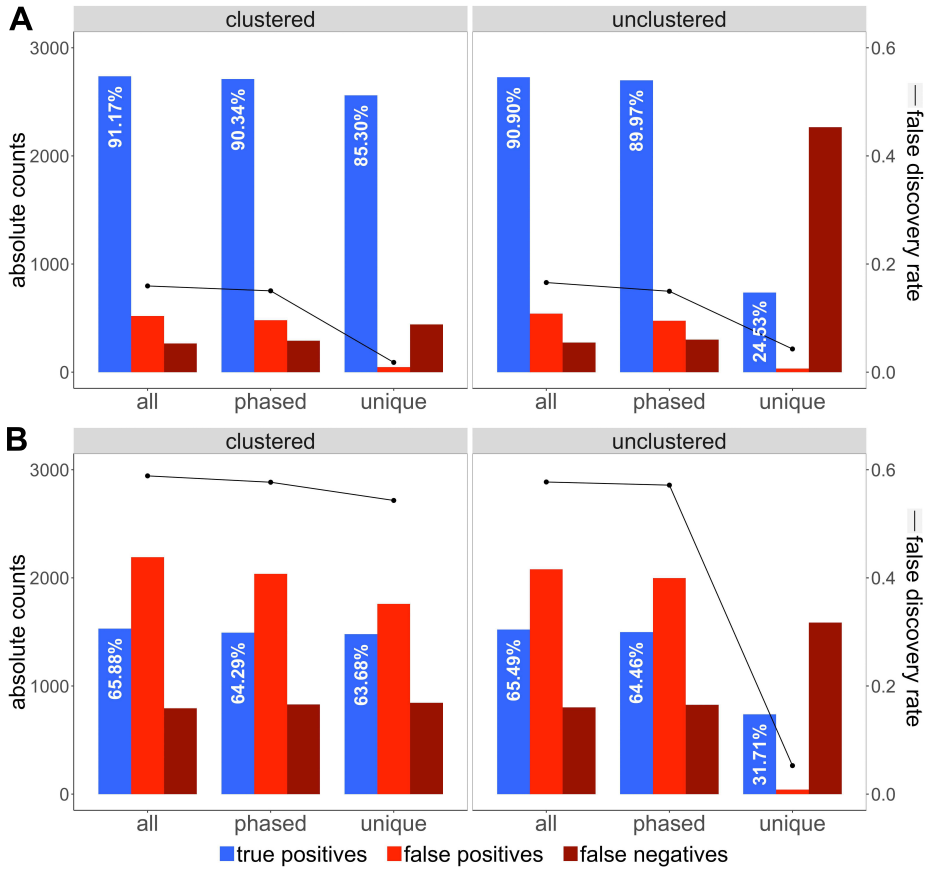

Figure 25: Comparison of read filtering strategies. Absolute numbers of true positive (TP, blue), false positive (FP, red), and false negative (FN, dark red) modification calls from the simulated datasets as well as the false discovery rate (FDR) of **(A)** *identical modifications* or **(B)** *random modification* sites. For the *unclustered* as well as the *clustered* method, the results of the called significant misincorporation sites using the `GATK`'s UnifiedGenotyper for the different read filter strategies (*all*, *phased*, and *unique*) are shown, respectively. Regarding different read filter strategies, the *unique* reads showing the best balance between the detected TPs and the errors (FPs, FNs). Using *unique* filtered reads, the *clustered* method is more sensitive and shows less errors, especially FNs, in comparison to the *unclustered* method

*unique*ly to *unclustered* tRNA results in TPR of only 0.25 (see **Fig. 25**). Correspondingly, the FNR is increasing, but the FDR remain comparable. This shows that using only *unique*ly mapped reads against a *clustered* tRNA reference genome collapsed into a single representative outperforms all alternative approaches tested here, provided specificity is the main concern.

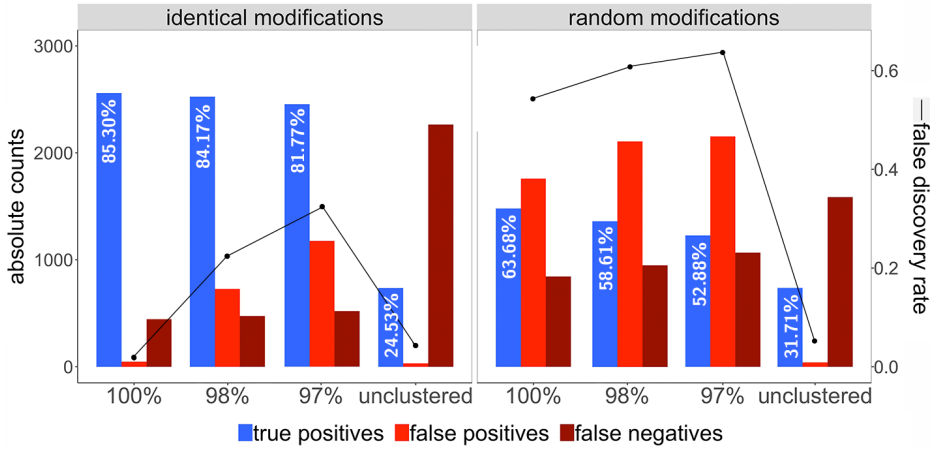

Figure 26: Comparison of clustering methods. Absolute numbers of counted true positives (blue), false positives (red), and false negatives (dark red) by analyzing simulated reads containing **(A)** *identical modification* or **(B)** *random modification* sites. Significant misincorporation sites for *uniquely* mapped reads were called using `GATK`'s `UnifiedGenotyper`. The statistical calling results for clustering with the different identity thresholds of 97%, 98%, and 100% are shown. In comparison to that, the numbers for the *unclustered method* are visualized. All examined thresholds exhibited a very similar (although smaller) true positive rate but were much less specific compared to the clustering with 100% identity in terms of the false positive rate. Indeed, the *unclustered method* shows only a few false positives, but a really low sensitivity.

Since clustering identical tRNA together seems to be a worthwhile strategy, we wondered if allowing also non-identical tRNAs to be represented in the same cluster could improve the performance even further. We hypothesized that reducing several very similar sequences to a single consensus would reduce the difficulty of read mapping, and the accumulation of reads for very similar tRNA sequences could improve the signal-to-noise ratio in the modification detection step – at least in the *identical modifications* scenario. We therefore allowed one up to three mismatches (100%, 98%, and 97% sequence similarity) between tRNA sequences assigned to the same cluster. Empirically, however, we did not observe an improvement: All examined thresholds exhibited a very similar (although smaller) TPR but were much less specific compared to the clustering with 100% identity in terms of the FDR, see **Fig. 26**.

After the read mapping procedure, alignments can be filtered with respect to the number of loci they map to. The effect of filtering on performance in general is as expected: sensitivity decreases and specificity increases from *all*, over *phased*, to *uniquely* mapped reads. The choice of filtering strategy seems to be the best way for the user to tweak the trade-off between

sensitivity and specificity. At least for the simulated data, using only *unique*ly mapped reads seems to yield the best balance between FPs and false negatives (FNs). The FDR drops from 0.16 (*all*) to 0.02 (*unique*), while the TPR only drops from 0.91 (*all*) to 0.85 (*unique*) for the *clustered* method (see **Fig. 25**).

We defined the optimal alignment(s) as the one(s) with the minimal edit distance between read and reference sequence. Base misincorporations in the reads and similarities can cause incorrect alignments whenever a misincorporation is compensated by a difference in an alternative reference location. Such cases cannot be recognized by filtering strategies. In the simulated data, $\sim$1% of the *unique*ly mapped reads do not map to the correct position (see **Suppl. Fig. A2**). The applied mapping tool `segemehl` can also report suboptimal alignments. Using this feature shows that the correct alignment scored only a single mismatch worse that incorrect one in these cases. Conversely, correct optimal alignments have incorrect suboptimal alternatives that differ by a single mismatch in many cases. Thus the performance of the mapping cannot be improved by either including suboptimal read alignments or by requiring a large score gap between best and next-best alignment.

The different filter strategies (*all*, *phased*, and *unique*) produce consistent patterns of misaligned reads for both real and simulated data. This can ultimately lead to incorrect calls of modification sites (see **Suppl. Figs. A3** and **A4**). In both data sets the *unique* filtering strategy appears to be the best-practice to reduce the calling of false positive misincorporations caused by multiply mapped reads. Furthermore, we observed that tRNAs that vary only in one or a few individual nucleotides may already show different modifications even within the same sample. We conclude that tRNA modification patterns depend on the mature sequence. This suggests that our simulated data scenario containing *identical modifications* for identical sequences fits better to the biological reality than the *random modification* scenario.

We compared the two different modification site calling approaches: `GATK`'s `UnifiedGenotyper` and our *ad hoc* `Pfropfen` approach. We observed that `Pfropfen` seems to be more sensitive at the expense of reduced specificity (see **Suppl. Figs. A5** and **A6**). It seems that `UnifiedGenotyper`'s more sophisticated handling of mapping artifacts outweighs the benefit of applying a statistical model that reflects the expected counts from a random misincorporation process. Nevertheless, the slightly increased FDR indicates that there is still room for improvement of the calling procedure by tailoring it to the underlying processes.

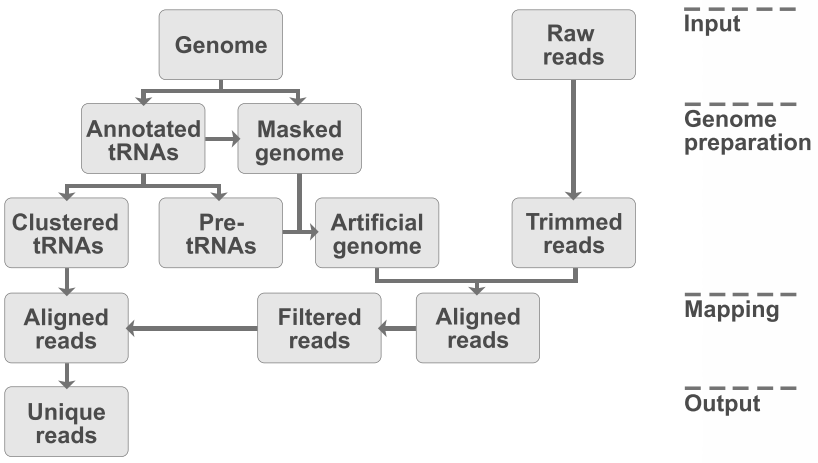

Figure 27: Scheme of the best-practice workflow for accurate mapping of tRNA reads. The top part describes the construction of the masked and artificial genome. The middle section refers to the mapping and filtering steps and the bottom layer shows the final *unique*ly filtered reads. The output reads can be used for, e.g., modification site detection.

In conclusion, we propose a best-practice workflow to detect tRNA modification sites in RNA-seq data depicted in **Fig. 27**: tRNA genes are annotated by `tRNAscan-SE`, masked in the reference genome and subsequently supplemented by tRNA sequences. In a first step pre-tRNAs were added and reads displaying specific precursor hallmarks are separated. In a second step sequences representing identical tRNA sequences are added. Only *unique*ly mapped reads are used for the follow up modification site calling using `GATK`'s `UnifiedGenotyper`. Using our simulated data and a tRNA specific mapping to handle the high density of modification induced mismatches at the reads we received a FDR of 0.02, a TPR of 0.85, and a true negative rate (TNR) > 0.99 for the *identical modifications* scenario.

8.2 Discussion

The general problem of determining the genomic origin of transcript fragments deriving from multi-copy or repetitive regions did only recently get the deserved attention [319]. In this respect, tRNAs are, due to their well defined boundaries, a special case. Nevertheless, the lessons learned can be generalized to some degree. Given that clustering of very similar tRNAs,

as long as they are not identical, leads to a decrease in modification calling performance, it becomes evident that clustering-based approaches are not suitable for more divergent classes of multi-copy elements such as Alu-repeats [320]. For the mere purpose of transcript quantification different probabilistic approaches were presented to assign reads to the most likely origin [321, 322]. Those are unfortunately not suitable for nucleotide variants, and in this sense also modification calling. An exciting proposed strategy could be to dynamically update the reference sequence based on already seen variants [323]. Such a strategy could be used to layaway from discriminable regions into indistinguishable by using the co-occurrence of modification site, if such occur densely enough.

For the moment the best available strategy to analyze tRNA-seq data consists of collapsing identical sequences together to reduce the search space and use only *unique*ly mapped reads. This strategy however can only be advised for relatively short RNA families, such as microRNAs (miRNAs) or tRNAs, due to their well defined boundaries and their convenient gene length to read length relationship. If the reference sequences to be clustered are much longer than the produced RNA-seq reads, a local clustering has to be applied, since differences at the far distant beginning can not be used to discriminate reads mapping to the very end of the region. In this thesis we surmised that our *phased* read filter, where we allowed reads to be multiple mapped but only if displaying identical misincorporation patterns, could potentially come up to such a local clustering strategy. Unfortunately, it did not live up to our expectations. Although, using *phased* reads performs half way between using *all* reads and only *unique*ly mapped reads in the native reference genome approach with respect to sensitivity and specificity (see **Fig. 24**), it does not reach the same quality of modification site calling than applying a pre-clustering of identical reference genome sequences (see **Fig. 25**). Nevertheless, it seems to be a viable option for research questions where a global clustering is not possible and sensitivity is of more interest than specificity.

8.3 Data Sources and Workflow Availability

An implementation of the best-practice workflow is available as bash script and as Galaxy workflow at `https://github.com/AnneHoffmann/tRNA-read-mapping`, respectively. For the workflow application to real data strand-specific small RNA-seq data from rRNA-depleted total RNA $>$ 200 nucleotides in size were obtained from the Encode project [324, 325]. Here the RNA-seq data of human esophagus muscularis mucosa tissue (female 51 years: GEO:GSE88169, female 53 years: GEO:GSE88236, male 37 years: GEO:GSE88128) were used.

CHAPTER 9

Specific Selection of tRNAs for RNA Sequencing

Contents

Currently available tRNA-seq methods such as YAMAT-seq [216] and tRNA-seq described in Pang et al. [217] do not sample the entire tRNA pool or lack specificity for tRNAs (see Section 5.2), since these are mainly designed for the detection of specific tRNA modifications. A disadvantage of such methods is that only full-length tRNAs are analyzed, while tRNA fragments or incomplete cDNAs due to reverse transcription stops at modified nucleosides (see Section 5.7) are lost. In this chapter the benchmark of a highly tRNA-specific RNA-seq method for an efficient and comprehensive analysis of tRNAs is demonstrated. We could point out that LOTTE-seq combines the benefits of existing methods and is able to handle various challenges arising for high-throughput analysis of the tRNAs (see Section 5.2 for the technical background).

The remainder of this chapter based on L. Eber and A. Hoffmann et al. [406] with the title *LOTTE-seq (Long hairpin oligonucleotide-based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing*. The laboratory implementation of LOTTE-seq has been carried out by L. Erber. Beside my contribution to the experimental design of LOTTE-seq, I mainly performed the data analyses described below.

9.1 LOTTE-seq Works for Species from All Domains of Life

In the LOTTE-seq protocol, a hairpin-shaped specific 3'-adapter is first ligated to mature tRNAs without previous purification. Thus, total RNA can be used without prior laborious and possibly bias-introducing tRNA enrichment, where usually a considerable amount of material is lost. A second adapter is ligated to the resulting cDNA 3'-ends. The adapter ligation to the cDNA 3'-end and not to the 5'-end leads to a considerable increase in sequence reads of the tRNA pool. Further, this specific cDNA 3'-end adapter ligation allows the amplification of both full-length as well as shorter cDNA fragments which results from read terminations (RTs) at nucleoside modifications (for more details see Erber et al. [406]). A schematic overview of the LOTTE-seq workflow is shown in **Fig. 28**.

To evaluate the performance of LOTTE-seq, the method has been conducted for representatives from each domain of life. In detail, LOTTE-seq has been performed for HEK293T cells (human), *Spinacia oleracea* (plant), *Saccharomyces cerevisiae* (fungi), *Dictyostelium discoideum* (Amoeba), *Escherichia coli* (Gram-negative bacteria), and *Geobacillus stearothermophilus*

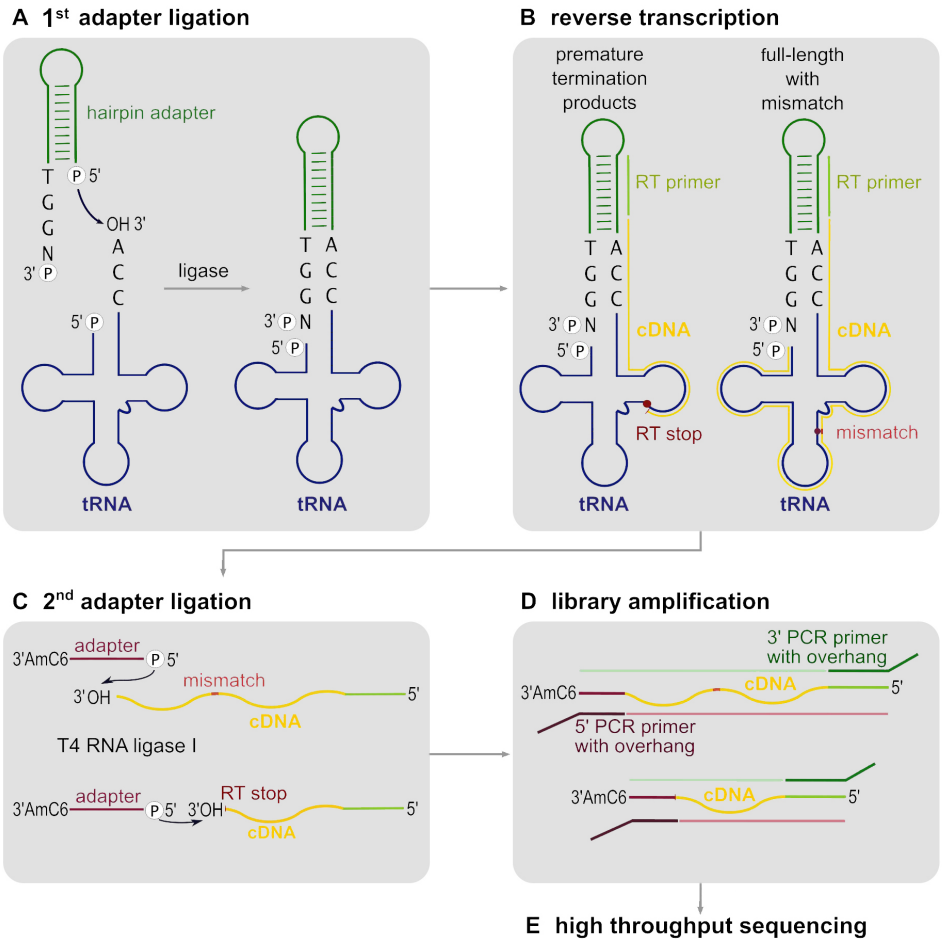

Figure 28: Schematic workflow of the LOTTE-seq procedure. (A) A DNA hairpin-oligonucleotide (green) with a 3'-TGGN overhang hybridizes to the complementary tRNA 3'-CCA end (tRNA in blue). T4 DNA ligase fuses the 3'-end of the CCA terminus to the phosphorylated 5'-end of the adapter. **(B)** The tRNA is reverse transcribed with parts of the hairpin oligonucleotide serving as primer binding site. Secondary structure and modified bases can lead to premature RTs and partial cDNA (yellow). **(C)** Using T4 RNA ligase I, a 5'-phosphorylated and 3'-blocked second adapter (red) is fused to the 3'-end of the cDNA, leading to the generation of cDNA product with adapters on both sides (red and green). **(D)** This product is amplified with indexed primers binding to the adapter overhang sequences. **(E)** The cDNA library consisting of full-length as well as prematurely terminated tRNA sequences is analyzed by high-throughput sequencing.

(Gram-positive bacteria) in two independent experiments. The samples were analyzed on an Illumina MiSeq device. Additionally, to compare LOTTE-seq with other RNA-seq methods, a standard sRNA TruSeq approach (5'- and 3'-adapter ligation followed by reverse transcription)

and an optimized sRNA TruSeq protocol (3'-adapter ligation followed by cDNA synthesis and subsequent cDNA adapter ligation) have been conducted for each of the six species (for details see E. Erber and A. Hoffmann et al. [406]). Annotation of tRNAs was performed as described in Section 7.1. For *S. cerevisiae*, only 16 of 24 mt-tRNAs could be annotated via `tRNAscan-SE`. Missing mt-tRNAs were added from the YeastMine database [326]. The number of annotated tRNAs for each species is listed in **Suppl. Tab. B3**. Our RNA-seq data analysis was prepared on the basis of the best-practice workflow for accurate mapping of tRNA reads [397] (see Chapter 8). Finally, the performance evaluation was implemented as outlined in Section 7.6.

In all investigations, LOTTE-seq shows the highest content of tRNAs in general with an average of 97%. In contrast, the average tRNA content in the optimized TruSeq sRNA method is 81%, while the average is only 6% in the standard Illumina TruSeq sRNA method. An overview of the comparisons is depicted in **Fig. 29** and the exact numbers are given in **Suppl. Tab. B4**. Further, LOTTE-seq specifically selects the highest amount of tRNAs with a 3'-CCA end in all six species (human: >55%; plant: >75%; fungi: >97%; Amoeba: >98%; Gram-negative bacteria: >98%; Gram-positive bacteria: >97%). In the optimized TruSeq sRNA procedure, the tRNA content with 3'-CCA end range from about 78% (Gram-negative bacteria) to less than 30% (human). The lowest amount of reads corresponding to tRNAs with 3'-CCA end is found in the standard Illumina TruSeq sRNA procedure with a range from $< 10\%$ (plant) to $< 1\%$ (human). In addition, we received a high amount of reads mapped to tRNAs as well as other genomic regions. Without exceptions, the ambiguous reads are very short in length and carry the 3'-CCA end of the tRNA which may result from read terminations during cDNA synthesis. In this case, the true origin of such reads cannot be determined unambiguously. The highest amount of ambiguous tRNA reads is found in both eukaryotic mulitcellular organisms (human: <38%; plant: <16%) using LOTTE-seq. In both unicellular eukaryotic organisms (fungi, amoeba) the ambiguous tRNA content is less than 2%, while in bacterial species the amount is close to zero.

In a further investigation we compared the tRNA content of our LOTTE-seq data to published tRNA-seq approaches described by Shigematsu et al. [216] (YAMAT-seq) and Pang et al. [217], see **Tab. 1**. A direct comparison to the data of Pang et al. [217] is not feasible, since neither the fraction of tRNA reads nor the fraction of tRNA reads with 3'-CCA end was

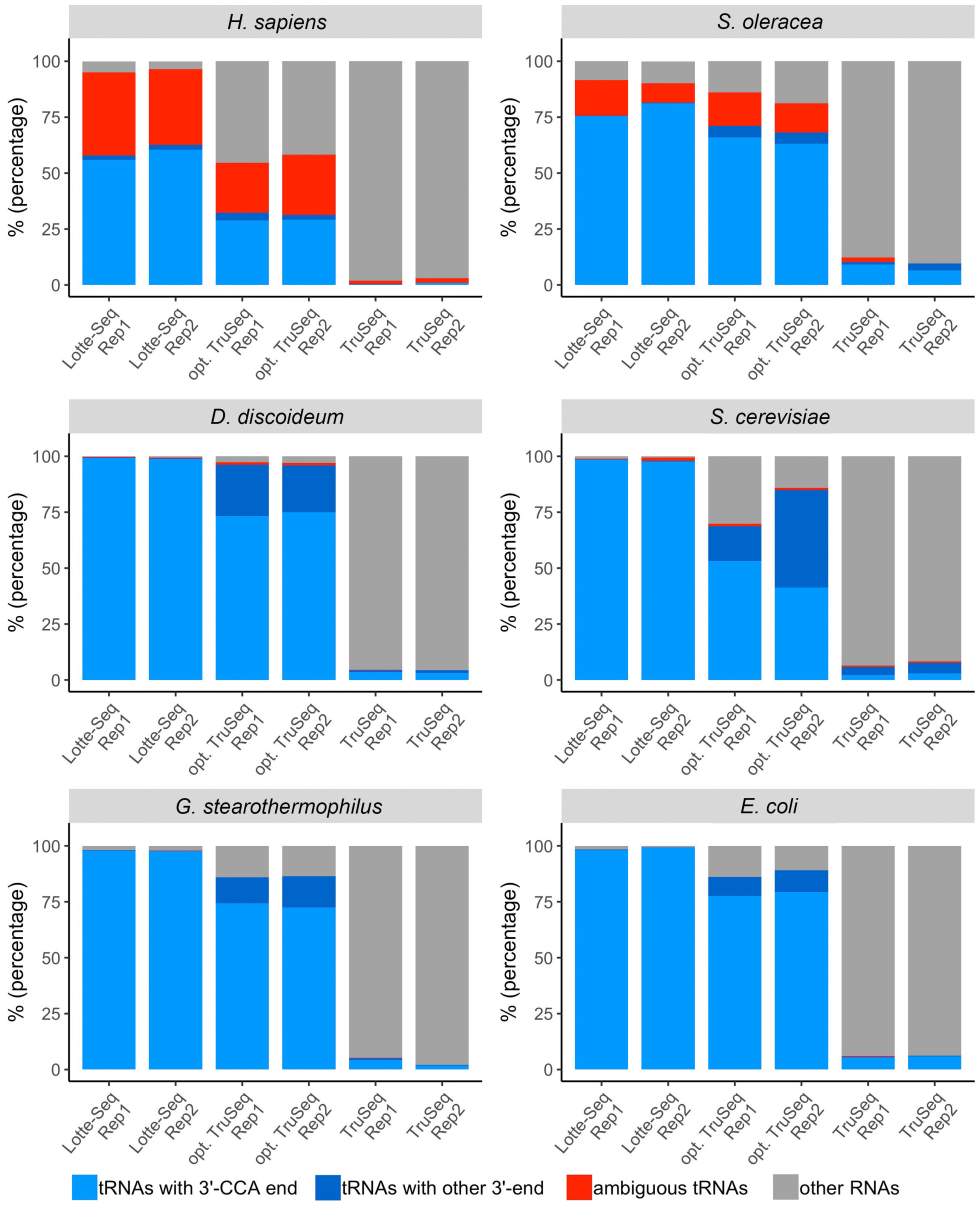

Figure 29: Comparison of LOTTE-seq to other RNA-seq methods. The tRNA content of LOTTE-seq compared to the optimized TruSeq sRNA protocol as well as to the standard Illumina TruSeq sRNA procedure is shown concerning tRNA content and 3'-CCA end. The percentage of tRNAs with a 3'-CCA end (light blue), with 3'-ends other than a CCA (dark blue) and non-tRNA reads (grey) are depicted for the individual organisms. The percentage of reads mapped to tRNAs as well as other genomic regions (ambiguous tRNAs) are highlighted in red. Two replicates of MiSeq-based sequence analyses of each species were investigated for each RNA-seq method. In all investigations, LOTTE-seq shows the highest content of tRNAs with CCA end.

reported. We, therefore, used sequencing data on *D. discoideum* and *G. stearothermophilus* that was generated by a procedure highly similar to the Pang approach [327]. Instead of HPLC separation, the tRNA-containing small RNA fraction was isolated by high salt precipitation as described in Cathala et al. [328] and Eichinger et al. [329]. The subsequent steps were identical to the Pang strategy. Compared to the published tRNA-seq methods, LOTTE-seq shows the highest amount of specific tRNA reads as well as tRNA reads carrying a 3'-CCA terminal end. While YAMAT-seq and LOTTE-seq show similarly high values for tRNA reads, adapter ligation by T4 DNA ligase is more selective for nick sealing in CCA sequence hybrids than truncated T4 RNA ligase used in YAMAT-seq. Due to the lack of CCA-specific adapter ligation, the Pang-like approach shows the lowest amount of tRNA reads and 3'-CCA end ligation, illustrating the importance of efficient separation of the tRNA fraction from other transcripts in the preparation procedure.

Taken together, LOTTE-seq is a highly robust and versatile approach that combines the pros of two - also very valuable - alternative procedures, while avoiding their cons. Combined with unique molecular identifiers, LOTTE-seq is a useful method to investigate the tRNA pools of different sources in a fast, convenient and reliable way.

Table 1: tRNA-specific reads in tRNA-seq methods. The average number of sequences mapped to tRNA genes and their proportion of reads carrying a 3'-CCA triplet sequence are shown for different tRNA-seq methods. Here, we compared our LOTTE-seq approach to YAMAT-seq [216] and a procedure closely related to Pang et al. [217]. For both criteria, LOTTE-seq shows the highest amount of specific tRNA sequences. Species abbreviations: *D. discoideum*: Ddi; *E. coli*: Eco; *G. stearothermophilus*: Gst; human HEK293T cells: Hsa; *S. cerevisiae*: Sce; *S. oleracea*: Sol.

tRNA-seq	% tRNA	% tRNA with 3'-CCA	Species
LOTTE-seq	97.0	99.4	Ddi, Eco, Gst Hsa Sce, Sol
YAMAT-seq	96.8	93.7	Hsa
Pang-like tRNA-seq	56.6	42	Ddi, Gst

9.2 Discussion

In recent years, the investigation of tRNAs or tRNA pools and their correlation to translation efficacy and regulation, stress conditions, and diseases developed into an important area of research [330–335]. There are many indications that tRNA abundance is associated with certain diseases [336–338]. However, the special features of tRNA molecules make library preparation quite complicated and error-prone, and standard Illumina approaches are not very practical for their analysis. This can be seen in **Fig. 29** where we compared LOTTE-seq with a standard sRNA TruSeq approach and an optimized sRNA TruSeq protocol. A reason for these difficulties is the high amount of modified bases [141, 339, 340] as well as stable secondary and tertiary structures of tRNAs [341, 342]. In comparison to both TruSeq sRNA procedures, LOTTE-seq shows the highest amount of tRNA reads and selects more specific tRNAs with a 3'-CCA end, including prematurely terminated cDNA fragments that represent the tRNA 3'-part. Reads from non-tRNA sequences were found at a very low abundance with an average value of 3%. tRNA sequences lacking the mature CCA end were found only in 0.6% of the all reads, indicating the high selectivity of our LOTTE-seq approach.

Compared to the procedures described by Shigematsu et al. [216] and Pang et al. [217] in terms of specificity, LOTTE-seq shows a selectivity for tRNAs similar to YAMAT-seq [216], see **Tab. 1**. However, the use of T4 DNA ligase leads to an increased specificity for complete CCA ends (only 0.6% non-CCA ends), while T4 RNA ligase 2 that was used in YAMAT-seq also accepts unpaired single-stranded 3'-ends, leading to 6.3% non-CCA ends. A direct comparison of the tRNA pool composition identified by YAMAT-seq and LOTTE-seq, however, is not reasonable, as Shigematsu et al. [216] used breast cancer cell lines (BT-474, SK-BR-3, MCF-7) in their analysis, while we used human embryonic kidney cells (HEK293T). There is growing evidence that the cellular tRNA pool composition is not stable, but is actively adjusted to individual growing conditions or cell type requirements, resulting in specific tRNA pools in different cells or organs [330, 331, 338, 343]. As a result, these cell-type-specific differences render a direct comparison of the data obtained by YAMAT and LOTTE-seq impossible. Thus, a direct comparison to the data of Pang et al. [217] is not feasible, since no tRNA fraction information has been reported. We, therefore, compared our data to tRNA-seq data which are highly similar to that approach [327]. While the Pang-like tRNA-seq procedure also led to a

considerable amount of tRNA reads (56.5%), the number of non-tRNA reads is much higher compared to YAMAT or LOTTE-seq. Furthermore, the use of T4 RNA ligase in the Pang-like tRNA-seq procedure leads in to 42% of sequences ending with sequences other than CCA. This is a further indication that CCA-specific 3'-adapters are highly selective. A combination of the CCA-specific 3'-adapters with T4 DNA ligation reaction in LOTTE-seq result in the highest number of reads with mature tRNA 3'-ends.

When analyzing our samples, we were faced with the problem of accurate mapping of tRNA reads as described by A. Hoffmann et al. [397]. Especially multicellular eukaryotic organisms show a high amount of tRNA genes, i.g., we annotated 732 and 2111 tRNAs in human and spinach, respectively. Their isodecoders only differ in a few nucleotides [12, 301, 344–346]. This complicates the allocation of reads to the corresponding gene. To this end, we applied the best-practice workflow for the accurate mapping of tRNA reads as discussed in Chapter 8. Since the specific cDNA 3'-end adapter ligation allows the amplification of both full-length as well as shorter cDNA fragments, which is caused by RTs at nucleoside modifications, a high amount of very short reads is available in the tRNA samples. When performing LOTTE-seq for different organisms, we found that the amount of full-length tRNA reads differed dramatically between species. Higher amounts were obtained for bacterial samples. This might be due to a less complex pattern of base modifications in these organisms [27, 347]. In contrast, the relative amount of full-length tRNA was smaller in human and plant samples, where tRNAs are usually modified to a greater extent. In higher organisms in particular, short reads map to genomic regions in addition to tRNAs, since a large number of tRNA-like structures occur, e.g., tRNA-derived fragments (tRFs) [65]. Determining the true origin of these short reads is an error-prone task. Therefore, we considered these ambiguous reads separately, as it cannot be excluded that they do not originate from tRNAs.

9.3 Data Sources and Availability

Genomes of *D. discoideum* (assembly dicty 2.7), *E. coli* (strain K-12 substr. MG1655), *G. stearothermophilus* (strain ATCC 12980), *H. sapiens* (assembly hg38), *S. oleracea* (assembly KY768855.1), and *S. cerevisiae* (strain BY4741) were downloaded from NCBI, release 90 [22]. All investigated RNA-seq data are available at NCBI BioProject: PRJNA541863.

CHAPTER 10

Detection of Chemical tRNA Modifications

Contents

Chemically modified nucleotides are ubiquitous in RNA and DNA sequences. They are incorporated post-transcriptionally at the nucleobase and/or the sugar unit (see Section 3.5). The intensive analysis of high-throughput experiments has led to a rapid increase of the knowledge of modifications in RNAs in particular. Transfer RNAs (tRNAs) are a hotspot with modifications contributing to the formation of the functional three-dimensional structure. Nevertheless, little systematic attention has been focused on the research of chemically modified residues in tRNAs. Some modified nucleotides leave specific signatures in RNA sequencing (RNA-seq) data which originated from the reverse transcriptase during complementary DNA (cDNA) synthesis (see Section 5.7). In this chapter we introduce a combination of specialized deep sequencing approaches and sophisticated bioinformatic methods enabling sensitive and precise detection of signatures generated by chemically modified nucleotides in tRNAs.

For an accurate detection and quantification of modified tRNA residues at the transcriptome level using RNA-seq data the reads have to be first precisely mapped to the reference genome. For this purpose we applied our newly developed best-practice workflow for accurate mapping of the short tRNA reads prior to each analysis described in this chapter. The mapping workflow is discussed in Chapter 8 and the tRNA annotation method is explained in Section 7.1.

In the remainder of this chapter Section 10.1.1 is based on A. Hoffmann et al. [397] with the title *Accurate Mapping of tRNA Reads*. Section 10.2 refers to the publication of L. Erber and A. Hoffmann et al. [406] titled *LOTTE-seq (Long hairpin oligonucleotide-based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing*. Section 10.2.1 is based on the publications of A. Hoffmann and L. Erber et al. [409] titled *Changes of the tRNA modification pattern during the development of Dictyostelium discoideum* and Erber et al. [408] titled *Dictyostelium discoideum: Unusual occurrence of two active CCA-adding enzymes*. The performance and results from the publication A. Hoffmann et al. [410] with the title *Temperature Dependence of Bacterial tRNA Modifications* are pointed out in Section 10.3.

10.1 Detecting tRNA Modifications by Base Misincorporations

In the simplest case tRNA modifications affect reverse transcription during cDNA synthesis leading to a visible position-specific increase of misincorporations in RNA-seq data (see Sec-

tion 5.7). Our first starting point when calling tRNA modifications via base-calling errors was to find a suitable modification caller. For this purpose we evaluated three different tools, namely `bcftools` [267, 268], `GATK` [269], and `HAMR` [261]. The technical background of these tools is discussed in Section 5.7 and their implementation is described in Section 7.3.1. Both `bcftools` and `GATK` are well established for germline short variant discovery from whole genome and exome sequencing data. We expected that the tools could also be used for our purpose to identify genomic positions with increased levels of sequence mismatches. `HAMR`, on the other side, is directly designed for RNA modification discovery. For performance evaluation (described in Section 7.6) of the three different tools we used ribo-minus RNA sequencing (rmRNA-seq) data (see Section 5.1) generated from human cerebellum. In short, all detected modification sites were visually examined to exclude possible hits calling due to mapping artifacts or wrongly interpreted reverse transcriptase signatures. True candidate sites were characterized according to known human tRNA modifications stored in the tRNAmodviz database [119]. Since only 26 of 754 human tRNA sequences are included in tRNAmodviz and the databases derive their modification information from different tRNA-seq experiments performed under different conditions, a tRNA-specific validation for all human tRNAs is not possible. Thus, we generally characterized called modifications as true sites if they overlap with known modified positions of the tRNAmodviz tRNA reference set (position-specific validation).

Surprisingly, our performance evaluation revealed that the tools produced very different results. The total numbers of called true positive (TP) and false positive (FP) sites for each tool are listed in **Tab. 2**. Not only the number of called modifications sites but also the amount of modified tRNAs varies greatly between the applied tools. With `GATK` we found 428 candidate modifications at 14 different tRNA positions and in 294 tRNAs. Among these 14 positions, 11 coincided perfectly with a known modified position as listed in the tRNAmodviz database. In contrast, we detected only 373 candidate modification sites in 276 different tRNAs using `bcftools`. These modification sites relate to 12 tRNA positions, 9 of which match the database entries. Using `HAMR`, we were able to detect only 110 candidate modification sites in 106 different tRNAs and at 6 tRNA positions. Out of these 6 positions 5 can be assigned to known ones.

Table 2: Comparison of three tools used for tRNA modification discovery. Overview of true transfer RNA (tRNA) modifications (black) and false positive called sites (red) for each of the three tools `GATK`, `bcftools`, and `HAMR` are shown. For each tool, the number of tRNAs that display a significant base misincorporation rate at the affected positions are given. Known human tRNA modifications listed in the tRNAmodviz database are assigned by position. Callings were performed on ribo-minus RNA sequencing data generated from human cerebellum tissue. The number of discovered tRNA modifications varies greatly between the three tools. `GATK` is the only tool that calls false positive sites (15%). `HAMR` finds very few candidate sites. Most suitable for tRNA modification calling seems to be `bcftools` which is moderately less sensitive than `GATK` but does not detect false positive sites. Abbreviations of the first column: AC – anticodon; ACC – acceptor; D – dihydrouridine; T – TΨC; V – variable. Nucleobase abbreviations: A – adenine; C – cytosine; G – guanine; T – thymine.

Area	Position	Alteration	GATK	bcftools	HAMR	Modification
5'-ACC-stem	1	G→T	8	0	0	-
	6	T→A	2	0	0	-
	6	G→C	12	0	0	m^2G
	7	G→T	1	0	0	-
-	9	A→(C|G|T)	19	10	18	m^1A
		G→T	35	19	5	m^1G
5'-D-stem	10	G→T	2	0	0	-
	12	A→C	1	0	0	-
D-loop	16	A→G	1	0	0	-
	20	T→G	1	0	0	D
3'-D-stem	23	A→C	11	6	0	unknown
-	26	G→T	31	48	22	m^2G, $m^2{}_2G$
5'-AC-stem	31	A→G	10	10	0	unknown
AC-loop	32	C→T	6	6	1	Cm, m^3C
	34	A→G	38	38	0	I
	37	A→(G|T)	21	23	0	t_6A, i^6A, m^1I
		G→T	32	12	0	m^1G, o_2yW
V-region	2e	C→(A|T)	4	4	0	m^3C
5'-T-stem	49	A→T	1	1	1	unknown
T-loop	56	C→T	1	0	0	-

Continued on next page

Table 2 – continued from previous page

Area	Position	Alteration	GATK	bcftools	HAMR	Modification
	57	A→T	20	0	0	-
	58	A→(G|T)	207	196	63	m^1A
	60	C→T	19	0	0	-
3'-T-stem	64	G→A	1	0	0	-
		T→A	1	0	0	-
	65	C→A	1	0	0	-
		G→C	1	0	0	-
3'-ACC-stem	66	C→A	1	0	0	-
	68	G→C	1	0	0	-
	69	C→T	1	0	0	-
		T→A	1	0	0	-

The overlap of called modification sites also varies considerably between the tools as depicted in **Fig. 30**. Only 72 candidate sites, which match the tRNAmodviz database entries, were found by all three tools. In addition, GATK and bcftools overlap at 231 modification sites with the tRNAmodviz database, while their overlap with HAMR is very small due to the small number of TPs called by HAMR. However, the numbers of modification sites found by only one tool are very high. In detail, GATK and bcftools called 76 and 45 sites, respectively, which overlap to the modified positions of the database entries but not with the results of the other tool. This suggests that none of the three tools is sensitive enough to find all TPs. A direct comparison of the modification patterns of the 26 sequences stored in the tRNAmodviz database with the results of the three tools confirmed this assumption (tRNA-specific validation; see **Fig. 30**). 18 of 41 modified sites detectable by accumulations of base misincorporations for the tRNAmodviz tRNA reference set cannot be found with any applied tools. It is possible that these modifications do not occur in our dataset given that the database is based on RNA-seq experiments created under different conditions.

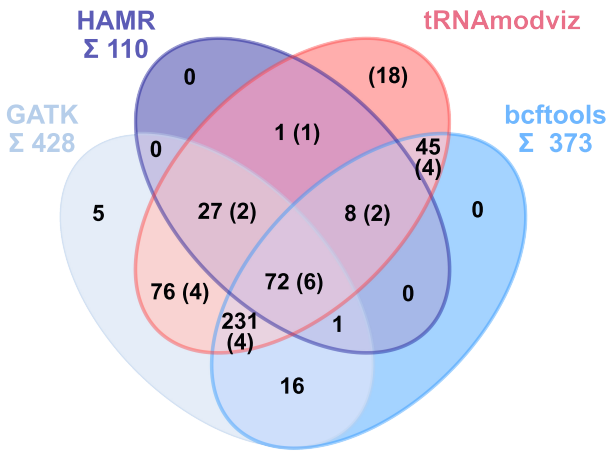

Figure 30: Overlap of called and known modification sites. Venn diagram illustrating the overlap between called candidate modification sites in human cerebellum obtained from the tools `GATK` (lightblue), `bcftools` (blue) and `HAMR` (darkblue) with known modifications of human transfer RNAs (tRNAs) stored in the tRNAmodviz database (red). True candidate sites were called based on accumulations of sequencing errors produced by the modified nucleotide during complementary DNA synthesis. An overlap to the tRNAmodviz database was counted (numbers without brackets) if the called true modification site is known in at least one human tRNA of the database (position-specific validation). The overlap of called true sites between the three tools is very low (# 72). Most similar in their modification calling specifications are `GATK` and `bcftools`, as they have an overlap of 231 true sites. Numbers in brackets show the overlap of the called true modification sites of the 26 tRNA sequences which serve as a reference set for human tRNA modifications in the tRNAmodviz database (tRNA-specific validation). These 26 tRNAs show 66 modified sites which are detectable by accumulations of base-misincorporations in RNA sequencing data. Only 41 of the 66 modified sites were included in the counting, as we received coverage of more than 10 reads only for these sites. `GATK` and `bcftools` achieved the greatest overlap (# 16) with these 41 sites.

In our analysis, `GATK` already identifies candidate sites with a base misincorporation rate of $>$ 10%, while the smallest rate of `bcftools` is 18% (see **Fig. 31**). We are not able to reduce this minimum prediction of the base misincorporation rate of `bcftools` by parameter adjustments. Since tRNAs can also be weakly modified [119], a higher false negative rate (FNR) results for `bcftools` than for `GATK`. However, the low base misincorporation rate predicted by `GATK` leads to increased detection of FPs (see **Tab. 2**). The tool is vulnerable to call mapping artifacts incorrectly as true candidate sites. In total, these FP hits amount to 15%. Read coverage does not seem to have a big influence on the sensitivity of `GATK` and `bcftools`, since true candidate sites with a low read coverage were also recognized. It is still unclear why sites showing high modification rates and read coverage are not always found by

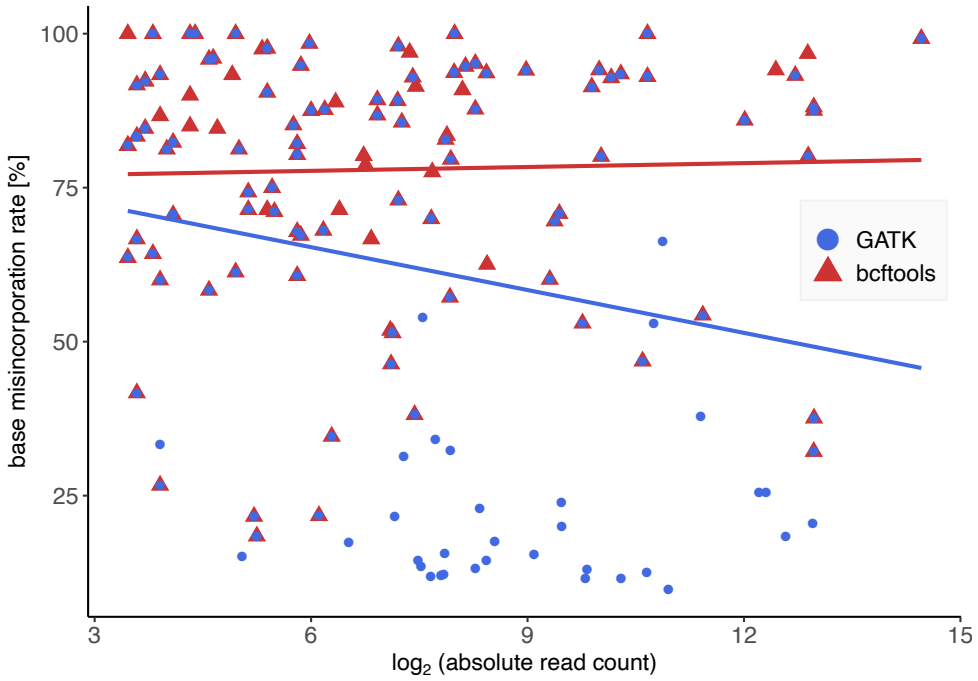

Figure 31: Performance of `GATK` and `bcftools`. Each true candidate modification site called by `GATK` (blue circles) or `bcftools` (red triangles) is illustrated regarding base misincorporation rate and read coverage. Overlapping symbols indicate that these candidate sites were detected by both tools. In comparison to `bedtools`, `GATK` also finds true candidate sites with a base misincorporation rate of less than 18%. Read coverage does not show a strong effect on the sensitivity of both tools, since even true candidate sites with low coverage were detected.

both tools. In summary, even though `GATK` called the most true candidate modification sites, it is the only tool that additionally called FPs. Since `HAMR` is not as sensitive as the other two tools, it is also not suitable for our analyses. `bcftools` is only slightly less sensitive and does not call any FPs compared to `GATK`, so we decided to use `bcftools` for our further analysis.

10.1.1 Tissue-specific Modifications in Human tRNAs

A recent study based on liquid chromatography-mass spectrometry quantification demonstrated that the relative abundance of nucleotide modifications in tRNAs varies substantially between different tissues in mouse and pig [348]. Thus, we were interested in determining whether such differences are also detectable in rmRNA-seq data of six human tissues (cerebellum,

diencephalon, ovary, skeletal muscle, esophagus muscularis mucosae, and testis) and if so, whether there are also qualitative differences in the sense that different locations are modified.

The answer to both questions is affirmative. We observe seven positions (9, 26, 32, 34, 37, 49 and 58) which are frequently modified in all six tissues. In contrast 15 positions are modified in a tissue-specific manner, 7 of which being present in more than one tissue. Candidate sites of ten modified positions coincided perfectly with known modifications which is illustrated in **Fig. 32A** and listed in detail in **Suppl. Tab. B5**. In all six tissues we detected the well known 1-methyladenosine (m^1A) modification at positions 9 and 58 indicated by an A-to-(C|G|T) substitution. At the anticodon adjacent position 37 we noticed a further mismatch pattern of adenines (As) which can be assigned to 1-methylinosine (m^1I), N^6-threonylcarbamoyladenosine (t_6A), and N^6-isopentenyladenosine (i^6A). In almost all reads modified tRNA sites at position 34 of all six tissues show a A-to-G transition which is typically for inosine (I). Additionally, we observed increased error rates in all tissues for the methylguanosine modifications 1-methylguanosine (m^1G) at position 9 as well as N^2-methylguanosine (m^2G) and N^2,N^2-dimethylguanosine ($m^2{}_2G$) at position 26. The occurrence of other observed guanosine modifications vary between tissues. For example, m^2G at position 6 is only present in ovary, testis, and esophagus muscularis mucosae tissues. Whereas m^1G and peroxywybutosine (o_2yW) is not present in diencephalon and skeletal muscle. We recognize the methylguanosine modifications by a G-to-C transversion or a G-to-T transition. Modified methylcytosines show specific mismatch patterns of 2'-O-methylcytidine (Cm) and 3-methylcytidine (m^3C) modifications at position 32 in each tissue. Cm is only modified at position 61 in diencephalon, whereas m^3C at position 2e is modified in all tissues except for testis.

However, the number of modified single tRNA genes varies according to human tissue (see **Fig. 32B**). Only 91 tRNAs are modified in each tissue, 38 of which have the same modification pattern. The highest amount of 155 identically modified tRNAs is observed for the tissues cerebellum and esophagus muscularis mucosae. In comparison, diencephalon and testis contain only 55 identically modified tRNAs. An example of a tissue-specific modification pattern is given in **Fig. 32C**. At this example, the alanine $tRNA^{Ala}_{AGC}$ is modified at positions 34 and 37 in each tissue, while positions 26 and 58 are not. These data strengthen the assumption that tRNAs are modified in a tissue-specific manner.

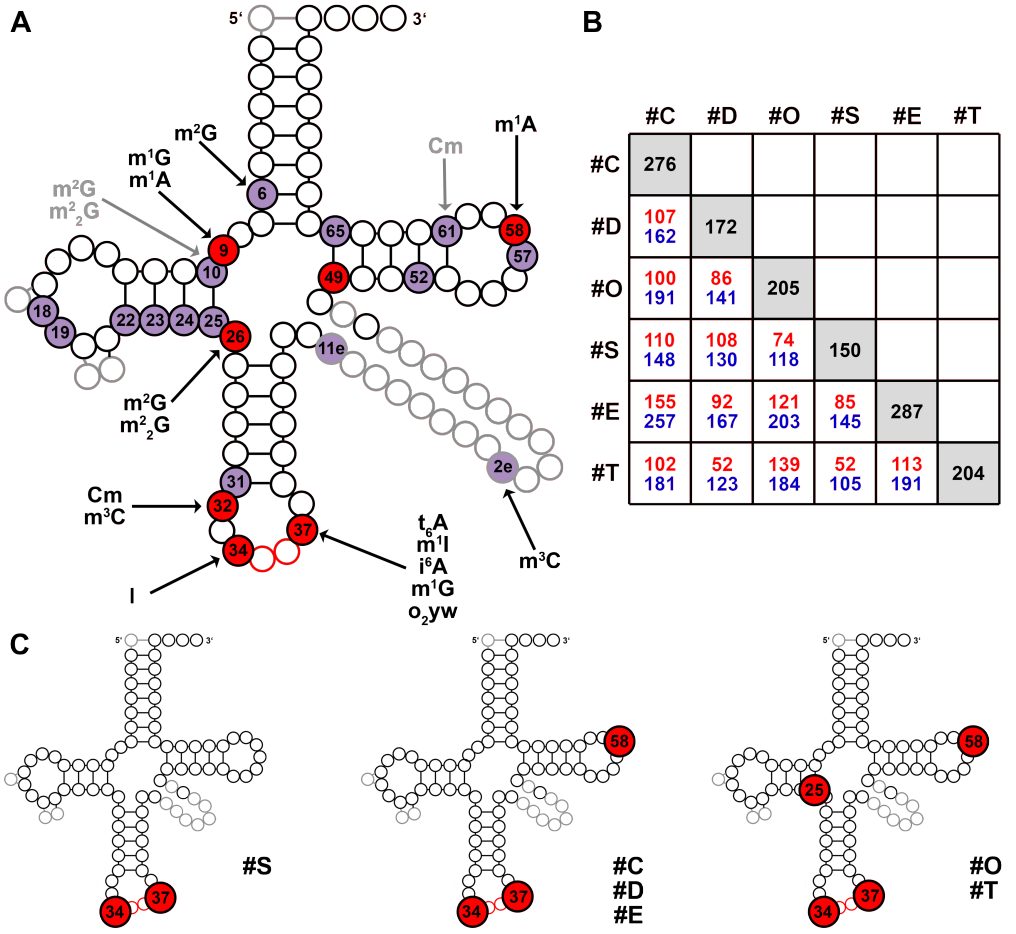

Figure 32: Summary of tissue-specific modification patterns in human tRNAs. (A) Overview of transfer RNA (tRNA) positions which are modified in each of the investigated tissues (red) or in at least one tissue (purple). The six investigated human tissues are cerebellum (C), diencephalon (D), ovary (O), skeletal muscle (S), esophagus muscularis mucosae (E), and testis (T). Known modifications of human (black) or other species (gray) listed in the tRNAmodviz database are assigned to the modified positions. In total, seven positions are modified in each tissue and 15 positions are modified in at least one tissue. **(B)** Overlapping numbers of modified tRNAs per human tissue. The matrix displays numbers of identical modified tRNAs (red) and tRNAs which are modified in different manners (blue) between tissues. In each tissue the same 91 tRNAs are modified, only 38 of which are identically modified. This strengthens the assumption that tRNAs are modified in a tissue-specific manner. An example of a tissue-specific modification pattern is given in **(C)**. Here, the pattern of an alanine tRNA ($tRNA^{Ala}_{AGC}$) is shown. In each tissue, the position 34 and 37 in the anticodon arm are modified. Position 58 in TΨC-arm is modified in each tissue except for skeletal muscle. Position 25 is modified only in the gonads (testis, ovary).

10.2 Read Terminations Provide Indications for Modification Profiling

Since base modifications can cause read terminations (RTs) during cDNA synthesis, it is highly likely that accumulations of apparent RTs represent such base modifications. A high sequencing depth is required to be able to determine tRNA modification from accumulations of RTs. On average, in the rmRNA-seq data used in Section 10.1, we could only assign 0.95% of all reads to tRNAs leading in a low tRNA read coverage. For a sufficiently high read coverage it is necessary to use tRNA-specific RNA-seq data. Thus, we compared reverse transcription patterns of human HEK293T cells from long hairpin oligonucleotide-based tRNA high-throughput sequencing (LOTTE-seq) and optimized sRNA TruSeq data (see Chapter 9).

Regarding the fraction of RTs over all tRNA reads in both tRNA-specific RNA-seq methods, we obtained strong peaks (> 0.10) at tRNA positions 9, 20a, 26, 32, 34, 37, 2e, 4e, and 58 (see **Fig. 33**). However, the patterns differ in their intensity. Except for tRNA position 20a LOTTE-seq always finds the strongest peaks. Differences in the intensity of RT fractions can be explained by the number of tRNA reads specifically selected by both methods. On average over the two human replicates, LOTTE-seq finds 95.8% tRNA reads, while sRNA optimized TruSeq only contains 56.4% tRNA reads (see **Suppl. Tab. B4**). However, both tRNA-specific methods are suitable for the identification of tRNA modifications by RTs, as both display apparent RT patterns.

Several experimental approaches exist that use an induced RT stops to specifically investigate the presence of individual types of modifications. For example, ARM-seq [242] as well as DM-seq [265] compare untreated with enzymatically demethylated samples to identify certain base methylations in transcriptome data (see Section 5.7). With these methods they are able to assign RT accumulation to m^1A (positions 9 and 58), m^1G (positions 9, 37), $m^2{}_2G$ (position 26), i^6A (position 37), and to m^3C (position 34) modifications. Our observed RT patterns perfectly correlate with the outcome of these studies. Additionally, we found significant base misinorporations for the apparent RT accumulations at distinct positions applying `bcftools` (see Section 7.3.1). Allocation of mismatch patterns to specific modifications (see **Suppl. Tab. B2**) is consistent with the results of the RT signals (see **Fig. 33**). At position 2e we further received reverse transcription signatures which can be assigned to m^3C.

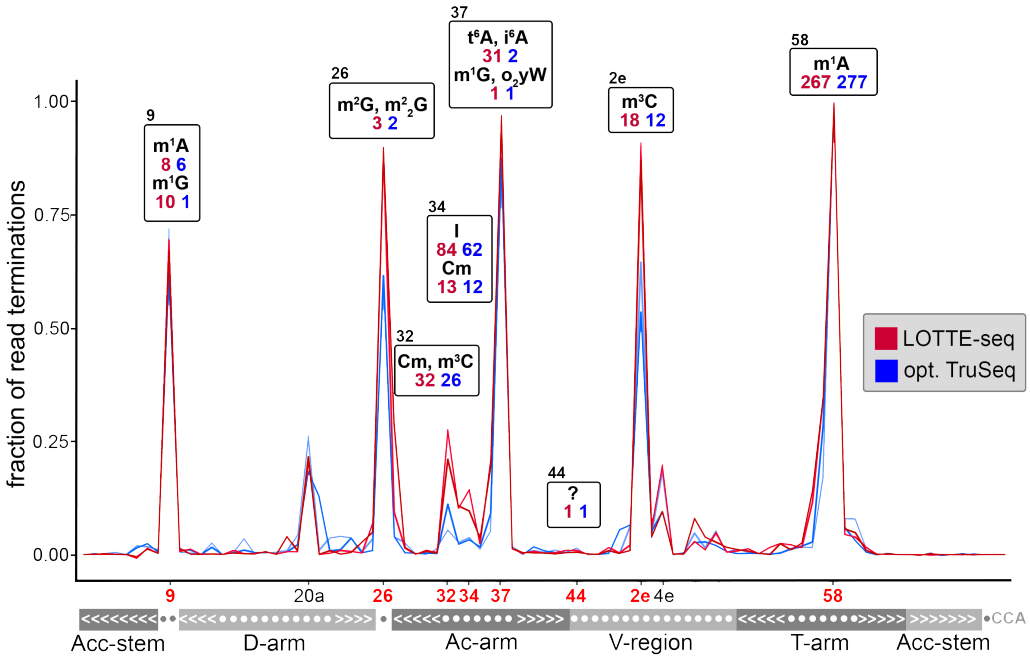

Figure 33: Fraction of read terminations identified in LOTTE-seq and optimized TruSeq data. Fraction of read terminations (RTs) over all transfer RNAs (tRNAs; y-axis) are shown for each tRNA position (x-axis). RT fractions are calculated from human LOTTE-seq (red) and optimized TruSeq (blue) data. The secondary structure of tRNAs is given in dot-bracket notation (bottom). Only positions of tRNAs that display peaks (> 0.10) are denoted. Despite tRNA positions 20a and 4e we observed a position-specific increase in the rate of sequencing errors for each peak. These positions are highlighted in red at the x-axis. Modifications that can be classified based on the individual mismatch pattern (see **Suppl. Tab. B2**) and the tRNAmodviz database [119] are specified as well as the number of tRNAs displaying this modification in each RNA-seq method. Positions 20a and 4e are represented by highly conserved uridine residues. Here we assume dihydrouridine modifications as they are not visible as conspicuous accumulations of mismatches but show high fractions of RTs. At position 44 we observed that adenine-to-guanine transitions exhibited RTs, but no modification is known at this position for human tRNAs. Both methods produced modification-specific RT signatures demonstrating that tRNA-enrichted RNA-seq data can be used for the identification of certain base modifications. However, the highest fractions of RTs are observed for Lotte-seq data (positions 26, 32, 34, and 2e), while the remaining pattern is quite similar to the optimized TruSeq method.

We found accumulation of RT fractions where we could not observe any mismatch patterns. Since RTs at positions 4e and 20a arise exclusively from uridines, a possible dihydrouridine modification can be assumed at this position. Usually, dihydrouridine (D) modifications cannot be identified by mismatch pattern rather than by RT signals or specific chemical treatments of the sequencing library. We recommend the use of chemical treatments for detection of D

modifications, since the signal we received is very weak and other well known modified sites of D (16, 17, 20, and 47) do not show strong RT signals.

The classification of reverse transcription signals requires prior knowledge. For example, we observed adenine-to-guanine transitions at position 44. However, there is no strong peak at this position indicating that the modification is either very weak or does not cause RTs. Since adenine alteration can show different types of modifications and no appropriate modification is published in human tRNAs at this position, an exact classification is practically impossible. It is reported that Cm and I modifications do not produce RTs during cDNA synthesis. Nevertheless, we observed a strong peak at position 34, where only Cm and I match the base misincorporation pattern. Still, it is unclear whether there is any other kind of modification at this position that is not listed in the tRNAmodviz database, or whether any other influences generate RTs at the position, e.g., secondary structure peculiarities.

10.2.1 tRNA Modifications Vary During the *D. discoideum* Life Cycle

A recent study demonstrates that gene expression of several methylated tRNA genes differ significantly in *Oryza sativa* (rice) and *Arabidopsis thaliana* (thale cress) between different stages of development [349]. Therefore, we are interested in whether variations in the intensities of RTs for specific modifications occur at different developmental stages using tRNA-enriched RNA-seq data. Since tRNA modifications for *Dictyostelium discoideum* have not been investigated yet, we used this well-studied model organism for our analysis.

D. discoideum is a slime mold that belongs to amoeba. When starving (0 hours), the slime mold is able to undergo a complex development to produce differentiated cells of spores and stalks depicted in **Suppl. Fig. A7**. Simultaneously, the organism changes from an unicellular to a multicellular organism. Development of *D. discoideum* is characterized by a highly regulated program of altered protein expression leading to complex formation of a new organism [350]. When the cells run out of sufficient nutrients, cyclic AMP (cAMP) signal release initiates a stream of surrounding cells to a central domain. Streaming leads to a multicellular aggregation (6 hours after starvation) and continues until a multicellular organism with the shape of a mound is formed (16 hours after starvation). A migrating slug then forms, which can move through the soil. When the culminant is formed (20 hours after starvation), cells begin to

differentiate into pre-stalk and pre-spore cells. Eventually, the fruiting body consists of stalk cells, which are intended to die, and spore cells, which are released under optimal growth conditions. A new unicellular amoeba then develops again through the spore cells (24 hours after starvation) [350, 351].

Cells of five different developmental stages (0 hours, 6 hours, 16 hours, 20 hours and 24 hours after starvation; see **Suppl. Fig. A7**) of *D. discoideum* were prepared [408] following the LOTTE-seq protocol (see Chapter 9). After read mapping, we obtained a high amount of tRNA reads in the samples (on average 98%) using LOTTE-seq data.

In order to gain an insight into the modification pattern of *D. discoideum* during its development, we identified candidate modification sites by accumulations of base misincorporations (see **Suppl. Tab. B6**) and apparent RT signals (see **Fig. 34**). We classified the detected modification sites according to our collection of modification-specific reverse transcription signals (see **Suppl. Tab. B2**). In addition, we used modification information of any species stored in the tRNAmodviz database for the classification of the tRNA modifications, since no data are available regarding tRNA modifications in the slime mold. In each investigated developmental stage, the same nine tRNA positions (9, 20, 26, 32, 34, 37, 47, 58, and 68) show base misincorporations as illustrated in **Fig. 35**. Despite positions 32 and 34, strong peaks ($>$0.10) of RT fractions occur. Base alterations at positions 9, 34, and 58 can clearly be assigned to m^1G, I, and m^1A, respectively, as no other tRNA modifications are known for these modified nucleotides. No unambiguous classification is possible for the other tRNA modifications, as several modifications are known in other species for the same altered nucleotides. Modified cytosines at tRNA positions 20, 32, 34 may be either m^3C or Cm displaying the same reverse transcription profile. The same applies to modified guanine residues which could probably be m^2G and $m^2{}_2G$ modifications (position 26) or m^1G and o_2yW (position 37). It is also indistinguishable whether modified adenines (A37) refer to t_6A, i^6A, or m^1I modifications. Since only RTs are reported for i^6A [352] and we observed a strong signal of RTs at position A37, tRNA genes showing RTs can be classified as i^6A. This can also be assigned to the identified adenine-to-guanine alterations at position 68 were no modification is reported in the tRNAmodviz database. Base alterations showing strong RT accumulations can be classified to i^6A or m^1A, while tRNA gene mapping without RTs could be an indication of I, m_1I, or t_6A. Typically, modified uridines cannot be detected by analyzing base misincorporation

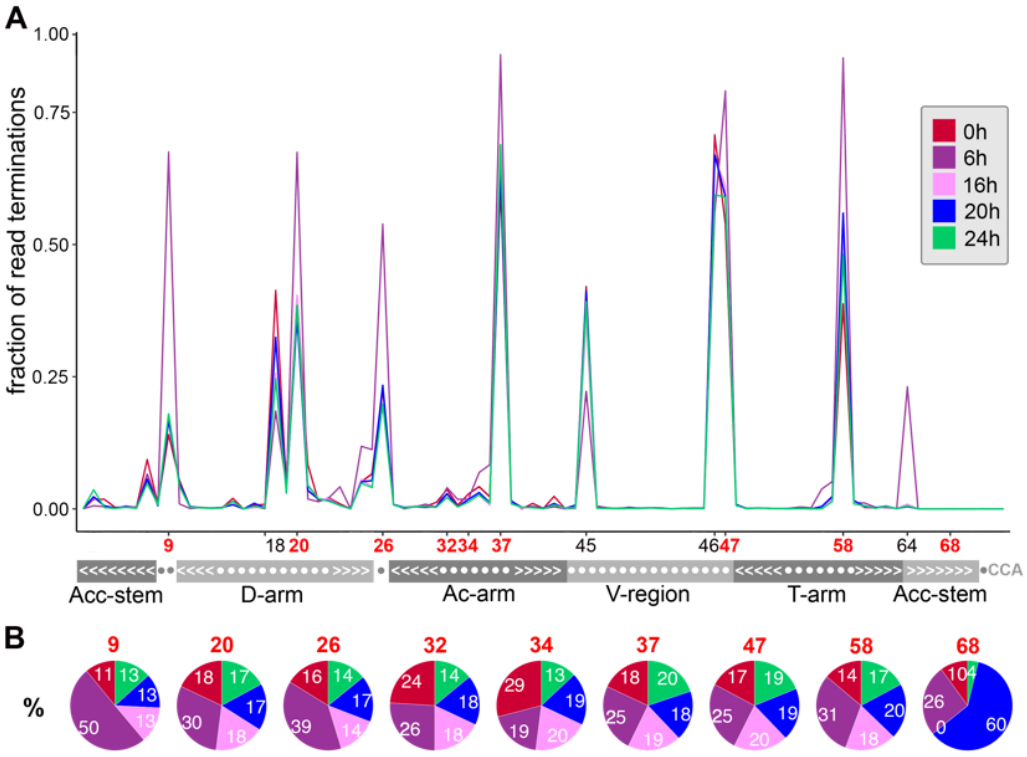

Figure 34: Comparison of read termination fractions during the life cycle of *D. discoideum*. **(A)** Fraction of read terminations (RTs) over all transfer RNAs (tRNAs) (y-axis) are shown for each tRNA position (x-axis). RT fractions are calculated for different developmental stages (0h, 6h, 16h, 20h, and 24h) of the life cycle of *D. discoideum*. Positions of tRNAs that display peaks (> 0.10) are highlighted in black, and peaks that also show base-calling errors are highlighted in red on the x-axis. The secondary structure of tRNAs is given in dot-bracket notation (bottom). **(B)** Relative fraction of RTs for each developmental stage is given in percentages. The color-coding is according to **(A)**. Generally, most RTs occur at time of 6h. At position 34, the highest relative proportion of RTs is observed from the starting point of developmental morphogenesis (0h). At 20h, the highest number of RTs (60%) is detected at tRNA position 68 while for 16h no RTs can be found. However, as only one tRNA is modified at each developmental stage at this position, the amount of RTs is very small and therefore no peak can be seen. Variations in the life cycle are a result from the strengths of RTs. This indicates different modification levels at individual positions for the investigated point in time.

sites, but by RT accumulations. However, we observed thymine-to-cytosine transitions and thymine-to-adenine transversions at positions 20 and 47. We assume that such base changes indicate D modifications, as uridines at position 20 are usually modified as D in several species. This can be confirmed as we achieved strong peaks of RT fractions at both tRNA positions

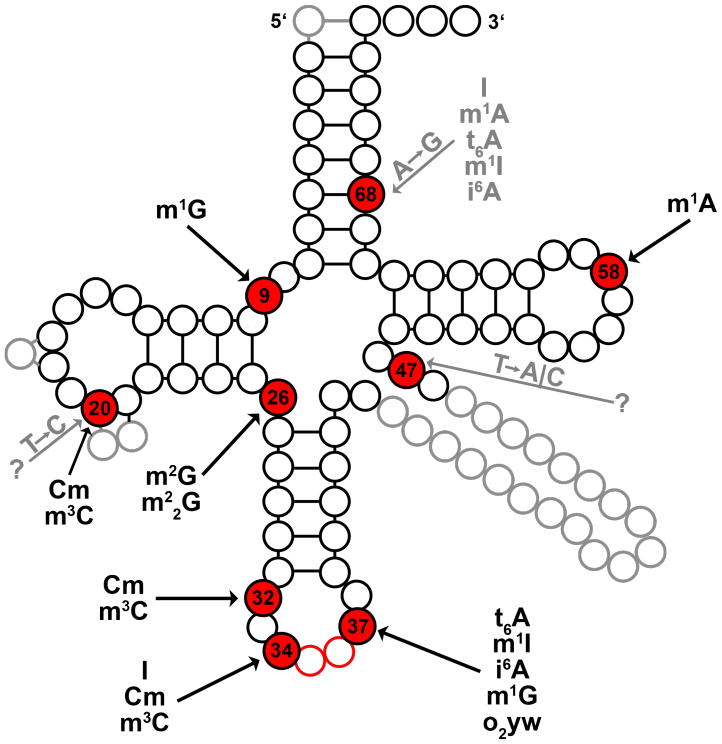

Figure 35: Modification pattern of *Dictyostelium discoideum*. At each investigated developmental stage (0h, 6h, 16h, 20h, and 24h after starvation) of the life cycle of *D. discoideum* the same transfer RNA (tRNA) positions are modified (red circles). Modifications are detected by base-calling errors. Since tRNA modifications have not been investiged yet for *D. discoideum*, known modifications (black arrows) of other species from the tRNAmodviz database [119] served as indicators for modification classification. Typically, uridine modifications cannot be detected by the analysis of base-calling errors. However, we observed thymine-to-cytosine transitions or thymine-to-adenine transversions at positions 20 and 47 (gray arrows). We assumed that such base changes indicate D modifications since uridines at position 20 are usually modified to D. At position 68 no tRNA modification is known in other species (gray arrow). The observed mismatch pattern (adenine-to-guanine) may be an indication of a possible I, m^1A, t_6A, m^1I, or i^6A modification at this position. Abbreviations: A – adenine; C – cytosine; Cm – 2'-O-methylcytidine; D – dihydrouridine; G – guanine; I – inosine; i^6A – N^6-isopentenyladenosine; o_2yW – peroxywybutosine; T – thymine; t_6A – N^6-threonylcarbamoyladenosine; m^1A – 1-methyladenosine; m^1G – 1-methylguanosine; m^1I – 1-methylinosine; m^2G – N^2-methylguanosine; $m^2{}_2G$ – N^2,N^2-dimethylguanosine; m^3C – 3-methylcytidine.

(U20, U47). Additionally, strong peaks can be recognized at tRNA positions 18, 45, 46, and 64. At these positions are no modifications known. For the classification of these candidate sites we recommend the use of chemically treated RNA-seq data.

The different time points vary in the number of modified tRNAs at certain positions (see **Suppl. Tab. B6**). In particular, 6h after starvation the life cycle of the slime mold differs strongly from the remaining investigated developmental stages by an increased or largely reduced number of modified tRNAs at the positions 20, 26, and 37. Time points 0h and 16h after starvation show a significant increase in the occurrence of modified tRNAs at position 9. The number of modified tRNAs at starvation increased compared to the other stages. However, the same amount of altered tRNA genes occurs at positions 32, 34, 58, and 68.

Normally, one would assume that the intensity of RTs correlates with the number of modified tRNAs. We cannot confirm these assumptions with our observations. Although, the number of modified tRNAs at four positions 6h after starvation is much lower compared to the other investigated time points, this time point usually displays the highest fraction of RTs (see **Fig. 34B**). For example, the starvation (0h) and mound formation (16h after starvation) stages of the development of *D. discoideum* show the highest amount of modified tRNA bases at position 9. At this tRNA position, however, RT fractions are nearly 4 times higher at 6h compared to 0h and 16h. Thus, the number of modified tRNAs does not necessarily provide a conclusion about the levels of modifications. Single tRNAs can be strongly modified and expressed at certain points in time, resulting in an increase in the relative fraction of reads over all tRNAs, but still only few tRNAs are modified in total.

Although the same tRNA positions are modified at all investigated stages of the life cycle of *D. discoideum*, they differ in the number of modified tRNA genes and their relative fraction of RTs. These results suggest that the chemical modification of distinct tRNA genes is regulated according to the stage of development. Thus, a potential function of the tRNA modification in the development of the slime mold can be expected.

10.3 Profiling tRNA Modifications in Treatment-Based Procedures

Dihydrouridine (D), pseudouridine (Ψ) and 7-methyl-guanosine (m^7G) modifications slightly affect reverse transcription during cDNA synthesis (see Section 5.7). Base pairing properties of D and Ψ modifications lead to being recognized as standard uridines resulting in no detectable misincorporation sites. Strong D modifications can lead to accumulations of position-specific RTs, however, the signals are often weak and can only be identified in few modified residues as

we saw from our previous analyses (see Section 10.2). In contrast, Ψ and m^7G modifications do not result in apparent RT signals in untreated samples.

To detect D, Ψ, and m^7G modification in tRNAs systematically, C. Lorenz developed RNA treatments designed to convert these modifications so that they yield a specific read-out in the subsequent sequencing step using tRNA-enriched pools. For the specific enrichment of tRNAs, they isolated and separated tRNAs by high salt precipitation of RNAs with higher molecular weight. The quality of the isolated tRNA pools was investigated on a BioAnalyzer device and on high-resolution denaturing polyacrylamide gels with subsequent staining. After quality assessment, the RNA pools were used for chemical treatment of the individual base modifications. They were able to establish chemical detection procedures for Ψ (detection by 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene (CMCT) treatment) and dihydrouridine (D) (detection by sodium borohydride ($NaBH_4$) treatment) as described in Section 5.7. $NaBH_4$ treatment enables also the detection of m^7G methylations. In all cases, untreated samples were produced to identify the modified positions by an increase occurrence of apparent RT accumulations compared to the untreated samples. Three biological replicates were prepared for each treatment as well as for the corresponding negative control for the bacterium *Bacillus subtilis* and submitted to Illumina sequencing. A detailed description of the library preparation protocol is given in A. Hoffmann and C. Lorenz et al. [410] and in the thesis of C. Lorenz [353].

Our task was to develop an evaluation strategy that sensitively and precisely recognizes D, m^7G, and Ψ sites in our treatment-based RNA-seq samples. The technical implementation of this post-mapping analysis strategy is described in detail in Section 7.3.2. In brief, for the direct comparison of the RT expression of the treated libraries and the negative control, the mapped reads are scaled sample- and replica-wise to the same size. Normalization is necessary to ensure that differences in library size do not affect the following down-stream analysis which prevents for inflated false positives in expression measures of RTs. For the profiling of modified sites we applied different statistical measurements. On one hand, we calculated the fold change (FC) for each tRNA and position. Using the FC we are able to test the null hypothesis which states that the logarithmic FC between two conditions for a gene's expression is zero. A zero value indicates modified sites which are not affected by the chemical treatment [354]. Thus, the greater the FC, the stronger is the effect of the treated

condition at the particular site. On the other hand, for each tRNA site we determined the statistical significance of the different expression intensities of RTs between both conditions using a Poisson distribution. We decided to use this distribution since in RNA-seq data each read is sampled independently and consistently from a pool of reads. Therefore, reads can be modeled as a random sampling process. Under this assumption the number of reads coming from a gene follows a binomial distribution and can be approximately described by a Poisson distribution [308, 355]. The fundamental property of the Poisson distribution is that its variance is equal to the mean, which is not generally given in RNA-seq data, especially for highly expressed genes. This so called *overdispersion* problem can be solved by using a generalized linear model framework. A linear model framework is commonly known as a "quasi-likelihood" approach, with Poisson-like assumptions or a negative binomial model. These distributions allow the calculation of an extra dispersion parameter that adjusts the variance independently from the mean [356, 357]. In our case, we do not address multiplicity problems, but rather use the RT counts as a single parameter for statistical regressions. We do not expect high variability, since our analysis focuses only on single tRNAs and not on whole genomic samples. Thus, the Poisson distribution fits well for our purpose. Poisson's regression estimations provide a specific p-value for each tRNA position that describes the probability of enriched RTs occurring in the treatment. In fact, sometimes small p-values happen by chance for multiple tests, which could lead to an incorrect rejection of the null hypothesis. To decrease the false discovery rate we adjust the p-values applying the Benjamini-Hochberg procedure [309].

10.3.1 Parameter Adjustments to Correct Background Noise

Considering all sites which display a statistically significant increase of RT expression in the treated samples (p-value < 0.01, $\log_2(\text{FC}) \geq 0.01$), a large number of hits is obtained. In both, the $NaBH_4$ (see **Fig. 36**) and the CMCT (see **Fig. 37**) treatment, highly-enriched RT sites cannot be distinguished from the background noise by analyzing only sites with a significant increase. Since reverse transcriptase can not only terminate at modified sites but can also react very sensitively to structural peculiarities or other modifications, RTs can occur beside D, m^7G, and Ψ. A background noise may originate because the reverse transcriptase

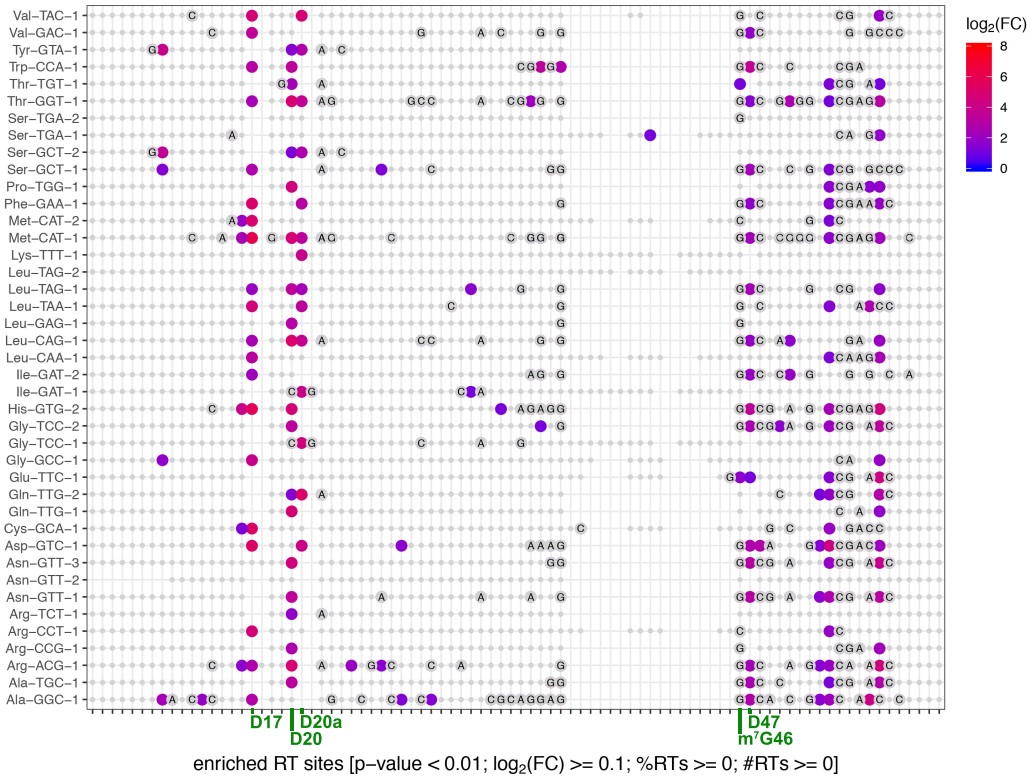

Figure 36: Significant candidate sites in the $NaBH_4$ treatment. All transfer RNA (tRNA) residues of *Bacillus subtilis* showing a significantly (p-value < 0.01) enriched amount of read terminations (RTs) in the sodium borohydride ($NaBH_4$) treatment in comparison to the untreated control sample are highlighted. Significant candidate sites of uridine are shown in color-coded (from blue to red) dots according to the logarithmic fold change (FC). Significantly enriched non-uridine sites are displayed in gray big dots and the corresponding nucleobases are abbreviated by: A – adenine; C – cytosine; G – guanine. tRNA sites that are not significantly enriched in terms of RT coverage are presented in small gray dots. $NaBH_4$ treatment should result in an increase of RTs at dihydrouridine (D; tRNA positions 17, 20, 20a, and 47) and 7-methyl-guanosine (m^7G; tRNA position 46) modifications. Using a minimal FC cutoff a huge amount of candidate sites is obtained. Modifications that should be visible through the treatment cannot be distinguished from background noise. A specific parameterization is necessary to reduce the background noise. Beside the logarithmic FC, the percentage number (%RTs) as well as the total number of RTs (#RTs) per site are suitable parameters for this purpose.

does not always terminate with the same probability and intensity, which leads to other sites producing significant differences in RT expression. Therefore, it is important to distinguish biological variability from such noise. This is possible by considering different parameters and adjusting them according to treatment and species.

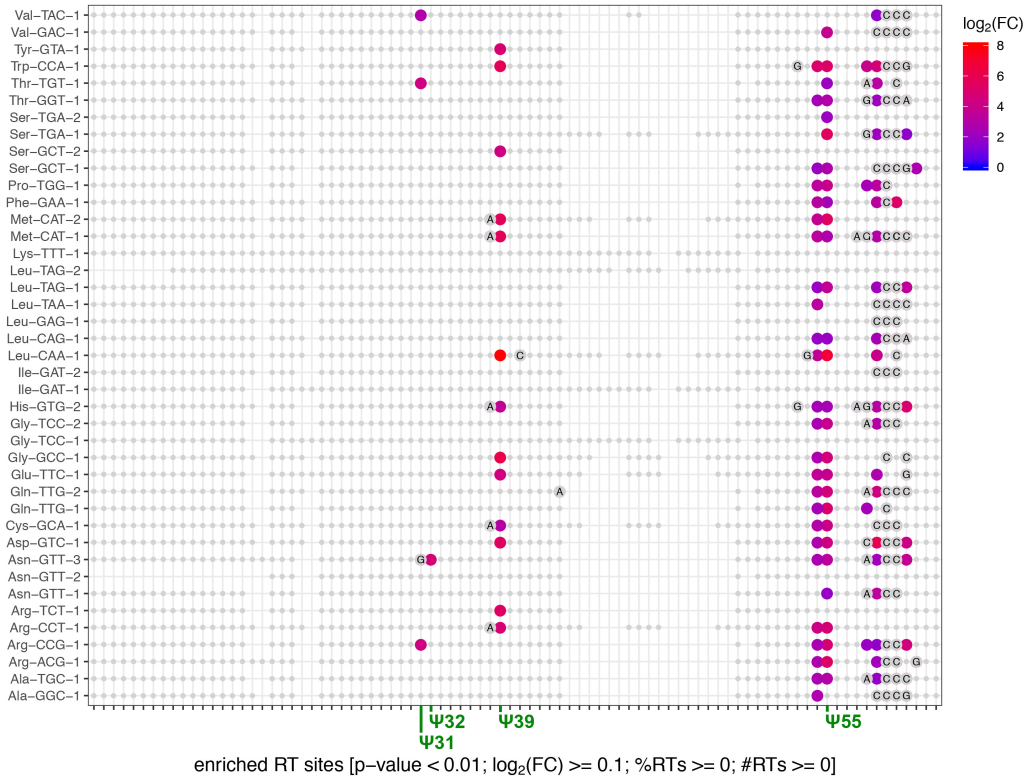

Figure 37: Significant candidate sites in the CMCT treatment. Transfer RNA (tRNA) residues of *Bacillus subtilis* which show a significant (p-value < 0.01) increase of read terminations (RTs) in the 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene (CMCT) treatment compared to the untreated control sample are highlighted. Significant candidate sites of uridine are shown in color-coded (from blue to red) dots according to the logarithmic fold change (FC). Significantly enriched non-uridine sites are displayed in gray big dots and the corresponding nucleobases are abbreviated by: A – adenine; C – cytosine; G – guanine. tRNA sites that are not significantly enriched in terms of RT coverage are presented in small gray dots. CMCT treatment should result in an increased amount of RTs at pseudouridine (Ψ) modifications frequently found in the bacterium at positions 31, 32, 39, and 55. Using a minimal FC cutoff a huge amount of candidate sites is retained. Modifications that should be visible through the treatment cannot be distinguished from background noise. In order to reduce the noise, a specific parameterization is necessary. Beside the logarithmic FC, the percentage number (%RTs) as well as the total number of RTs (#RTs) per site are suitable parameters for this purpose.

Parameter and cutoff selection should not result in a large reduction of TP hits, but should reduce FPs. For a systematic determination of cutoffs in the *Bacillus subtilis* samples, we classified each hit according to TPs and FPs. Hits matching the tRNA positions of D (U17, U20, U20a, and U47), m7G (G46), and Ψ (U31, U32, U39, and U55), which are already

annotated in the tRNAmodviz database, were declared as TPs.

Low-enriched sites can be filtered by a selected $\log_2$ FC cutoff, as treatments only cause a high accumulation of RTs at the specific modifications. It has to be considered, that not each modifications lead to a strong accumulation of RTs. Especially, low modifications may produce few RTs whose presence remains undetected by high FC cutoffs, even if they were enriched by the treatment. A suitable cutoff for the logarithmic FC seems to be 1 for the $NaBH_4$ and 2 for the CMCT treatment, while higher values lead to a strong decrease of TP hits as shown in **Fig. 38A**. Thus, only a 2-fold and a 4-fold enrichment of the $NaBH_4$ and CMCT treatments, respectively, were considered.

Highly enriched signals can also occur when a low RT coverage is present relative to the total number of reads at this position. In both treatments these signals are particularly frequent at the TΨC-arm (T-arm; 3'-end) of the tRNAs. We assume that treatment-based RT sites have a higher ratio of RT reads to non-RT reads at the corresponding position compared to noise signals. Thus, we filtered hits by the percentage of RTs at the respective position. While a cutoff of 1% already results in a low reduction of TPs in the $NaBH_4$ treatment, the number of TPs decreases by 4% in CMCT treatment (see **Fig. 38B**). The selection of the cutoff should not be too stringent, since the highest read coverage occurs in the 5'-range of the tRNAs due to the 3'-to-5'-directed activity of the reverse transcription in the used library preparation protocol. Since these non-RT reads mainly decrease only by modifications in 3'-direction of the tRNA, a high percentage of non-RT reads occur. Therefore, we consider a cutoff of 1% ($NaBH_4$ treatment) and 3% (CMCT treatment) as suitable for this parameter. The selected parameter cutoffs leads to a strong reduction of FPs at the T-arm of the tRNAs. In particular, 66% in the $NaBH_4$ and 76% in the CMCT treatment can be filtered out.

tRNA positions with generally low read coverage may result in significant signals even from a few RTs. This is often the case in the 3'-range (D-arm and 5'-acceptor stem) of tRNAs. In addition, we recognized that the used reverse transcriptase increasingly incorporates erroneous bases in the 3'-range and terminates prematurely. This results in a lower read frequency in the 3'-regions. Such noise can be filtered out by the minimum number of RTs per position. Based on the distribution of TPs and FPs at different cutoff values (see **Fig. 38C**), we set the cutoff for the $NaBH_4$ and CMCT treatment to 100 and 175 RT counts per site, respectively.

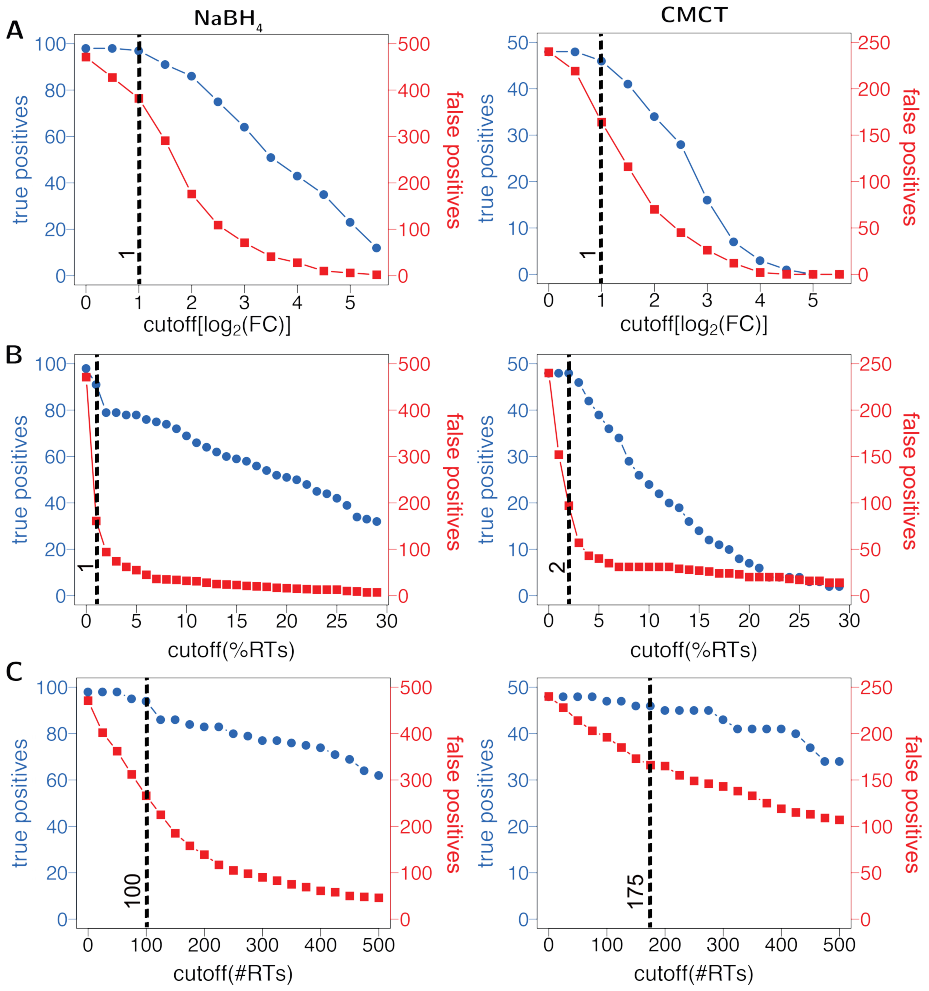

Figure 38: Parameter settings to reduce background noise. Numbers of true positives (known modifications sites in the tRNAmodviz database) and false positives (potential background noise) using different parameter cutoffs are displayed for the sodium borohydride ($NaBH_4$; left side) and the 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene (CMCT; right side) treatment. Cutoffs were defined for **(A)** the logarithmic fold change log_2(FC), **(B)** the percentages of read terminations (RTs) per site and **(C)** the total number of RTs per site. The chosen cutoffs (dashed horizontal lines) should balance the minimum loss of TPs and the maximum noise reduction.

Using the defined parameter settings we can reduce 76% noise in the $NaBH_4$ treatment and 88% in CMCT treatment. In contrast, 87/98 (89%) truly modified sites remain in the $NaBH_4$ treatment and 43/48 (90%) TPs in the CMCT treatment. These parameter settings provide a good balance between preserving TPs and reducing FPs. As a result, we can now

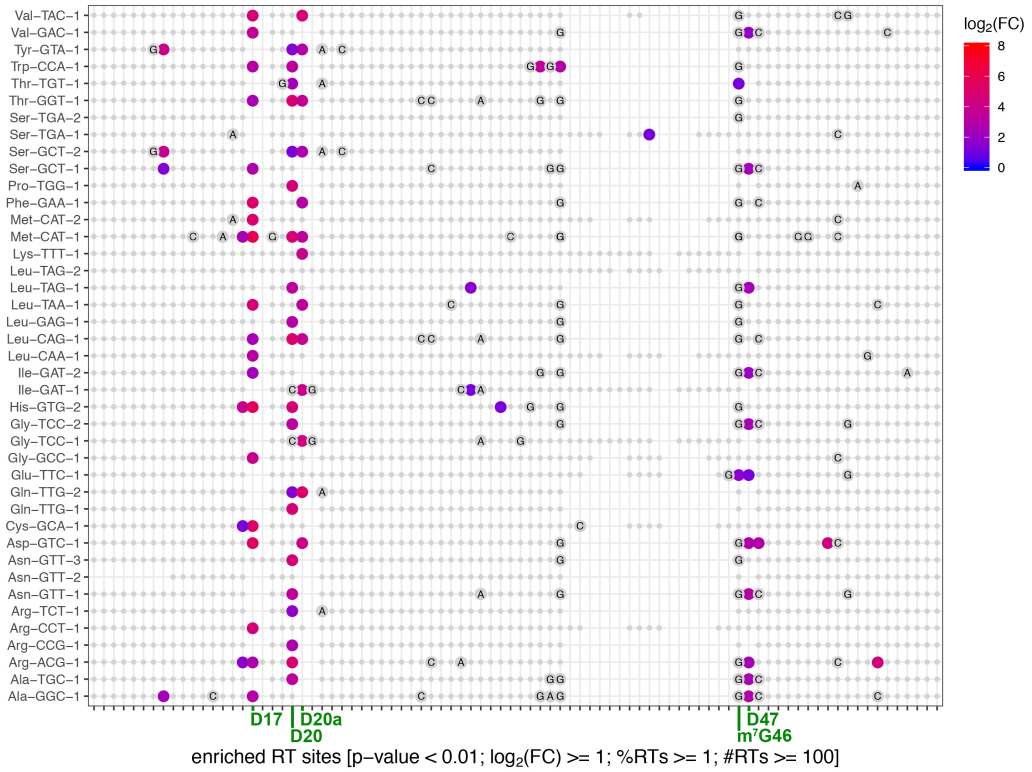

Figure 39: Filtered candidate sites in the NaBH$_4$ treatment. The following parameter adjustments are used to reduce background noise and filter out not strongly enriched RT sites: logarithmic fold change ($\log_2$(FC)) ≥ 1; percentage number of RTs (%RTs) per site ≥ 1; total number of RTs (#RTs) per site ≥ 100; p-value < 0.01. All transfer RNA (tRNA) residues of *Bacillus subtilis* which surpass the parameter cutoffs in the sodium borohydride (NaBH$_4$) treatment in comparison to the untreated control sample are highlighted. Candidate sites of uridines are shown in color-coded (from blue to red) dots according the logarithmic fold change. Non-uridine sites are displayed in gray big dots and the corresponding nucleobases are abbreviated by: A – adenine; C – cytosine; G – guanine. tRNA sites which were filtered out showing not significantly enriched RT coverage are presented in small gray dots. NaBH$_4$ treatment should results in an increase of RTs at dihydrouridine (D; tRNA positions 17, 20, 20a, and 47) and 7-methyl-guanosin (m^{7}G; tRNA position 46) modifications. At precisely these sites we achieved highly enriched RT sites: D17 (#19), D20 (#20), D20a (#13), D47 (#21), and m^{7}G (#11). This results indicate that our analysis is sensitive for detecting both D and m^{7}G modifications.

more clearly separate TP sites from the background noise in both NaBH$_4$ (see **Fig. 39**) and CMCT treatment (see **Fig. 40**).

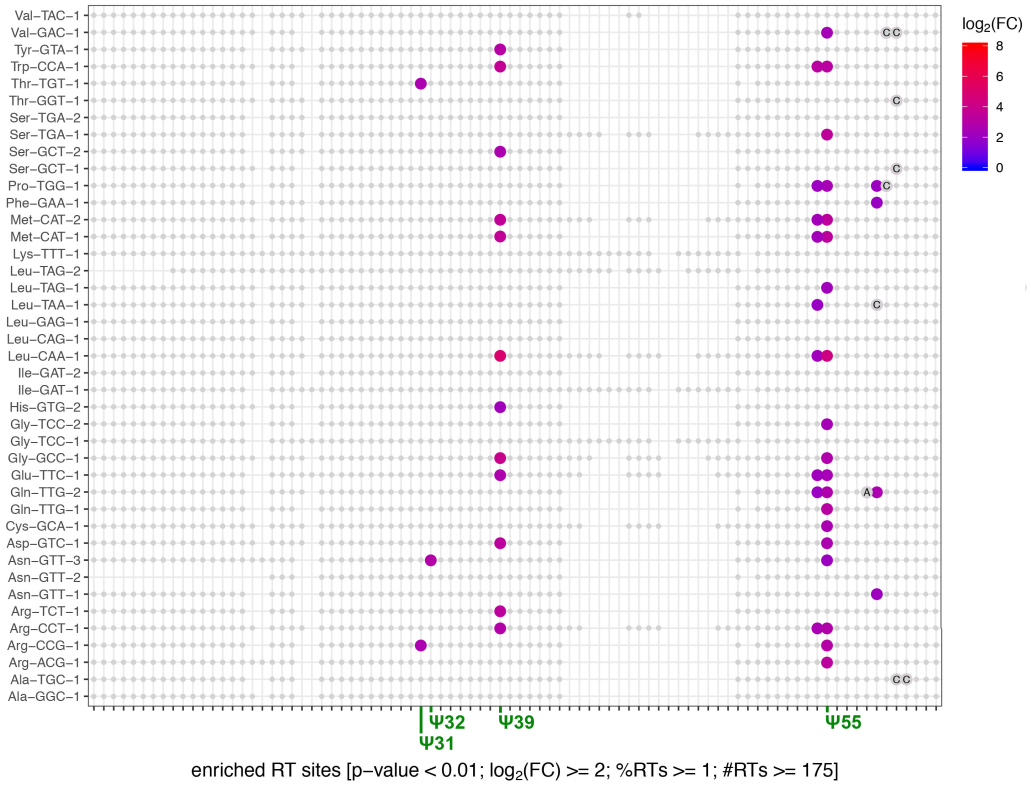

Figure 40: Filtered candidate sites in the CMCT treatment. The following parameter adjustments are used to reduce background noise and filter out not strongly enriched RT sites: logarithmic fold change ($\log_2(FC)$) ≥ 1; percentage number of RTs (%RTs) per site ≥ 3; total number of RTs (#RTs) per site ≥ 175; p-value < 0.01. Transfer RNA (tRNA) residues of *Bacillus subtilis* when they surpass the parameter cutoffs in the 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene (CMCT) treatment are highlighted. Candidate sites of uridine are shown in color-coded (from blue to red) dots according to the logarithmic fold change. Non-uridine sites are displayed in gray big dots and the corresponding nucleobases are abbreviated by: A – adenine; C – cytosine; G – guanine. tRNA sites which were filtered out showing not significantly enriched RT coverage are presented in small gray dots. CMCT treatment should result in an increase of RTs of pseudouridine (Ψ) modifications which are frequently found at positions 31,32, 39, and 55. At precisely these known Ψ sites from the bacterium, we achieved highly enriched RT sites: Ψ31 (#2), Ψ32 (#1), Ψ39 (#19), and Ψ55 (#26). These results indicate that our analysis is sensitive for detecting Ψ modifications.

10.4 Discussion

Detecting tRNA modifications has become a very timely topic in current research. Coupled with new detection methods, new insights have been gained into the function of RNA modifications in the regulation of RNA stability [27, 133], of protein biosynthesis [139, 140, 143, 145], and

of immunity [124]. The combination of reverse transcription-based methods and RNA-seq has emerged as a powerful instrument for the profiling of RNA modifications. Early studies of tRNA sequencing data showed that tRNA modifications in plants are readily detectable by a position-specific increase in the rate of sequencing errors [358, 359]. Later analyses demonstrated that the profiles of sequencing errors are at least approximately reproducible between experiments [261, 263]. It has been proven challenging to call modifications from next-generation sequencing (NGS) data since reads are not only erroneous (with error rates of around 0.1-10% in various sequencing technologies), but result from a complex error process that depends on the properties of the instrument, the preceding mapping tools, the genome itself [266], and from the used reverse transcriptase. To reduce errors during the read mapping step, we applied the best-practice workflow for the accurate mapping of tRNA reads described in Chapter 8. For a sensitive and accurate profiling of tRNA modifications the choice of a suitable modification caller, which is able to divide errors from true modifications, is crucial.

Our benchmark of the appropriate callers `GATK` [269], `bcftools` [267, 268], and `HAMR` [261] revealed that the tools vary strongly in their true positive rate (TPR; sensitivity) and false discovery rate (FDR) as listed in **Tab. 2**. We received the same number of modified tRNAs for only one tRNA position in all three investigated tools. Although `HAMR` was developed directly for transcriptome-wide discovery of tRNA modifications with single nucleotide resolution, it displays the lowest sensitivity in our benchmark. In contrast, the general purpose variant caller `GATK` shows the highest sensitivity closely followed by `bcftools`. The biggest drawback of `GATK` is that it achieves high but still imperfect accuracy, since more than 15% of all found sites could be classified as FPs by visual inspection of the mapping patterns. Similar inaccuracies could be found when using `GATK` for single-nucleotide polymorphism (SNP) calling [360]. The different accuracies could be explained by their implemented algorithmic features when calling modifications, e.g., required misincorporation rate and read coverage (see **Fig. 31**). Our attempt to adjust the cutoffs of `bcftools` via its parameter settings in order to detect weaker modifications remained unsuccessful. No tool was sensitive enough to detect all modifications since the number of true modification sites found by only one of the tools was very high (see **Fig 30**). Since `bcftools` is only slightly less sensitive and does not call any false positive candidate sites compared to `GATK`, we recommend to use `bcftools` for tRNA modification calling based on misincorporation sites.

We did not consider other tools based on data-adaptive methods in our benchmark. Such tools, e.g. `haarz` from the `segemehl` suite [266], are designed for general RNA-seq data which cannot be used for tRNAs without further modifications. In contrast to the variant callers `GATK` and `bcftools`, `haarz` particularly recognized expanded polymerase chain reaction (PCR) artifacts as an overabundance of reads with the same start and stop positions. However, tRNAs do not allow much variation in the positioning of read starts and ends. Reverse transcription stops lead to sharp changes in coverage and hence to many reads with the same end position. This often supports the presence of a chemical tRNA modification rather than an indication of an PCR artifact. Therefore, a data-adaptive method needs to be aware of the peculiar patterns produced by short, heavily structured transcripts.

By applying our analysis strategy to open–source rmRNA-seq data of six different human tissues (cerebellum, diencephalon, ovary, skeletal muscle, esophagus muscularis mucosae, and testis) we find compelling evidence for tissue-specific differences of tRNA modification patterns. However, a functional explanation remains elusive. At this point we can only note that neither the developmental stage (cerebellum, diencephalon, skeletal muscle from fetal sources and testis, ovary, esophagus muscularis mucosae from adult sources) nor the tissue turnover rate [361] seem to be a convincing cause for the differences. It is well known that tRNA expression patterns vary up to 10-fold between different tissues. However, testis and ovary show distinct tRNA modification patterns but very similar tRNA expression patterns relative to brain tissue samples [362]. This makes it very unlikely that our reported modification differences are an artifact resulting from ascertainment biases caused, e.g., by different sensitivities for the detection of varying modification patterns. On the other hand, some biological explanations for expression differences might also account for tissue-specific modifications. One hypothesis postulates that modifications may be introduced to reduce misfolding and the subsequent degradation and/or dysfunctioning of tRNAs [363]. Differences in the modification patterns may influence the relative abundance of functional tRNA and favor the expression of messenger RNA (mRNA) with a suitably adapted codon bias [364]. It is also conceivable, however, that the tissue-specific modification patterns are the result of different expression levels of specific modification enzymes that have evolved for reasons unrelated to tRNA biology.

It is important to note in this context that by no means all chemical modifications are visible in untreated RNA-seq data due to the accumulation of misincorporation rates

produced by the reverse transcription during cDNA synthesis. Some base modifications lead to reverse transcription terminations and become detectable as apparent accumulations of RTs by sequencing techniques [262, 335]. Since a high tRNA read coverage allows quantitative statements about accumulations of RTs, we used tRNA-enriched RNA-seq data for our analyses, e.g., LOTTE-seq and optimized sRNA TruSeq data. The analysis of the tRNA sequence reads identified a series of accumulations of RT signals resulting from reverse transcription termination at specific modified base positions. However, gradations are recognizable in both tRNA-enriched RNA-seq data. The number of modified tRNAs and the relative fraction of RTs are considerably increased in human LOTTE-seq data compared to the optimized sRNA TruSeq reads illustrated in **Fig. 33**. The specific protocol of LOTTE-seq, in which a two-stage adapter ligature procedure has been implemented, enables the capture of a much higher number of (partial) cDNA fragments resulting from premature RT termination. While this deceases the number of reads spanning certain modifications, the corresponding pileup of RTs in LOTTE-seq is an excellent indicator for these base modifications.

Our analysis of base-specific RT signals in combination with the corresponding base misincorporation pattern allows the unambiguous classification of several tRNA methylations in human tRNAs, e.g., m^1A, m^1G, $m^2{}_2G$, i^6A, and m^3C. Our observed RT patterns perfectly correlate with the outcome of previews studies [242, 265, 365]. The investigated tRNA-enriched data appear sufficient to display the listed tRNA modifications due to their effect on RT incorporation. Other modifications, i.e., t_6A, I, o_2yW, and Cm can only be classified by their base misincorporation sites and/or by the use of specific chemical treatments. For D modifications, which usually show no altered Watson–Crick face in the sequencing reads, we only observed weak RT signals. Therefore, profiling of D modifications via RTs is unsuitable and we recommend to apply a chemical treatment for their detection. In general, for the classification of tRNA modification we used the obtained reverse transcription signals. For this purpose, we expanded and improved knowledge about modification-specific reverse transcription signals of previous studies [119, 241, 242, 261–263, 265, 365] based on our observations. This allows an easier classification of tRNA modification patterns by analyzing reverse transcription-based RNA-seq data. Our improved collection of common tRNA modifications and how they become visible in sequencing data is given in **Suppl. Tab. B2**. Since some reverse transcriptase signals also fit to several modification types, an unambiguous assignment without

prior knowledge is challenging. Therefore, we incorporated the information stored in the tRNAmodviz database [119] for classifying the modification patterns. The database contains information on tRNA modifications collected from various scientific literature and tRNA-seq experiments. Since many tRNA modifications for single tRNAs and species are not well investigated or not included in the database, e.g., *D. discoideum*, a stringent classification is not possible for all retained signals using untreated RNA-seq data (see **Fig. 35**). Therefore, it cannot be excluded that we have misclassified some tRNA modifications or assigned RTs as true modifications, e.g., resulting from robust secondary structures. In order to be able to classify the modifications unambiguously, specific chemical treated RNA-seq data are imperative. Since most treatments can only display a single modification type, it is very cost-intensive to cover a wide range of different types of modifications.

When analyzing time dependent reverse transcription signals in tRNA genes during the development of *D. discoideum*, we always observed modified sites at the same tRNA positions as illustrated in **Fig. 35**. However, the numbers of modified tRNAs and RT signal intensities vary drastically between the investigated developmental stages (see **Fig. 34**). Of the 12 detected types of tRNA modification, only five are modified within the same tRNAs (see **Suppl. Tab. B6**). The identification of positions modified in the same tRNAs in all investigated samples indicates the applicability of our LOTTE-seq data and that the found differences in the life cycle of *D. discoideum* are probably not artifacts. Interestingly, modified tRNAs during multicellular aggregation (6 hours after starvation) of the slime mold development display the highest relative fraction of RTs for almost all modified positions. Nevertheless, the number of modified tRNAs is often lower at these positions compared to the other stages. Some tRNAs are probably more strongly modified leading to these increased RT fraction sites during cell aggregation. A possible biological interpretation of the developmental-specific modification patterns is given in the thesis of L. Erber.

Since only D-modified tRNA residues show weak fractions of RT signals in untreated samples, specific chemical treatments are necessary for their visibility in mapped RNA-seq reads. To this end we developed a treatment-based RNA-seq method and an analysis strategy allowing the systematic profiling of D, m^7G ($NaBH_4$ treatment), and Ψ (CMCT treatment) modifications in the transcriptome with single-nucleotide resolution. Based on our statistical quantification of the data, we detected a variety of RT sites which are significantly enriched

(p-value < 0.01) in the treated samples compared to the negative control. Classifying the significantly enriched RT sites as modified sites is not trivial. A great challenge was that the treatments are performed for up to two different modifications. In contrast to gene expression analysis, where statistical robustness and performance can be increased by aggregating reads from an entire gene, measurements of tRNA modifications comprise only single nucleotides [260]. Another challenge was the separation of background noise from the true modified sites which is often the case in comparable analyses (see **Fig. 36** and **37**). Enriched RT sites corresponding to background noise can be due to RNA structure, reverse transcription error rate, complex processing of the RNA, genomic misalignments of sequencing reads and, technical errors of the sequencing platform. Specific parameter cutoffs can be used to reduce the background noise as shown in **Fig. 38**. We assume that the modified sites are considerably more enriched due to the treatment than signals from the background noise. The use of a specially adapted FC cutoff is, therefore, suitable. Other background noise which occurs from low RT read coverage, primarily located at the 5'-ACC-stem and the 3'-region of the tRNAs, can be filtered out by only considering sites showing an absolute number and percentage of RTs at the respective position over a determined threshold. The adjustment of these parameters cutoffs allows to reduce 76% background noise in the $NaBH_4$ treatment (see **Fig. 39**) and 88% noise in CMCT treatment (see **Fig. 40**). However, it should be noted that this filtering strategy of background noise leads to a reduction of TP sites ($\leq 11\%$). We were not able to completely reduce the background noise without accepting an increased false negative rate by more stringent parameter settings.

10.5 Data Sources

Strand-specific rmRNA-seq data were obtained from the Encode project [324, 325] for Section 10.1. The ENCODE data sets were chosen to represent three different organs: brain, muscle and gonades of *Homo sapiens*. For each organ two different tissues were considered. Biosamples showing approximately the same age were used as replicas. The tissues cerebellum (female 19 weeks and female 37 weeks: GEO:GSE78291) and diencephalon (female 20 weeks and male 22 weeks: GEO:GSE78292) were selected for brain, esophagus muscularis mucosa (female 51 years: GEO:GSE88169, female 53 years: GEO:GSE88236, male 37 years:

GEO:GSE88128) and skeletal muscle (female 19 weeks and male 22 weeks: GEO:GSE78300) for muscle organ as well as testis (male 54 years: GEO:GSE88414, male 37 years: GEO:GSE88124) and ovary (female 51 year: GEO:GSE87965) for gonade.

Genomes of *D. discoideum* (assembly dicty 2.7), *H. sapiens* (assembly hg38) and *B. subtilis* (strain NCIB 6310) were downloaded from NCBI, release 90 [22]. Numbers of annotated tRNA genes for each genome are given in **Suppl. Tab. B3**.

CHAPTER 11

Synteny-Based Orthology Identification of tRNAs

Contents

Gene families evolving under concerted evolution are not amenable to classical phylogenetic analyses since paralogs maintain identical, species-specific sequences, precluding the estimation of correct gene trees from sequence differences. This leaves conservation of syntenic arrangements with respect to "anchor elements" that are not subject to concerted evolution (see Chapter 4) as the only viable source of phylogenetic information. However, our newly developed, purely synteny-based workflow is quite capable of solving this problem.

Our workflow distinguish orthologs and paralogs in transfer RNA (tRNA) families that are subject to concerted evolution in a more systematic than previous studies [196, 197]. The workflow based on the use of uniquely aligned adjacent sequence elements as anchors to establish syntenic conservation of sequence intervals. In practice, anchors and intervals can be extracted from genome-wide multiple sequence alignments (MSAs) (see Section 5.5.1 and Section 6.1.1). To this end a so-called synteny map was implemented which harbors information about syntenic tRNA gene clusters which are subdivided by the genomic anchors (for the methodical implementation see Section 7.4). Syntenic clusters of concertedly evolving genes of different families were then subdivided by list alignments, leading to usually small clusters of candidate co-orthologs as described in Chapter 6. On the basis of recent advances in phylogenetic combinatorics, these candidate clusters were further processed by cograph editing to recover their duplication histories. The workflow can be conceptualized as step-wise refinement of a graph of homologous genes.

This chapter is based on Velandia-Huerto et al. [376] titled by *Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies*. In a further work, the described workflow was refined to a fully automatized pipeline (`SMORE`) and is published in Berkemer et al. [377] with the title *SMORE: Synteny Modulator Of Repetitive Elements*. I Implemented the creation of synteny maps as described in Section 7.4. Subsequent steps of the workflow were implemented by S. J. Berkemer based on the concept described in Section 6. For further details of the implementation and discussion of the workflow see the dissertation of S. J. Berkemer [366]. In this chapter, the biological applications of the workflow is given to revisit the evolution of tRNAs in primates, as an example for a phylogenetically very narrow range, and in fruit flies, as an example for a phylogenetically already very diverse system.

11.1 Evolution of Primate tRNAs

Starting from the primate MULTIZ alignments [287] (see Section 6.1.1) we obtained 1665 connected components of the candidate graph Γ_c, including 961 singletons. In 168 connected components, tRNAs of only a single species were found. 536 connected components formed non-trivial graphs showing the orthology relation between tRNAs in distinct species. Almost all of the connected components of the estimated orthology graph Γ_o were already cographs. Only 3 of the 536 graphs had a non-cograph structure. This appears to be related to pseudogenization of parts of the cluster, which causes some of the pairwise distances of the pseudogenized tRNAs to drop below the threshold value for orthology assignment.

The connected components based on the MULTIZ MSA blocks are typically small and show very few tandem duplications. This may be caused by the choice of one particular copy of duplicated sequence flanking a tRNA in the MULTIZ pipeline. The corresponding gain and loss events are mapped to the primate phylogeny in **Fig. 41**. To investigate this effect we therefore joined connected components of the MULTIZ-based Γ_c that share boundary MSA blocks. This reduced the number of synteny regions by about a third to 1079 connected components and about halved the number of singleton from 961 to 482. Still, we found 64 components comprising tRNAs of only a single species. Of 533 non-trivial connected components only 2 did not have cograph structure. The main effect joining adjacent synteny groups, i.e., considering larger syntenic groups in the initial step, is that events are assigned to evolutionary more ancient events. This is a consequence of reconstructing larger clusters as the ancestral state, so that more deletions from these clusters are inferred instead of evolutionary more recent events of seeding novel clusters.

Of the cographs, 327 were subsets of adjacent vertices (cliques) and thus did not contain duplication events. The remaining 206 include duplication events that increased the total number of tRNAs by 66. In addition, 60 duplications were detected in the connected components containing only tRNAs of the same species.

In summary, we observe that between about a third and a half of the tRNAs in extant primate genomes have been syntenically conserved since the last common ancestor of human and macaque. The seeding of new tRNA locations, on the other hand, is clearly an ongoing process. A surprisingly large number of loci is gained and lost in a lineage-specific manner.

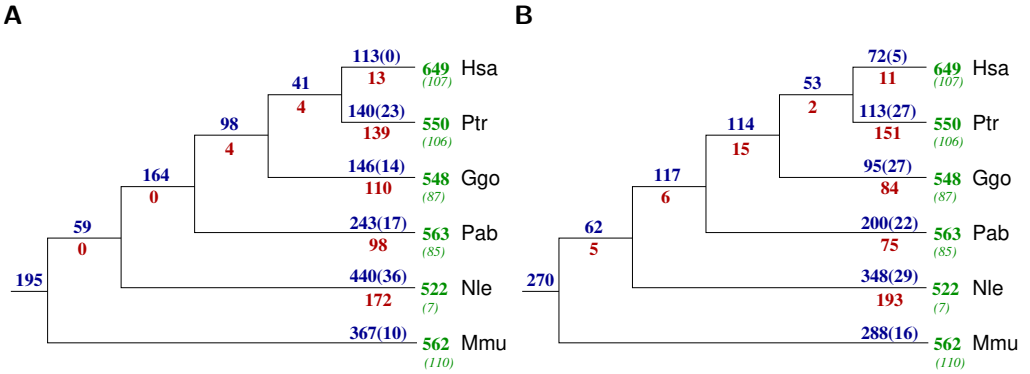

Figure 41: Gain, loss, and duplications of tRNAs in primates. Gain, loss, and duplications of transfer RNAs (tRNAs) in primates computed from the most fine-grained synteny definition based on individual multiple sequence alignment (MSA) blocks **(A)** and by joining adjacent blocks **(B)**. Gain and duplication events were assigned to the edge leading to the last common ancestor of all observed co-orthologs, except for groups that contained only a macaque and a human or a chimpanzee tRNA; in these cases we assigned two lineage-specific gains. Green numbers refer to the total number of tRNAs detected by `tRNAscan-SE`; green numbers in parentheses count the pseudogenes found in the set of all tRNAs. Blue numbers refer to the total gain, i.e., the sum of event seeding new connected components and duplication events with a connected component. The number of identified local duplication events is given in parentheses in blue. The red numbers indicate the loss events on the corresponding branch. Species abbreviations: human, *Homo sapiens*: Hsa; chimapanzee, *Pan troglodytes*: Ptr; gorilla, *Gorilla gorilla gorilla*: Ggo; orangutan, *Pongo abelii*: Pab; gibbon, *Nomascus leucogenys*: Nle; rhesus macaque, *Macaca mulatta*: Mmu.

This effect can be attributed to the rapid formation and erasure of pseudogenized copies. Errors in the genome assembly and the genome-wide sequence alignments will lead to false negatives in the synteny assessment and thus to unrecognized orthologies.

11.2 Evolution of tRNAs in Drosophilids

Fruit flies cover evolutionary distances comparable to the entire vertebrate phylum [367]. Nevertheless the synteny-based method of ortholog identification remains applicable since the much smaller genomes still provide a sufficient density of anchors with unique sequences. Based on the `MULTIZ` alignments provided through the UCSC genome browser we identified 1889 connected components including 1235 singletons. 375 connected components contained tRNAs of just one species. The remaining 280 connected components were graphs showing the orthology relations between tRNAs of distinct species. Out of these, 275 graphs have a

cograph structure and in only 5 cases the graph structure had to be edited to get the closest possible cograph structure. Analogously, for the primate case, clusters were then joined such that two clusters sharing the same border became one cluster. This reduced the number of connected components by about 40% to 1042, of which 722 did not have any edges. 602 of these graphs were singleton tRNAs and in the remaining 110 only tRNAs of the same species were found. All the 320 non-trivial graphs were cographs. Out of these, 190 cographs were cliques. In the remaining 130 graphs, 205 duplicated tRNAs could be detected. Additionally, 349 duplications were detected in the graphs containing tRNAs of the same species.

As in the case of primate tRNAs, a substantial fraction of tRNAs can be traced back to the drosophilid ancestor and has been syntenically conserved since then (see **Fig. 42** for the joined MSA and **Suppl. Fig. A8** for the individual MSA-based approach). The seeding of new loci that subsequently are conserved in most species is again an ongoing process, accompanied by a relatively small rate of losses. As in the case of primates, the overwhelming part of the turnover is lineage-specific and involves nearly half of the extant tRNA complement.

Due to the different genome version used in [197] and the UCSC `MULTIZ` alignments only about 90% of the tRNA genes can be related unambiguously between the two data set. Thus, a comparison of the coordinates systems was not possible for *Drosophila willistoni*, *Drosophila sechellia*, and *Drosophila persimilis*. In the remaining species we were able to establish $1:2$ correspondences for 2196 tRNAs. In the case where tRNAs could not be matched, $1:1$ the `Liftover` tool [312] and sequence similarity information were used to identify the most likely corresponding tRNA sequences. The remaining 216 tRNAs of Rogers et al. [197] could not be unambiguously assigned to 246 tRNAs appearing in our tRNA data. For the total of 2196 tRNAs, we identified 796 pairwise orthology relations with the `MULTIZ`-anchored approach. The orthology map of Rogers et al. [197] restricted to the same tRNAs comprises 5493 edges, 644 of which coincide with our much more restrictive orthology assignments. When clusters are joined, we increased the number of co-orthologs, thus increasing the number of ortholog pairs to 1808 of which 1061 coincide with the $1:1$ assignments of Rogers et al. [197]. Since the `BLAST`-regions (see Section 5.5.1) used in Rogers et al. [197] often correspond to very distant anchors, their orthology assignments are much more inclusive.

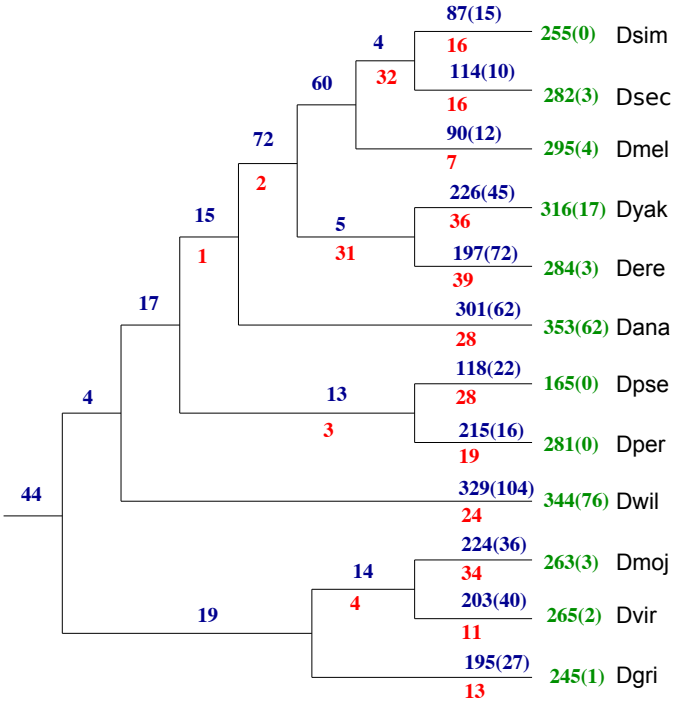

Figure 42: Gain, loss, and duplications of tRNAs in drosophilids. Gain, loss, and duplications of transfer RNAs (tRNAs) in drosophilids computed based on synteny definitions given by joined adjacent multiple sequence alignment (MSA) blocks. Gain and duplication events were assigned to the edge leading to the last common ancestor of all observed co-orthologs, except for groups that contained only one tRNA sequence of two species; in these cases we assigned two lineage-specific gains. Green numbers refer to the total number of tRNAs detected by `tRNAscan-SE`; green numbers in parentheses count the pseudogenes found in the set of all tRNAs. Blue numbers refer to the total gain, i.e., the sum of event seeding new connected components and duplication events with a connected component. The number of identified local duplication events is given in parentheses in blue. The red numbers indicate the loss events on the corresponding branch. Species abbreviations: *Drosophila simulans*: Dsim; *Drosophila sechellia*: Dsec; *Drosophila melanogaster*: Dmel; *Drosophila yakuba*: Dyak; *Drosophila erecta*: Dere; *Drosophila ananassae*: Dana; *Drosophila pseudoobscura*: Dpse; *Drosophila persimilis*: Dper; *Drosophila willistoni*: Dwil; *Drosophila mojavensis*: Dmoj; *Drosophila virilis*: Dvir; *Drosophila grimshawi*: Dgri.

11.3 Numerous tRNA Remolding Events Occur

Numerous remolding events summarized in **Fig. 43** were detected in both, primates and drosophilids. The remolding events identified here are largely congruent with those reported in Rogers et al. [197] for fruit flies and Rogers et al. [196] in primates, see **Tab. 3**. A detailed overview of all identified remolding events and their comparison to the named previous work

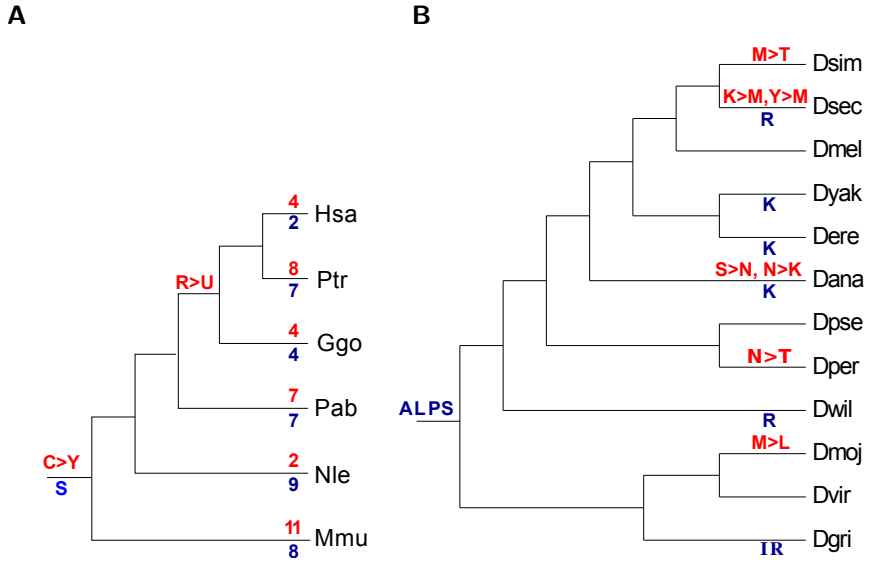

Figure 43: Remolding events in primates and drosophilids. Remolding events in primates **(A)** (summary statistics only) and drosophilids **(B)** (affected isoacceptor classes). Isoacceptor remoldings are shown in dark blue, alloacceptor remoldings are given in red. Details of all anticodon changes are given in **Suppl. Figs. B7** and **B8**. The tRNAs isoacceptor classes are indicated by their 1-letter codes: A (alanine); C (cysteine); G (glycine); H (histidine); I (isoleucine); K (lysine); L (leucine); M (methionine); N (asparagine); P (proline); R (arginine); S (serine); T (threonine); U (selenocystein); Y (tyrosine). Species abbreviations: S *Drosophila simulans*: Dsim; *Drosophila sechellia*: Dsec; *Drosophila melanogaster*: Dmel; *Drosophila yakuba*: Dyak; *Drosophila erecta*: Dere; *Drosophila ananassae*: Dana; *Drosophila pseudoobscura*: Dpse; *Drosophila persimilis*: Dper; *Drosophila willistoni*: Dwil; *Drosophila mojavensis*: Dmoj; *Drosophila virilis*: Dvir; *Drosophila grimshawi*: Dgri; human, *Homo sapiens*: Hsa; chimapanzee, *Pan troglodytes*: Ptr; gorilla, *Gorilla gorilla*: Ggo; orangutan, *Pongo abelii*: Pab; gibbon, *Nomascus leucogenys*: Nle; rhesus macaque, *Macaca mulatta*: Mmu.

are provided in **Suppl. Tab. B7** and **B8** for primates and fruit flies, respectively. Our method is somewhat more sensitive and predicts significantly more tRNA remolding events. The overwhelming majority of events maps to the terminal branches of the phylogenetic trees. This concerns in particular almost all alloacceptor remoldings. Most likely, most or all of these "terminal" remolding events lead to non-functional tRNAs and already constitute pseudogenes. Despite the much greater phylogenetic depth of the drosophilid clade [367], we observe fewer remolding events. This may be explained at least in part by the larger total number of tRNAs in the primate genomes. Most of the detected remolding events occur close to the leaves of the phylogenetic tree. In principle they might be artifacts deriving from sequencing errors. While

we cannot strictly rule out this interpretation, we deem it unlikely. First, the observed number of events would be unusually high: for primates we observe in total 73 remolding events at the leaves of the tree, in a sample of 1985 tRNAs. Within the three anticodon positions this would amount to a sequencing error rate of about 0.012, compared to an expected error rate in assembled contigs of $\ll 10^{-4}$/nt [368]. For drosophilids we observed 14 remoldings in 3348 tRNAs, amounting to a substitution rate of ≈ 0.0014. This also cannot be explained by sequencing errors. Second, we recover between 85% and 95% of the results in [196, 197] and detected additional putative remolding events although in part different genome assemblies were used. A much more plausible explanation therefore is that most remoldings affect tRNA function so that remolded tRNAs are unlikely to survive longterm and are rapidly pseudogenized and removed from the genome.

Interestingly, a small set of remoldings of threonine (Thr) tRNAs was observed multiple times in primates, namely $\text{tRNA}^{\text{Thr}}_{\text{AGT}}$ which was changed to $\text{tRNA}^{\text{Thr}}_{\text{CGT}}$ and $\text{tRNA}^{\text{Thr}}_{\text{TGT}}$. Surprisingly, we identified one alloacceptor remolding whose descendants persisted in primate genomes since the common ancestor of human and rhesus. An ancestral cysteine (Cys) $\text{tRNA}^{\text{Cys}}_{\text{GCA}}$ gave rise to a remolded $\text{tRNA}^{\text{Tyr}}_{\text{GTA}}$ whose sequence is still nearly identical to the Cys-decoding ancestor, see **Fig. 44**. While we have no direct evidence that this $\text{tRNA}^{\text{Tyr}}_{\text{GTA}}$ is a functional tRNA, its evolutionary conservation is at least suggestive of some functional role.

Table 3: Comparison of remolding events with previous studies. All annotated tRNA remolding events of primates **(A)** and fruit flies **(B)** were compared to the previous studies of Rogers et al. [197] and Rogers et al. [196], respectively. Remolding events were grouped corresponding to their annotation origin: found in our and the previous studies (common), only in the previous studies (Rogers et al.), and only in our study (novel).

(A) Primates	**Common**	**Rogers (2014)**	**Novel**
Isoacceptor remolding	9	0	9
Alloacceptor remolding	17	3	17
(B) Drosophilids	**Common**	**Rogers (2010)**	**Novel**
Isoacceptor remolding	7	1	5
Alloacceptor remolding	4	1	3

Figure 44: Example alignment of a tRNA remolding event. Alignment of transfer RNAs (tRNAs) deriving from the cysteine (Cys) $tRNA^{Cys}_{GCA}$ to the tyrosine (Tyr) $tRNA^{Tyr}_{GTA}$ remolding event predating the last common ancestor of human and rhesus. Descendants of both tRNAs have survived in all investigated genomes except gibbon. The secondary structure is given in dot-bracket notation. Species abbreviations: chimapanzee, *Pan troglodytes*: Ptr; gibbon, *Nomascus leucogenys*: Nle; gorilla, *Gorilla gorilla gorilla*: Ggo; human, *Homo sapiens*: Hsa; orangutan, *Pongo abelii*: Pab; rhesus macaque, *Macaca mulatta*: Mmu.

11.4 Intron-Containing tRNAs are Genomically Clustered

Some tRNAs contain short introns. Introns are removed by a dedicated enzymatic machinery which is not only fundamentally different from spliceosomal splicing but also differs between Archaea and Eukarya [369]. Nevertheless, most tRNA introns are located in the "canonical position", one nucleotide 3' to the anticodon [180]. We use tRNAs as an independent test for orthology assignment. We expect that either all or none of the members of a groups of (co-)orthologous tRNAs have an intron. This is indeed the case: In primates, there are 87 clusters of predicted orthologs in which all members carry an intron. In all other clusters none of the tRNAs has an intron. In drosophilids we found 49 clusters containing tRNAs with introns. All but a single one comprise tRNAs with introns only. The only exception is a leucine (Leu) tRNA cluster, namely $tRNA^{Leu}_{CAA}$, that also include single $tRNA^{Leu}_{CAG}$ sequences from the highly diverging *Drosophila grimshawi*. It remains unclear whether this case constitutes a true change in intron structure, or whether the *D. grimshawi* tRNA is a false positive ortholog assignment. Despite a possible concerted evolution effect we observe that tRNA introns typically exhibit multiple substitutions and some insertions and deletions. In a small number of clusters of orthologous tRNAs in drosophilids we observe a considerably variation in intron length; in the extreme case introns have lengths between 21 and 52 nucleotides. This may not be unusual given that the phylogenetic depth of the drosophilids exceeds that of the mammalian radiation [367].

11.5 Discussion

Gene families that are subject to mechanisms of concerted evolution cannot be studied with traditional phylogenetic methods because concerted evolution rapidly erases all information about their evolutionary relationships from the sequences of paralogs. We investigated how synteny information can be harnessed in a systematic manner for this purpose. We demonstrated that synteny *in principle* provides the necessary information as long as syntenically conserved sequence blocks are long enough to contain unique sequences that can be used as anchors. While it may seem desirable to use a full-fledged sequence-based model such as `OrthoAlign` [290] to track genome evolution in detail, such approaches do not scale to genome-wide surveys because of the computational efforts required. We reason that a step-wise workflow that first localizes the problem to individual gene clusters is a good compromise. These still can be prohibitively large, in particular in mammalian genomes. We therefore opted for a strategy that uses synteny information as much as possible.

A surprising observation is that a large part of the inferred gain and loss of tRNAs is species-specific (see **Figs. 41** and **42**). While this observation may be partially confounded by residual noise in the synteny assignments, it can be explained by a rapid copying of tRNAs followed by rapid pseudogenization. The tRNA model implemented in `tRNAscan-SE` [255] is very specific and distinguishes very stringently between tRNAs and tRNA pseudogenes that may differ by only a few point mutations from their functional ancestors. Reconstruction gain and loss events are largely consistent between the three levels of stringency in the definition of synteny, with most of the differences concentrated to the species-specific gains and losses.

Remolding events are observed predominantly at shallow phylogenetic depth (see **Fig. 43**), indicating that most of them occur in pseudogenes. In contrast, remoldings that persist over large phylogenetic distances are rare and almost never change the isoacceptor class. Only a single deep alloacceptor remolding was observed. While it is unlikely that the remolded tRNA is functional in translation, it is well conceivable that the gene serves one of the recently described secondary functions of a tRNA, as a source for microRNA (miRNA)-like small RNAs [25, 65], as sponge [370], or as a genomic insulator element influencing chromatin organization [371].

11.6 Data Sources and Workflow Availability

The following genomes and assemblies were used for primates: *G. gorilla* (assembly gorGor3), *H. sapiens* (assembly hg38), *M. mulatta* (assembly rheMac3), *N. leucogenys* (assembly nomLeu3), *P. abelii* (assembly ponAbe2), and *P. troglodytes* (assembly panTro4) as well as for drosophilids: *D. ananassae* (assembly droAna3), *D erecta* (assembly droEre2), *D. grimshawi* (assembly droGri2), *D. mojavensis* (assembly droMoj3), *D. melanogaster* (assembly dm6), *D. persimilis* (assembly droPer1), *D. pseudoobscura* (assembly droPse3), *D. simulans* (assembly droSim1), *D. sechellia* (assembly droSec1), *D. virilis* (assembly droVir3), *D. willistoni* (assembly droWil2), and *D. yakuba* (assembly droYak3). For the `MULTIZ` alignments we used the multiple genome-wide alignments of 19 mammalian (16 primate) genomes with human (assembly hg38) and the multiple genome-wide alignments of 26 insects with *D. melanogaster* (assembly dm6). All genomes and both `MULTIZ` alignments were downloaded from the UCSC genome browser [312]. Numbers of annotated tRNA genes for each species are given in **Suppl. Tab. B3**. The `SMORE` pipeline is freely available from `https://github.com/AnneHoffmann/Smore`.

CHAPTER 12

nm-tRNAs: Could They be Functional?

Contents

The biological relevance of nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) is still unknown, indicating that they have been poorly investigated so far. However, their presence raises intriguing questions about their possible functionality. Only one nm-tRNA annotation strategy (see Section 5.6) has been published [24, 57]. In this perspective, the following chapter includes our systematic method to annotate nm-tRNAs that enables to detect them even if their are strongly degraded. In addition, we observed evidence that nm-tRNAs serve as binding sites for RNA-binding proteins (RBPs).

This chapter is based on S. Hoser and A. Hoffmann et al. [411] titled *Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing*. The biological background can be found in Section 2.3 and the technical background is given in Section 5.6. The methodological implementation is described in Section 7.5.1 and the performance evaluation is specified in Section 7.6.

12.1 Many Unidentified nm-tRNAs are Present in Nuclear Genomes

To scan, in particular, the mouse and human genomes for nm-tRNAs, we tested different combinations of annotation tools and strategies. Although `tRNAscan-SE` is not intended to annotate nm-tRNAs, we applied the integrated mt-tRNA search mode not to mitochondrial genomes (mt-genomes), but to nuclear sequences. In another approach we used the covariance models (CMs) from `MiTFi`. These CMs contain information on aberrant mitochondrial tRNAs (mt-tRNAs) in addition to the normal mt-tRNA sequence and structure consensus which can help to detect nm-tRNAs exposed to high selection pressure. We used `Infernal` as search engine for the CMs from `MiTFi`. For each nm-tRNA annotation strategy, we tested two different approaches, which we term *NUMT-based* and *genome-based* approach. In the *NUMT-based* approach we only used the nuclear mitochondrial DNA (NUMT) sequences as reference. In the *genome-based* approach, however, the entire nuclear genome was applied.

All methods yielded very different results. With the *NUMT-based* approach for the human genome, we received 775 hits from `Infernal` and 726 hits from `tRNAscan-SE`. In contrast, the *genome-based* approach provides a large variance. Here we got only 367 hits from `Infernal`, whereas `tRNAscan-SE` scored about 2.65 times more hits (977 hits). We found very similar relations in the analysis of the mouse genome. We got 105 hits from `Infernal` and 79 hits

from `tRNAscan-SE` within the *NUMT-based* approach. The hits from the *genome-based* approach vary from 75 (`Infernal`) to 246 (`tRNAscan-SE`). In general, we assume that all hits located in NUMTs are nm-tRNAs. It cannot be excluded that the NUMT annotation is incomplete on the basis that NUMTs are probably not very conserved. Thus, hits outside of NUMTs, which do not overlap with known tRNA annotations, are potential nm-tRNAs. Since we cannot clearly identify these hits as nm-tRNAs and cannot rule out that they are other types of unannotated tRNAs, we refer to them as mt-tRNA-lookalikes.

The performance evaluation is based on synteny information given by each NUMT, as the primordial origin of each NUMT is traceable (see **Tab. 4**). Within the *NUMT-based* approach, `Infernal` found 2% more nm-tRNAs in human than `tRNAscan-SE` resulting in a true positive rate (TPR) of 0.91. Despite the lower sensitivity of `tRNAscan-SE`, the tool counts only 29 false positives (FPs) compared to the 68 FP hits of `Infernal`. The difference is even stronger in the *NUMT-based* approach for mouse, where `Infernal` found 13% more nm-tRNAs, but also 11% more FPs compared to `tRNAscan-SE`. `tRNAscan-SE` shows the highest sensitivity in the *genome-based* approach with a TPR of 0.88 and 0.72 in human and mouse, respectively. `Infernal` delivers much less true positives (TPs) in both species and is therefore not suitable for this method. In both species, the number of FPs annotated by `tRNAscan-SE` is less than 1% if only hits within NUMTs are considered. However, `tRNAscan-SE` finds over 4 times more mt-tRNA-lookalikes when applying the *genome-based* approach compared to `Infernal`. In order to obtain as many true nm-tRNAs as possible, we used the results of both tools in our subsequent analysis. Taken together all tested annotation strategies, we identified 355 novel nm-tRNAs in human (total 731) and 43 in mouse (total 92). Compared to Telonis et al. [57], we identified 45% more human nm-tRNAs. In mouse, the predicted nm-tRNAs in Telonis et al. [24] show a low TPR of 0.47, while our TPRs range from 0.72 to 0.85. These previous studies only found a comparatively low number of hits (497 in human and 53 in mouse), which explains the reduced sensitivity and the smaller false discovery rate (FDR). In summary, our method is much more sensitive compared to the previous studies allowing the identification of a high number of novel nm-tRNAs in human and mouse.

Table 4: Performance evaluation of different nm-tRNA annotation strategies. Absolute counts of true positives (TPs), false positives (FPs) and false negatives (FNs) of annotated nuclear mitochondrial tRNA (nm-tRNAs) of **(A)** *H. sapiens* and **(B)** *M. musculus* are given for the *NUMT-based* and *genome-based* approach for single tools and their combination. The true positive rate (TPR), false discovery rate (FDR) and false negative rate (FNR) is given as well. Counts were calculated based on synteny information given by the mitochondrial origin of the nuclear mitochondrial DNAs (NUMTs). The same validation was carried out with data already published by Telonis et al. [24, 57]. mt-tRNA-lookalikes are marked in blue and are classified as potential FPs. In each approach, `tRNAscan-SE` shows the best balance between TPR and FDR. Although `Infernal` got the highest count for TPs in the *NUMT-based* approach, the tool shows a reduced sensitivity in the *genome-based* approach. A combination of both increases the TPR, at the expense of FPs. Compared to currently published data, our implemented methods found > 1.8 times more TPs.

Method	Tool	TPs	FPs	FNs	TPR	FDR	FNR
(A) *H. sapiens*							
NUMT-based	`Infernal`	707	68	72	0.91	0.09	0.09
NUMT-based	`tRNAscan-SE`	697	29	82	0.89	0.04	0.11
NUMT-based	both	726	95	53	0.93	0.12	0.07
genome-based	`Infernal`	300	5+62	479	0.39	0.02 (0.18)	0.61
genome-based	`tRNAscan-SE`	689	7+281	90	0.88	0.01 (0.29)	0.12
genome-based	both	689	12+315	90	0.88	0.02 (0.32)	0.12
Telonis (2014)	`BLAST`	376	0 +121	403	0.48	0 (0.24)	0.52
(B) *M. musculus*							
NUMT-based	`Infernal`	88	17	17	0.84	0.16	0.16
NUMT-based	`tRNAscan-SE`	75	4	27	0.71	0.05	0.29
NUMT-based	both	89	17	16	0.85	0.16	0.15
genome-based	`Infernal`	33	2 +40	72	0.31	0.06 (0.56)	0.69
genome-based	`tRNAscan-SE`	76	3 +167	29	0.72	0.04 (0.69)	0.28
genome-based	both	76	4 +199	29	0.72	0.05 (0.72)	0.28
Telonis (2015)	`BLAST`	49	0 +4	56	0.47	0 (0.08)	0.53

37% of our detected nm-tRNAs in human and 39% in mouse are part of an annotated transcript (protein-coding genes, non-coding genes, long non-coding RNAs (lncRNAs), pseudogenes, intergenic, or exonic). Previous computational studies have also demonstrated

the presence of nm-tRNAs within introns of nuclear protein-coding genes in humans [57] and also in the mouse and opossum genomes [372]. nm-tRNAs located in introns of known transcripts are termed nuclear-encoded intronic mitochondrial-derived tRNAs (nim-tRNAs). We identified a total of 273 human nim-tRNAs in the introns of 76 different host genes, of which 30 were protein-coding, 28 were constituted of long intergenic non-coding RNAs (lincRNAs), 13 were non-coding RNAs (ncRNAs) and 5 were pseudogenes. The JAK2 (#18) and the DYNC2H1 (# 14) genes as well as the LINC00630 (# 13) lincRNA and the GUSB pseudogene 6 (# 13) contain the highest accumulation of human nim-tRNAs. In total 121 of the identified human nim-tRNAs are novel. In mouse, 14 of our annotated 36 nim-tRNAs, which are located in 12 different host genes (9 in protein-coding genes and 3 in lncRNAs), are novel. The Myo3a (# 7) gene and the Cep295-201 (# 5) intron harbor the largest amount of nim-tRNAs. However, nim-tRNAs are not present in any homologous transcripts of the others species.

12.2 Are nm-tRNAs Target Sites for RNA-Binding Proteins?

We observed that the overwhelming majority of human nm-tRNAs show evidence of negative selection in their host genomes, since their evolutionary conservation in mammals measured by phylogenetic p-value (PhyloP) scores is very low. While we found that PyloP scores are slightly enhanced in nm-tRNAs compared to the surrounding NUMT sequences (see **Fig. 45**), the selection pressure is not strong enough to identify nm-tRNAs under strong negative selection. As we received higher evolutionary conservation for a few nm-tRNA fragments, we interpret these as possible binding sites that have emerged from the inserted mt-tRNA sequences. We found that binding sites of 31 proteins, which have a function in splicing or play other regulatory roles (see **Suppl. Tab. B9**), overlap with nim-tRNAs. Of these proteins, DHX30, G3BP1, and NSUN2 have a more than 2-fold enriched binding site coverage in nim-tRNAs.

However, testing the conservation of (parts of) a NUMT is not trivial. While it is simple in principle to use conservation measures such as the PyloP score computed for genome-wide alignments, one has to take into account that NUMTs, due to their quasi-repetitive nature, may have an incurred problem in genome assemblies and/or may be misaligned. We, therefore, used a complementary approach to measure evolutionary conservation of nm-tRNAs and

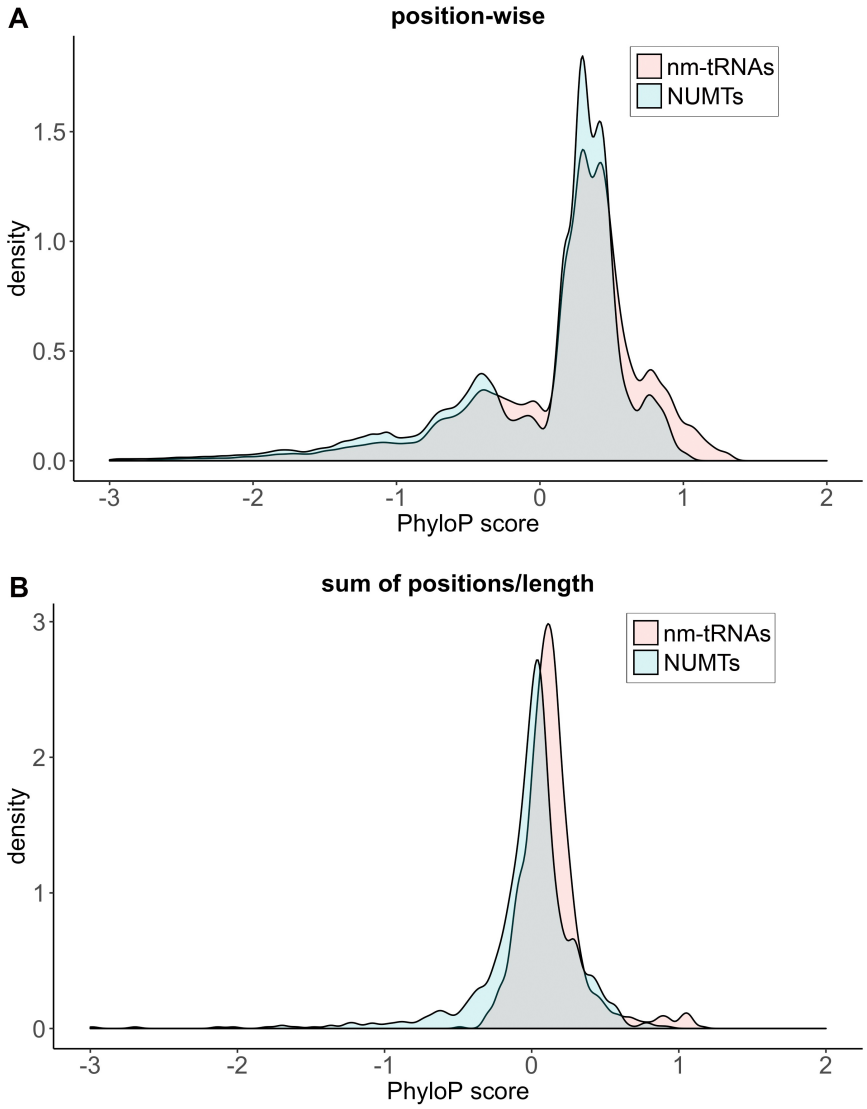

Figure 45: Conservation densities of nm-tRNAs and NUMTs. Densities of conservation are shown for nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) and the surrounding nuclear mitochondrial DNA (NUMT) sequences. Conservation is measured by phylogenetic p-values (PhyloP) where **(A)** displays the density of PhyloP scores for each single nucleotide of the sequences and **(B)** the normalized scores (sum of PhyloP scores per sequence divided by sequence length). PhyloP scores were taken from multiple genome alignments of 29 mammalian genomes to human. Although most nm-tRNAs are not subject to negative selection, the PhyloP scores in nm-tRNAs are slightly enhanced compared to the surrounding NUMT sequences. A higher density for larger PhyloP values can be seen in **(A)** which suggests that a few short nm-tRNA fragments are more conserved than the remaining nm-tRNA sequences.

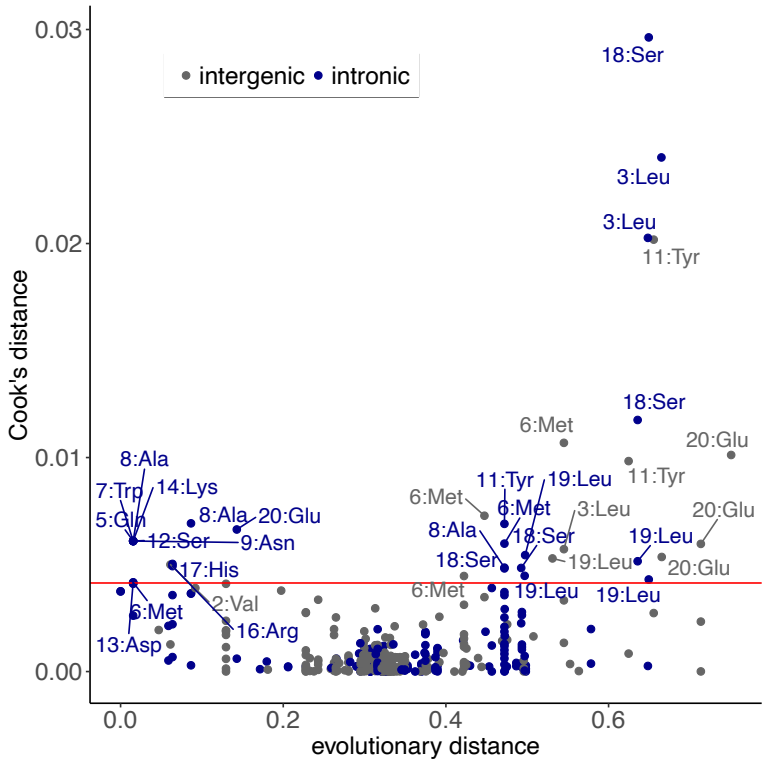

Figure 46: Evolutionary conserved nm-tRNAs. Outliers of nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs) are shown above the red line. These outliers are subject to a stronger stabilizing selection after their insertion into the nuclear genome relative to nuclear mitochondrial DNAs (NUMTs). We defined all observations with Cook's distance larger than three times the mean Cook's distance as outliers. The majority (25 of 36) of the more extreme outliers are intronic.

NUMTs in human. Within this approach we identified about a dozen nm-tRNAs that appear to have evolved significantly slower than the adjacent NUMT sequences (see **Fig. 46**). It is also interesting to note that nim-tRNAs are represent the majority of the more extreme outliers.

Finally, we found computational evidence that nm-tRNAs and nim-tRNAs are a source for functional binding sites. As expected in such a scenario, most nm-tRNAs and nim-tRNAs have not attained functional significance because they are simply not in a useful genomic context or there is no selective advantage to be gained from a nm-tRNA or nim-tRNA-derived binding site at the position of the insertion.

12.3 Discussion

A great advantage of our nm-tRNA annotation strategy is that we considered synteny information, which can be determined from the mitochondrial origin of the NUMTs. Thus, false positive hits can be discarded to obtain the most accurate set of true positive nm-tRNAs. However, this method depends strongly on the accuracy of the annotated NUMTs. The best methodology for locating NUMTs in nuclear genomes has not been carefully examined yet. The established NUMT annotation methods show strong differences in the length and number of NUMT sequences [56, 373, 374]. The loss of highly selective pressure and the resulting low conservation of the NUMT sequences make annotations difficult. In our analysis we used the NUMT annotation from Tsuji et al. [373] as it is the only one that takes into account the low identity when aligning older NUMTs to mtDNA. However, we also found inaccuracies here. For example, we found a highly conserved nm-tRNA-like sequence outside of the NUMT boundaries of the human catenin–ß–like 1 (CTNNBL1) gene. However, the high conservation of the nm-tRNA sequence suggests that the NUMT sequence was annotated too short due to the strong degradation. Such inaccuracies in the NUMT annotation may increase the false negative hits within our analysis.

Within our analysis we applied two different tools (`tRNAscan-SE` and `Infernal`). We got the best TPRs in the *NUMT-based* method using `Infernal`. The disadvantage of `Infernal` is that the results always show the highest FDRs. This method is useful when synteny information are available and FPs can be filtered out. Thus a large set of nm-tRNAs can be obtained. If no synteny information is available, only `tRNAscan-SE` should be used, since `Infernal` delivered very small numbers of TPs. This can be explained by the different scoring systems of the tools. While `tRNAscan-SE` is able to detect hits for complete genome sequences sensitively, `Infernal` seems to be designed for shorter sequences. The majority of the recovered nm-tRNAs in the *genome-based* approach using `tRNAscan-SE` also follows the genomic distribution of NUMTs, which is in favor of the method. Overall, `tRNAscan-SE` shows the best balance between TPs and FPs in each approach. Of course it is also possible to combine each tool and approach, which slightly increases the TPR but also the FDR.

A global analysis of mitochondrial-derived tRNA sequences within the human genome yielded a total of 726 loci of which 273 are located in introns of 76 different host genes. The

uneven distribution is explained by the fact that nim-tRNAs, like their mitochondrial sources, are often arranged in clusters within NUMTs. As a group, neither nim-tRNAs nor nm-tRNAs in general are under detectable stabilizing selection, suggesting that they usually have not acquired new functions after their insertion in the nuclear genome. However, there are a few nm-tRNAs whose evolutionary rates appear retarded (see **Fig. 46**). These are preferably nim-tRNAs. As most nm-tRNAs are not under negative selection is consistent with the fact that they are non-processed and in general do not have a function as an independent ncRNA. They rather function as evolutionary raw material for binding sites, which only acquire a detectable function if the NUMTs are inserted at a suitable site. The association with introns and the over-representation of binding sites for DHX30, G3BP1, and NSUN2 in nim-tRNAs suggests that nim-tRNA can acquire a function in the regulation of splicing provided the NUMT is inserted in fortuitous intronic locations.

12.4 Data Sources and Availability

Mitochondrial and nuclear genomes of *Homo sapiens* (assembly hg38) and *Mus musculus* (assembly mm10) were downloaded from NCBI, release 90 [22]. The annotation of NUMTs were obtained from Tsuji et al. [373] for the older assemblies mm9 and hg19. The NUMT coordinates were converted to the latest genome assemblies mm10 and hg38 for mouse and human, respectively, applying the UCSC `Liftover` utility [312]. PhyloP scores of the multiple alignments of 29 mammalian genomes to hg38 were also downloaded from UCSC. Transcript annotations were obtained from Ensemble release 96 [314]. RBP interaction sites were downloaded from the ENCODE [324, 325] eCLIP repository. A complete list of our annotated nm-tRNAs and nim-tRNAs of mouse and human can be found in S. Hoser and A. Hoffmann et al. [411].

CHAPTER 13

Conclusion and Outlook

Since 1950, when the central dogma of molecular biology [8] has been proposed, the manifold involvement of RNAs in protein synthesis has been widely recognized. Fundamental research has revealed the importance of different types of RNA for these processes. Especially transfer RNAs (tRNAs) represents the physical linkage between the genetic code and the amino acid sequence of proteins during translation [2, 7]. Even though tRNAs are one of the oldest molecules discovered in all areas of life, they are still intriguing study objects. Recent advances in the biology of tRNAs suggest that these classical non-coding housekeeping RNAs are key components of the small RNA-mediated gene regulatory system beyond translation [81]. As such, their functionality is linked to the presence of various tRNA base modifications, to tRNA sequence variants known as isoacceptors and isodecoders, and to the versatility of protein binding partners. This complexity offers a large repertoire of tRNA species that fulfill various functions in cellular homeostasis and in adapting cellular functions to changing environments. It is likely that the origin of these functions dates back to the fundamental role of RNAs in early evolution [375]. A strong impulse for these discoveries originates from the advances of RNA sequencing (RNA-seq) methods which enables the profiling of the transcriptome using deep sequencing technologies [209].

Sequencing of tRNA is challenging both experimentally and computationally. With respect to data analysis, the challenges include to overcome reverse transcription errors introduced by chemically modified nucleotides. Furthermore, it is intricate to map the reads to the true genomic tRNA origin, given their multiple identical and almost identical genomic loci. In order to overcome these pitfalls, we developed an innovative mapping strategy to accurately align short tRNA reads which identifies and solves mapping artifacts resulting from simpler mapping schemes. The workflow is discussed and evaluated on simulated and human RNA-seq data (see Chapter 8). In brief, the reads are mapped against a modified target genome in which known tRNA loci are masked and instead intronless tRNA precursor sequences are appended as artificial "chromosomes". In a first pass, reads displaying specific precursor hallmarks are filtered out. In the second pass, the mature tRNA reads are mapped against mature tRNA sequence cluster which assemble identical tRNAs helping to overcome the multi-copy nature of tRNA genes. Thus, uniquely mapped high confident tRNA reads can be used for downstream analyses. Since we adjusted the mapping parameters in each step individually, our pipeline is able to handle the high density of modification induced mismatches in the alignment which even

enables to reliably identify many of the chemical tRNA modifications. Using simulated data, the false discovery rate (FDR) to call tRNA modifications by base misincorporations is small as 2%. Although the method was developed specifically for tRNAs, in our pilot study (data not shown in this thesis) [397], we were able to recover some of the previously reported [376] modification and editing sites of microRNAs (miRNAs). In further work, with the method being applicable also to miRNA data, it becomes feasible to investigate modification patterns and their evolution also in other multi-copy RNA families with nearly identical paralogs, e.g., small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs).

Standard RNA-seq methods such as ribo-minus RNA sequencing (rmRNA-seq) or total RNA-seq only capture a relatively low amount of tRNA sequences [213]. For example, rmRNA-seq contains only ∼0.9% short ncRNAs, while other non-tRNA transcripts are mainly long non-coding RNAs (lncRNAs) (∼81.8%), mRNA-like RNAs (∼8.8%) and snoRNAs (∼2.9%). The amount of non-tRNA reads may lead to a falsification of the tRNA analysis results, especially when calling modified nucleosides or quantifying expression. Current tRNA-seq methods only cover full-length tRNA reads, while tRNA fragments or incomplete cDNAs are lost due to reverse transcription terminations at modified nucleosides. Therefore, as depicted in Chapter 9, we cooperated on benchmarking and improving long hairpin oligonucleotide-based tRNA high-throughput sequencing (LOTTE-seq), a method for efficient capturing of tRNAs for deep sequencing analysis. Our benchmark exhibited that LOTTE-seq combines the advantages of other valuable approaches [216, 217], while avoiding their disadvantages. The usage of a DNA hairpin adapter that specifically hybridizes to the tRNA 3'-CCA end ensures that exclusively mature tRNA transcripts are investigated. In the reaction catalyzed by T4 DNA ligase [362] only full-length CCA ends are accepted for ligation. tRNAs with partial or lacking CCA ends are efficiently excluded, as T4 DNA ligase does not tolerate single-stranded nucleic acids or double strands that carry a gap in the hybrid region between the CCA end and adapter overhang. The benchmark of LOTTE-seq with other tRNA-specific RNA-seq methods demonstrated that the content of tRNAs with CCA end is highest in LOTTE-seq data, ranging from 90% in *S. oleracea* to 100% in *D. discoideum*. LOTTE-seq renders the analysis of tRNA pools or individual transcripts including some modification more efficient and accurate. It is worth noting that the additional use of different chemical treatments should expand the range of modifications that are detectable by LOTTE-seq data. A combination of LOTTE-seq

data with methods that apply enzymatic treatment of specific modifications (ARM-seq [242], AlkAniline-seq [365], and DM-tRNA-seq [240, 265]) is a promising strategy for studying the variety of tRNA modifications. Such combinations would dramatically increase the number of reliable tRNA reads and thus facilitates the accurate identification of position-specific modifications which only become detectable by chemical treatments. We assume that this will not only improves the statistical relevance of tRNA expression data but also sets the stage to implementing tRNAs as powerful biomarker to detect various cellular states.

Mature tRNAs contain by far the highest density of chemically modified nucleotides of all known nucleic acids. These impact structure and function, and even contribute to regulation of translation. Nevertheless, they have received little systematic attention. In Chapter 10 we focused on the sensitive and precise detection of tRNA modifications in different kinds of RNA-seq data. After application of our "accurate mapping of tRNA reads" workflow, we used the mapped reads to call RNA–DNA differences, that are indicative of chemical tRNA modifications. In the simplest case, tRNA modifications affect the reverse transcriptase during cDNA synthesis leading to a position-specific increase in the rate of sequencing errors in their mapping profile. In order to find a suitable and freely available tool to call tRNA modification by mismatch patterns, we evaluated three tools (`HAMR`, `GATK`, and `bcftools`) that seemed to be qualified for our analyses. Surprisingly, the number of called modifications sites varies greatly between the tools, so does the amount of sites profiled by one tool alone. In conclusion, none of the three tools is sensitive enough to reliable call tRNA modifications. However, `bcftools` currently seems to be the best available tool for tRNA modification calling. It does not call any false positive candidate sites and it is only slightly less sensitive than `GATK`, and it found significantly more modifications than `HAMR`. In further work, our short benchmark should systematically be expanded in order to get detailed conclusions about individual error sources and to determine the limits of detectability of tRNA modifications. Simulated tRNA reads can be used for this purpose. The simulated reads should be modified according to different models of nucleotide modification distribution and abundance. This should essentially follow the suggestions of Tserovski et al. [316], who discovered that misincorporation signatures depend on the neighboring sequence context. Such a benchmark allows to adjust the parameters of the investigated tools to maximize the accuracy of modification calls. For our benchmark we used rmRNA-seq data. Only less than 1% of all reads in rmRNA-seq data mapped to

tRNAs [213] which increases the probability of detecting false positives (FPs) due to the low tRNA read coverage. Therefore, further benchmarking should also consider different types of tRNA-enriched RNA-seq data, e.g., YAMAT-seq [216] or LOTTE-seq.

By applying our analysis strategy to detect tRNA modifications by accumulations of base-calling to generic human small rmRNA-seq data showed that the resulting called modification sites can distinguish between very similar tRNAs. Furthermore, we could show that the vast amount of publicly available small RNA-seq data is a hitherto mostly untapped resource to examine the modification status of tRNAs. Thus, we used the data to discover that there are surprising differences in the modification patterns between human tissues. While most tRNA positions are frequently modified in each of the investigated tissues (cerebellum, diencephalon, skeletal muscle, testis, ovary, and esophagus muscularis mucosae), there exist drastic deviations at several modified tRNA residues. It has been well known that differences in tRNA modification can be associated with human diseases, e.g., mitochondrial dysfunctions, metabolic defects, neurological disorders, and cancer [377]. However, our results demonstrate that variation in modification patterns are not at all limited to dysfunction or immortalized cell lines but appear naturally in healthy tissues. This novel observation certainly deserves closer inspection in future work. In particular, it will be interesting to see if tissue specific differences are evolutionary conserved, which would suggest that they are directly related to functional differences between tissues. The detection of modifications due to their misjudgment can also be applied to other multi-copy RNA families, including miRNAs, snoRNAs, and snRNAs. This allows the investigation of specific modification patterns and their evolution in different types of RNAs.

Moreover, some tRNA modifications become visible as accumulations of read termination (RT) fragments in the mapping profile. RT fragments results from specific modified nucleosides that terminate the reverse transcriptase during complementary DNA (cDNA) synthesis. Using RNA-seq data such as LOTTE-Seq, which contain a high read coverage of RT fragments, we were able to identify strong RT signals at specific modified bases. The combination of mapping patterns, of mismatch incorporations, and RTs are good indicators to classify most tRNA modifications. Thus, we expanded and improved the current knowledge [119, 241, 242, 261–263, 265, 365] of reverse transcription signals for common modification types. However, we recognized that a unambiguous classification of tRNA modifications merely on the basis of

their reverse transcription signals is not always possible. As such 1-methylguanosine (m^1G), N^2-methylguanosine (m^2G), and N^2,N^2-dimethylguanosine ($m^2{}_2G$) show the same mapping profiles. An exact classification is only possible by incorporating a priori knowledge about the position of the specific modifications. For this purpose we used the information from the tRNAmodviz [119] database. However, it is problematic for less studied organism such as *Dictyostelium discoideum*, where no or little is known about tRNA modification sites. In ongoing work, our modification signature collection could be extended for easier classification of tRNA modifications in less investigated species. This extension should include context-sensitive reverse transcription signatures for specific tRNA modifications, since misincorporation signatures depend on the neighboring sequence context [316]. In addition, a possible context-dependent correlation of reverse transcription terminations during cDNA synthesis is an interesting topic for future research. Furthermore, in another pilot study (data not shown in this thesis) we observed differences in the reverse transcription signatures applying different types of reverse transcriptions (TGIRT and SuperScript III). This discovery certainly deserves closer inspection in future work. In particular, a systematic analysis of the accuracy and reproducibility of the modification dependent reverse transcription signatures and their dependency on the use of different commercially available types of reverse transcriptases should be investigated. This analysis will shed light on the usability and reliability of modification identification by reverse transcription-based misincorporations depending on the nature of the reverse transcription enzyme and the sequence context in the templating tRNA. Additionally, since our benchmark of variant callers showed that they are still far from perfect, an implementation of a improved modification caller should be part of future research directions. Beside base-calling errors, `HAMR` differentiates between different classes of modifications by the knowledge taken from the tRNAmodviz database [119]. A novel modification caller should consider the specific reverse transcription signatures of our improved collection and their context specificity, as well as signatures generated by several reverse transcription enzymes. Since these patterns will depend on the experimental conditions, the algorithm could follow the basic idea of the `haarz` variation caller [266] of the `segemehl` suit. The `haarz` algorithm estimates the relevant parameters from the data current set using the assumption that the vast majority of the covered sites is unmodified and hence can be used to parametrize the background model.

Already in 1984, Schachner et al. [378] provided evidence that tRNAs were modified in a development-specific manner in *D. discoideum*. The analysis of reverse transcription signals obtained from LOTTE-seq data designed for different developmental stages of slime mold revealed that the same tRNA positions are modified in each stage. However, at certain time points there are strong differences in the number of modified tRNAs and in the relative abundance of RTs. This findings indicates a potential regulatory function of tRNA modifications in the life cycle of the slime mold. The functional explanation is discussed in the thesis of L. Erber (*in preparation*).

In another study, we developed an innovative analysis strategy for sensitive detection of tRNA modifications in treated RNA-seq data. These RNA-seq data has been constructed to convert a specific read-out in the mapping profiles by introducing RTs at dihydrouridine (D) and 7-methyl-guanosine (m^7G), or pseudouridine (Ψ) modifications in *Bacillus subtilis*. We used data from the bacterium given that its tRNA modifications are well-studied which allowed us to roughly validate and adjust our methodology. After sample normalization, we scanned for RT sites that are significantly enriched in the mapping profile compared to the untreated control sample according the Poisson distribution. Most challenging was to distinguish true modification sites from the background noise. Background noise frequently occur since the reverse transcription is also reacts sensitive to structural peculiarities or to other modifications which are not specifically enriched by the treatment. Complex processing of RNA, genomic misalignments of sequencing reads, and technical errors of the sequencing platform contribute to background noise. Thus, we filtered out sites under a determined fold change (FC) threshold, since we assumed that the modified sites of D, m^7G, and Ψ are considerably higher enriched due to the treatment than signals from the background noise or other modifications. We discarded background noise that occurs from low RT read coverage, primarily at the 5'-ACC-stem and the 3'-region of the tRNAs, by considering only sites with an absolute number and percentage of RTs above a certain threshold. Adjusting these parameters cutoffs allows us to reduce over 76% background noise, while only less than 11% of truly modified sites got lost. However, further strategies are necessary to reduce the sources of noise. For example, Ψ occur at specific sequence and/or structural motifs [166, 167, 169]. Filtering potential Ψ sites an the basis of such motifs can help to reduce background noise [260]. It is also possible that the reverse transcriptase used (SuperScript IV) reacts very sensitivity

to sequence and secondary structure peculiarities of tRNAs causing an increased RT rate at unmodified sites compared to other enzymes. In general, the reverse transcriptase terminates one position before the treated tRNA residue. However, we observed sometimes that the SuperScript IV reverse transcriptase terminates directly at the modification or up to two residues after it. For example, the reverse transcriptase always produced significant and highly enriched accumulations of RTs one position before and directly at the m^7G46 modification only if the tRNA does not contain a variable loop. As this concerns two neighboring guanines, it is very difficult to clearly allocate the modification to the true site when no prior knowledge is available. Additional experiments based on other reverse transcriptases should be generated in further work to determine a suitable reverse transcriptase which does not produce such kinds of artifacts. A great advantage of our analysis strategy is that we also take background noise into account which is caused by non-uridine sites in order to reduced the FDR. This enables us to consider the full spectrum of background noise, rather than ignoring the noise of non-uridine sites, as is common in certain studies [166, 167, 169, 379].

Since we demonstrated that our analysis strategy allows the identification of most modified sites of Ψ, D, and, m^7G, our analysis should be expanded to other species in further work. D and Ψ modifications are well known to affect the flexibility of tRNA structures as an adaptation to the environmental temperatures [27, 133], similar to the effect of many chemical modifications in proteins [134, 135]. Thus, ongoing work should investigate the differences in modification patterns of Ψ and D in different psychrophilic, mesophilic, and thermophilic representatives at their corresponding minimal, optimal, and maximal growth temperatures.

In Chapter 11 we introduce a synteny-based framework to distinguish orthologs and paralogs in tRNA gene families. As most tRNAs are typically present as multi-copy genes, the members of the individual tRNA families evolve under concerted or rapid birth-death evolution. Thus, paralogous copies maintain almost identical sequences over long evolutionary time-scales. To a good approximation these are functionally equivalent. Thus, selective pressure on individual tRNA copies is low. Such tRNA genes are evolutionary unstable and can easily mutate into pseudogenes and get lost. This leads to a rapid turnover of tRNAs and often large differences in the tRNA complements of closely related species. Since tRNA paralogs can not be distinguished by their sequence, common methods cannot not be used to establish tRNA gene orthology. However, synteny which describes the maintenance of relative genomic

positions can be considered to disambiguate evolutionary relationships of tRNA genes. Within our framework we studied on the one hand whether pre-computed genome-wide alignment blocks can efficiently be used as syntenic conserved anchors for this purpose. On the other hand, we showed that on the basis of sequence-based synteny data it is possible to approximate the history of gene clusters that cannot be resolved further. To this end we combined an alignment-like approximation to multi-species synteny with recent advances in phylogenetic combinatorics [291, 292] that relate orthology with cographs.

An additional outcome of this thesis is to highlight the technical problems and difficulties associated with an accurate and quantitative analysis of the evolution of multi-copy genes. Not surprisingly, the quality of the available data sources play a critical role. While the annotation of tRNA genes and pseudogenes does not seem to pose much of a problem, there are several issues limiting genome-wide multiple sequence alignments. On the one hand, coverage of alignable sequences can be a problem. In addition, with increasing phylogenetic distances, the fraction of aligned DNA decreases, hence conserved sequence anchors will become sparser, making the synteny approach less accurate. Even more problematic is the question whether aligned sequence blocks are really unique and thus are suitable as anchors. The differences in the results obtained with different anchor types indicate that genome-wide alignments provide far from perfect synteny anchors. Several factors seem to be critical. Most importantly, currently available multiple sequence alignment (MSA) pipelines do not explicitly filter for unique sequences before computing alignment chains [287]. Therefore, highly conserved paralogous sequences may lead to spurious anchors, which in turn lead to false correspondences between homologous tRNAs. The concept of uniquely mappable sequence intervals [380, 381], originally developed for high-throughput screening data analysis, probably could be adapted to the construction of genome-wide MSAs which provide more accurate anchor sets. This issue of ambiguities in MSAs needs to be addressed in future work as the development of new genome-wide alignment pipelines goes far beyond the scope of this work.

To showcase the framework, we reconstructed the evolution of tRNAs of human and six primates as well es of twelve drosophilids. We found that a large fraction of the tRNAs are recent copies. This proliferation is probably compensated by rapid pseudogenization, since we identified a large number of tRNA remolding events concentrated at the tips of the phylogeny. The developed workflow is applicable not only to tRNAs but also to other gene families evolving

under concerted evolution or birth-death evolution showing patterns of rapid duplications and losses. This encompasses several families of ncRNAs, as well as rapidly evolving gene families such as olfactory receptors or the Krüppel-associated box domain zinc finger protein (KRAB-ZNF) family in primate genes [286, 382]. As described in Berkemer et al. [377] we demonstrated that our workflow is also applicable for gene families like Y RNAs.

The natural transfer of DNA from mitochondria to the nucleus generates nuclear mitochondrial DNAs (NUMTs) and is an ongoing evolutionary process, as genome sequences attest [56]. Through the transposition of mitochondrial fragments, sequence copies of mitochondrial tRNAs (mt-tRNAs), referred to as nuclear-encoded mitochondrial-derived tRNAs (nm-tRNAs), have been integrated in the genomes of most eukaryotic organisms as parts of NUMTs [383]. Previously, nm-tRNAs have not been thoroughly investigated. Just one approach for the annotation of nm-tRNAs has been developed [24, 57] and merely speculations about their function have been made. As depicted in Chapter 12, we developed and compared different nm-tRNA annotation strategies that are much more sensitive (human: true positive rate (TPR) > 0.88; mouse: TPR > 0.71) and systematic than the published one (human: TPR 0.48; mouse: TPR 0.47). Finally, we identified 335 and 43 novel nm-tRNAs in human and mouse, respectively. To evaluate our performance we made use of synteny information which was determined from the primordial mitochondrial origin of the NUMTs. One benefit of considering synteny information is that we are able to filter FPs from our final set of nm-tRNAs. The most successful strategy was to limit the annotation range to published NUMT [373] sequences. In another approach we used whole genomic sequences for nm-tRNA annotation. Therewith, we found a high amount of hits outside of NUMT boundaries. Since hits outside from NUMTs can not be clearly assigned as nm-tRNAs we defined such hits as mt-tRNA-lookalikes. However, it cannot be ruled out that mt-tRNA-lookalikes are nm-tRNAs as we recognized that NUMT annotation is far from accurate. As the accuracy of our analysis strongly depends on a precise NUMT annotation, further work has to focus on improving NUMT annotation. Most NUMT sequences appear to be too shortly annotated, probably caused by their lowly conserved 5'- and 3'-ends. Therefore, NUMTs could be improved by reconsidering flanking regions which would contain adjacent mt-tRNA-lookalikes.

Finally, we found computational evidence that nm-tRNAs contain many functional binding sites for RNA-binding proteins (RBPs). Interestingly, intronic nm-tRNAs comprise an over-

representation of binding sites for splicing associated RBPs suggesting that nm-tRNA can acquire a function in the regulation of splicing given that the NUMT is inserted in a fortuitous intronic locations. As expected in such a scenario, most nm-tRNAs have not attained functional significance because they are simply not in a useful genomic context or there is no selective advantage gained from a nm-tRNA-derived binding site at the position of the insertion.

In summary, this thesis overcome hurdles in tRNA analysis and facilitated insights into novel aspects of tRNA biology on the basis of the combination of specialized deep sequencing approaches and sophisticated bioinformatic methods. This includes the systematic profiling and characterization of tRNA modifications in a trancriptome-wide scale which have not been detected using standard methods so far. The accurate mapping of tRNA reads and the optimization of tRNA-specific RNA-seq protocol was the basic prerequisite to make tRNA modifications clearly recognizable in RNA-seq data. Furthermore, conclusions about the evolution and biological relevance of individual tRNAs were made. Additionally, computational evidence of a novel potential function of nm-tRNAs in splicing was found suggesting that nm-tRNAs are more than molecular poltergeists.

Appendices

APPENDIX A

Additional Figures

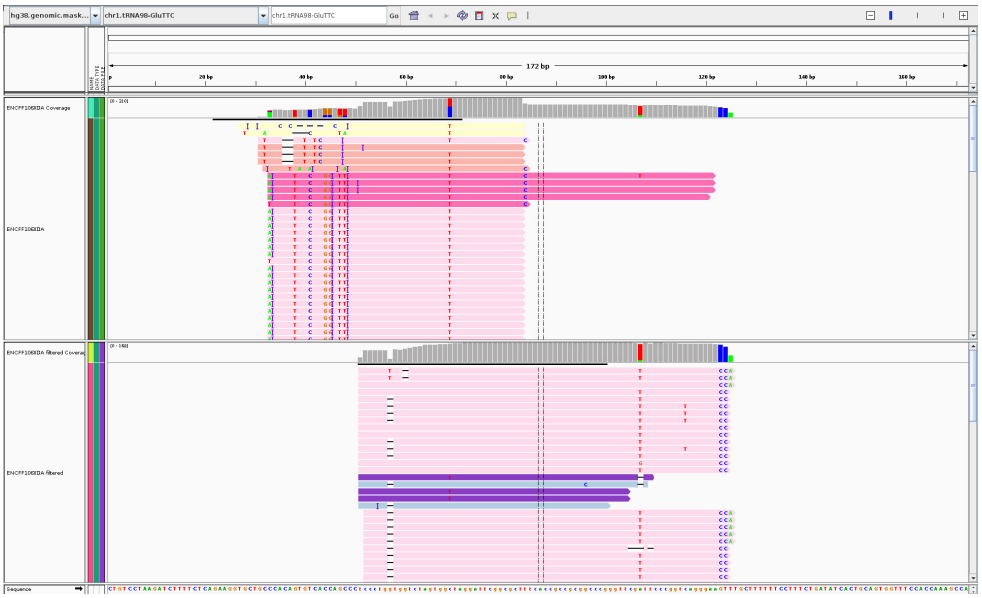

Figure A1: Genomic view of pre-tRNA filtering. Genomic view showing mapped reads from esophagus muscularis mucosae tissue to the glutamine (Glu) transfer RNA (tRNA) chr1.tRNA$^{Glu}_{TTC}$ after read mapping to the artificial genome. Unfiltered mapped reads (top) and precursor tRNA (pre-tRNA) filtered reads (bottom) are displayed. The pre-tRNAs mapping to the 5'-leader (residues 1-50) and 3'-trailer (residues 126-175) sequence introduce an additional error at position 69 at the tRNA (residue 51-125). Mismatches caused by mapping pre-tRNAs will be erroneously called as modification site. The exclusion of pre-tRNA reads therefore helps to reduce false positive hits.

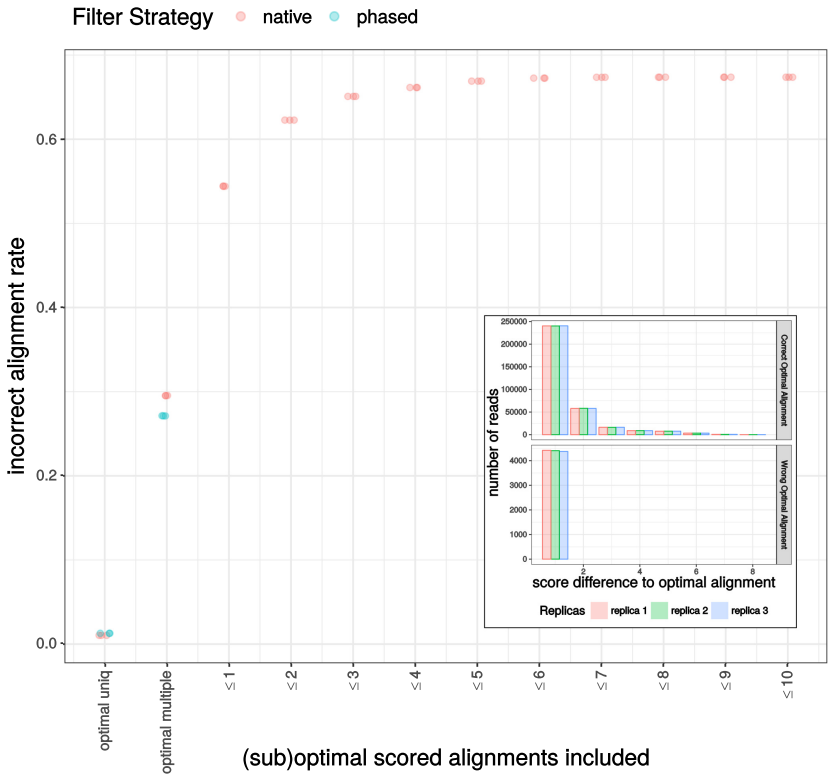

Figure A2: Incorrect alignment rates of suboptimal alignments. *Unique*ly mapped reads are, by definition, reads which map only to one location with the optimal alignment score. In the same line, multiply mapped reads are reads which map to more than one location with the same optimal score. The alignment score is measured as the edit distance between the reference genome and the read sequence. In addition, suboptimal read alignments can be used for variant calling. Given the high number of expected base misincorporations the true mapping location could also be amongst one of the suboptimally aligned positions. For the simulated reads used in this study it can be shown that doing so quickly increases the number of erroneously considered reads. If only optimal alignment positions of *unique*ly mapped reads are considered, ∼1% of the trusted alignments are mapped to the wrong position, and thus potentially produce wrong misincorporation patterns. If also multiple optimally aligned reads are allowed this number increases to ∼27% and ∼30%, for *all* and *phased* reads, respectively. If suboptimal read alignments are also included up to one mismatch worse than the optimal alignment, already ∼54% of the considered read alignment locations are misguided. This observation can be explained by the distribution of suboptimal alignment scores (see insert plot). Even though *all* reads whose optimal alignment did not correspond with the correct alignment had their true alignment only one mismatch away (insert, lower panel). The majority of reads whose optimal alignment was the correct one, had also the next suboptimal alignment only one mismatch away (insert, upper panel). So allowing for suboptimal read alignments up to one mismatch worse than the optimal alignment would rescue many correct alignments but for the cost of allowing two magnitude more wrong reads to be considered in the follow-up analysis.

Figure A3: Genomic view of the different filtering strategies using human RNA-seq data. Genomic view displaying mapped reads from the human tissue esophagus muscularis mucosae to one of the glutamine (Glu) transfer tRNA (tRNA) $tRNA^{Glu}_{TTC}$ clusters. The different read filter strategies (top) *unique*, (middle) *phased* and (bottom) *all* are shown. Misaligned reads can ultimately lead to wrong modification site calling, like at the positions 19, 34, 63, 64, and 66 for all mapped reads. Using *unique* filtered reads seems to be the most efficient method to reduce the misincorporation calling from false positives and false negatives caused by multiply mapped reads.

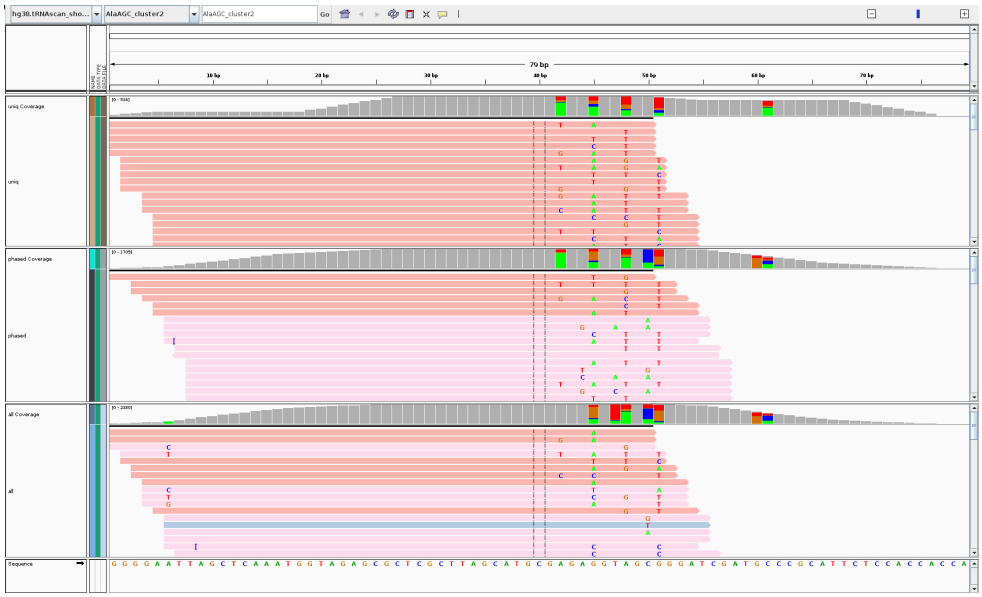

Figure A4: Genomic view of the different filtering strategies using simulated data. Genomic view displaying mapped reads from the simulated data containing *identical modifications* to one of the alanine (Ala) transfer RNA (tRNA) $tRNA^{Ala}_{AGC}$ clusters. The different read filter strategies (top) *unique*, (middle) *phased* and (bottom) *all* are shown. Misaligned reads can ultimately lead to wrong modification site calling, like at the positions 50 and 60 from *phased* filtered reads and the positions 47, 50 and 60 by *all* mapped reads. The position 42 will not be called in comparison to the two other filtering steps. Using *unique* filtered reads seems to be the most efficient method to reduce the misincorporation calling from false positives and false negatives caused by multiply mapped reads.

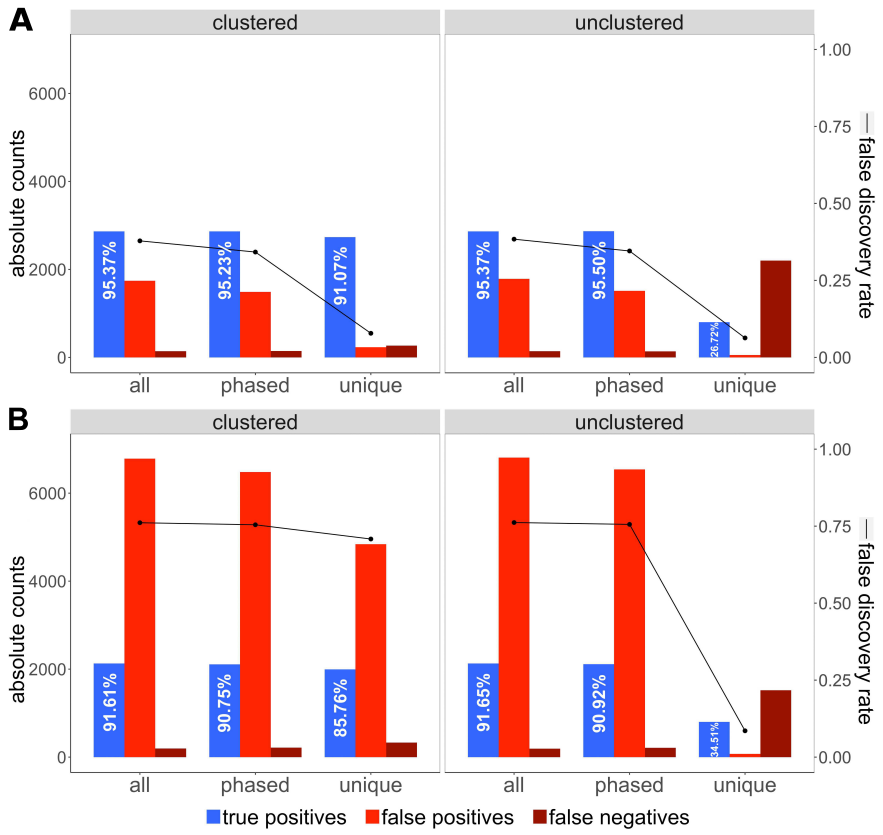

Figure A5: Evaluation of the `Pfropfen` variation caller. Absolute numbers of counted true positives (TPs; blue), false positives (FPs; red) and false negatives (FNs; dark red) by analysing simulated reads containing **(A)** *identical modification* or **(B)** *random modification* sites. For the *unclustered* (r.h.s.) as well as the *clustered method* (l.h.s.), the results of the different read filter strategies (*all*, *phased*, *unique*) are shown, respectively. Significant misincorporation sites for each filtering step were called using `Pfropfen`. For both simulated data sets, the ratio between the true positive rate and the false positive rate is balanced in favor of the *unique*ly mapped reads. The *unique* filtered reads are much more sensitive in regard to the *clustered method*, then the *unclustered method*.

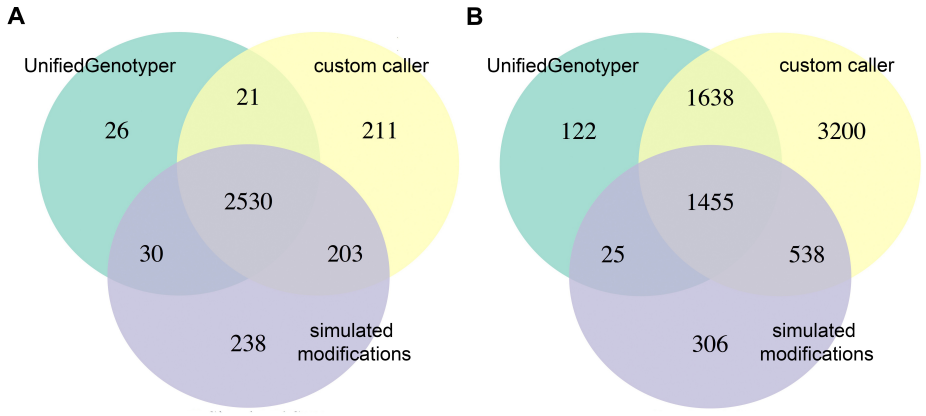

Figure A6: Comparison of two different modification caller. Overlap of the (purple) simulated transfer RNA (tRNA) modifications with the called modification sites resulting from the best-practice tRNA read mapping method using (green) `GATK`'s `UnifiedGenotyper` and `Pfropfen` (yellow). The analysis of the simulated reads containing *identical modification* sites is shown in **(A)**, in which 2,530 (84,3%) of the 3,001 generated modification sites were detected using both variation callers. Im comparison to that, in the analysis of the simulated reads containing *random modifications* **(B)** 1,455 (62,6%) modifications of the simulated 2,324 sites were detected overlapping both caller methods. The `UnifiedGenotyper` detected much less true positives than the custom *ad hoc* method, but also shows a reduced set on false positives. Thus, `Pfropfen` seems to be more sensitive at the expense of reduced specificity.

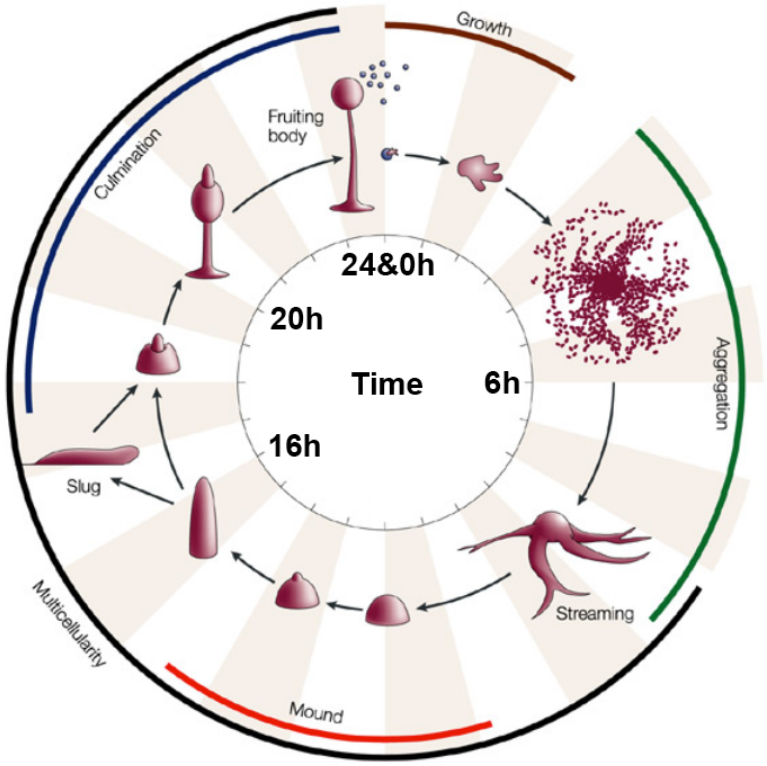

Figure A7: Life cycle of *Dictyostelium discoideum*. Developmental morphogenesis of *D. discoideum* starting from a single and vegetative amoebae (0h). Aggregation of the single amoebae is mediated by the chemotaxis of cells to form a multicellular aggregate (6h after starvation). During this process, multicellular aggregate streams toward a central domain or aggregation center. Aggregation results in the formation of a mound (multicellular organism, 12h after starvation). Mound forms then a tipped mound (14h after starvation). The tip extends and forms a finger which might fall over to form a phototactic migrating slug (16h after starvation) or begins culuminations (20h after starvation) to form a fruiting body. Finally, the fruiting body contains a sorus of spores on top of a stalk which germinate following dispersal, renewing the cycle (24h after starvation). For our analysis described in Section 10.2.1 we used cells from 0h, 6h, 16h, 20h and 24h after starvation. The figure is modified from Chisholm et al. [351].

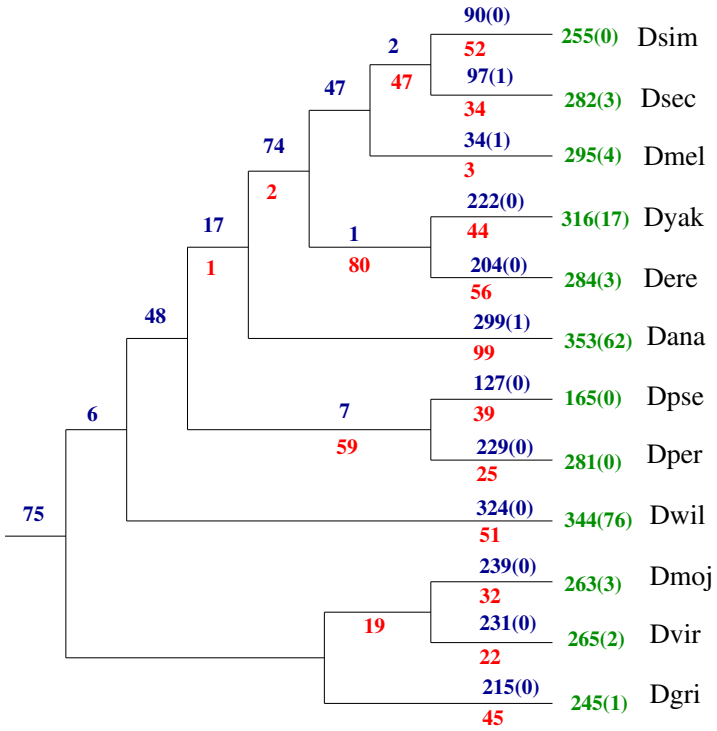

Figure A8: Gain, loss, and duplications of tRNAs in drosophilids. Gain, loss, and duplications of transfer RNAs (tRNAs) in drosophilids computed from the most fine-grained synteny definition based on individual multiple sequence alignment (MSA) blocks. Gain and duplication events were assigned to the edge leading to the last common ancestor of all observed co-orthologs, except for groups that contained only one tRNA sequence of two species; in these cases we assigned two lineage-specific gains. Green numbers refer to the total number of tRNAs detected by `tRNAscan-SE`; green numbers in parentheses count the pseudogenes found in the set of all tRNAs. Blue numbers refer to the total gain, i.e., the sum of event seeding new connected components and duplication events with a connected component. The number of identified local duplication events is given in parentheses in blue. The red numbers indicate the loss events on the corresponding branch. Species abbreviations: *Drosophila simulans*: Dsim; *Drosophila sechellia*: Dsec; *Drosophila melanogaster*: Dmel; *Drosophila yakuba*: Dyak; *Drosophila erecta*: Dere; *Drosophila ananassae*: Dana; *Drosophila pseudoobscura*: Dpse; *Drosophila persimilis*: Dper; *Drosophila willistoni*: Dwil; *Drosophila mojavensis*: Dmoj; *Drosophila virilis*: Dvir; *Drosophila grimshawi*: Dgri.

APPENDIX B

Additional Tables

Table B1: Overview of tRNA modifications. Collection of all 93 annotated transfer RNA (tRNA) modifications. Their symbols and common names are listed. The data are collected from the tRNAmodviz database [119].

Abbreviation	Common name
ac^{4}C	N^4-acetylcytidine
ac^{4}Cm	N^4-acetyl-2'-O-methylcytidine
ac^{6}A	N^6-acetyladenosine
acp^{3}U	3-(3-amino-3-carboxypropyl)uridine
Am	2'-O-methyladenosine
Ar(p)	2'-O-ribosyladenosine (phosphate)
C$^+$	agmatidine
chm^{5}U	5-(carboxyhydroxymethyl)uridine
Cm	2'-O-methylcytidine
cmnm^{5}s^{2}U	5-carboxymethylaminomethyl-2-thiouridine
cmnm^{5}U	5-carboxymethylaminomethyluridine
cmnm^{5}Um	5-carboxymethylaminomethyl-2'-O-methyluridine
cmo^{5}U	uridine 5-oxyacetic acid
D	dihydrouridine
f^{5}C	5-formylcytidine
f^{5}Cm	5-formyl-2'-O-methylcytidine
G$^+$	archaeosine
g^{6}A	N^6-glycinylcarbamoyladenosine
galQ	galactosyl-queuosine
Gm	2'-O-methylguanosine
Gr(p)	2'-O-ribosylguanosine (phosphate)
hn^{6}A	N^6-hydroxynorvalylcarbamoyladenosine
ho^{5}U	5-hydroxyuridine
I	inosine
i^{6}A	N^6-isopentenyladenosine

Continued on next page

Table B1 – continued from previous page

Abbreviation	Common name
imG	wyosine
imG-14	4-demethylwyosine
imG2	isowyosine
inm^{5}s^{2}U	5-(isopentenylaminomethyl)-2-thiouridine
inm^{5}U	5-(isopentenylaminomethyl)uridine
inm^{5}Um	5-(isopentenylaminomethyl)-2'-O-methyluridine
io^{6}A	N^6-(cis-hydroxyisopentenyl)adenosine
k^{2}C	lysidine
m^{1}A	1-methyladenosine
m^{1}Am	1,2'-O-dimethyladenosine
m^{1}G	1-methylguanosine
m^{1}Gm	1,2'-O-dimethylguanosine
m^{1}I	1-methylinosine
m^{1}Im	1,2'-O-dimethylinosine
m$^1\Psi$	1-methylpseudouridine
m^{2,7}Gm	N^2,7,2'-O-trimethylguanosine
m$^2{}_2$G	N^2,N^2-dimethylguanosine
m$^2{}_2$Gm	N^2,N^2,2'-O-trimethylguanosine
m^{2}A	2-methyladenosine
m^{2}G	N^2-methylguanosine
m^{2}Gm	N^2,2'-O-dimethylguanosine
m^{3}C	3-methylcytidine
m$^4{}_2$Cm	N^4,N^2,2'-O-trimethylcytidine
m^{5}C	5-methylcytidine
m^{5}Cm	5,2'-O-dimethylcytidine
m^{5}s^{2}U	5-methyl-2-thiouridine

Continued on next page

Table B1 – continued from previous page

Abbreviation	Common name
m^5U	5-methyluridine
m^5Um	5,2'-O-dimethyluridine
m^6A	N^6-methyladenosine
m^6t^6A	N^6-methyl-N^6-threonylcarbamoyladenosine
m^7G	7-methylguanosine
manQ	mannosyl-queuosine
$mchm^5U$	5-(carboxyhydroxymethyl)uridine methyl ester
mcm^5s^2U	5-methoxycarbonylmethyl-2-thiouridine
mcm^5U	5-methoxycarbonylmethyluridine
mcm^5Um	5-methoxycarbonylmethyl-2'-O-methyluridine
$mcmo^5U$	uridine 5-oxyacetic acid methyl ester
mimG	methylwyosine
mnm^5s^2U	5-methylaminomethyl-2-thiouridine
mnm^5se^2U	5-methylaminomethyl-2-selenouridine
mnm^5U	5-methylaminomethyluridine
mo^5U	5-methoxyuridine
ms^2hn^6A	2-methylthio-N^6-hydroxynorvalyl carbamoyladenosine
ms^2i^6A	2-methylthio-N^6-isopentenyladenosine
ms^2io^6A	2-methylthio-N^6-(cis-hydroxyisopentenyl) adenosine
ms^2m^6A	2-methylthio-N^6-methyladenosine
ms^2t^6A	2-methylthio-N^6-threonyl carbamoyladenosine
ncm^5U	5-carbamoylmethyluridine
ncm^5Um	5-carbamoylmethyl-2'-O-methyluridine
nm^5s^2U	5-aminomethyl-2-thiouridine
o_2yW	peroxywybutosine
OHyW	hydroxywybutosine

Continued on next page

Table B1 – continued from previous page

Abbreviation	Common name
OHyW*	undermodified hydroxywybutosine
oQ	epoxyqueuosine
preQ0	7-cyano-7-deazaguanosine
preQ1	7-aminomethyl-7-deazaguanosine
Q	queuosine
s_2C	2-thiocytidine
s_2U	2-thiouridine
s_2Um	2-thio-2'-O-methyluridine
s_4U	4-thiouridine
t_6A	N^6-threonylcarbamoyladenosine
πm^5s_2U	5-taurinomethyl-2-thiouridine
πm^5U	5-taurinomethyluridine
Um	2'-O-methyluridine
Ψ	pseudouridine
Ψm	2'-O-methylpseudouridine
yW	wybutosine

Table B2: Mapping signatures of common tRNA modifications. Chemical transfer RNA (tRNA) modifications and their possible detection by analyzing RNA sequencing (RNA-seq) data are listed. Modifications become visible in RNA-seq data as accumulation of base-calling errors, read terminations (RTs), and/or by chemically treated RNA-seq data. Typically, thymine modifications cannot be detected by analyzing base-calling errors. However, we observed thymine-to-cytosine transitions or thymine-to-adenine transversions at single tRNA positions which are known sites for dihydrouridine modifications. Common names of modification abbreviations are listed in **Suppl. Tab. B1**. Nulceobase abbreviations: A – adenine; C – cytosine; G – guanine; T – thymine.

Modification	Alteration	RTs	Treatment
Am	-	-	+
m^1A	A→(C|G|T)	+	-
I	A→G	-	+

Continued on next page

Table B2 – continued from previous page

Modification	Alteration	RTs	Chemical Treatment
m^1G	G→(A|C|T)	+	-
m^2G	G→(A|C|T)	+	-
$m^2{}_2G$	G→(A|C|T)	+	-
m^7G	-	-	+
Um	-	-	+
m^5U	-	-	+
D	T→(A|C|G)?	+	+
Ψ	-	-	+
$m^1\Psi$	-	-	+
Cm	C→(A|T)	+	+
m^3C	C→(A|G|T)	+	+
m^5C	-	-	+
i^6A	A →(G|T)	+	-
m^1I	A→G	-	+
o_2yW	G→(A|C|T)	-	-
t_6A	A→(C|G|T)	+	-
ms^2i^6A	A→(C|G|T)	+	-

Table B3: Number of annotated tRNA genes per species. The number of annotated cytosolic, pseudo or mitochondrial transfer RNAs (tRNAs) are listed for each species.

Species	# cytosolic tRNAs	# pseudo tRNAs	# mitochondrial tRNAs
Dictyostelium discoideum	403	4	18
Drosophila ananassae	304	165	–
Drosophila erecta	281	3	–
Drosophila grimshawi	258	1	–
Drosophila melanogaster	291	4	–

Continued on next page

Table B3 – continued from previous page

Species	**# cytosolic tRNAs**	**# pseudo tRNAs**	**# mitochondrial tRNAs**
Drosophila mojavensis	261	3	–
Drosophila persimilis	297	1	–
Drosophila pseudoobscura	293	1	–
Drosophila sechellia	296	13	–
Drosophila simulans	264	3	–
Drosophila virilis	267	2	–
Drosophila willistoni	291	166	–
Drosophila yakuba	324	51	–
Escherichia coli	88	1	–
Geobacillus stearothermophilus	62	1	–
Gorilla gorilla gorilla	431	87	–
Homo sapiens	732	98	22
Macaca mulatta	463	115	–
Nomascus leucogenys	441	119	–
Pan troglodytes	531	108	–
Pongo abelii	543	118	–
Saccharomyces cerevisiae	189	1	18
Spinacia oleracea	2111	450	24

Table B4: Comparison of LOTTE-seq with common tRNA-seq methods. Transfer RNA (tRNA) content of the long hairpin oligonucleotide based tRNA high-throughput sequencing (LOTTE-seq) method compared with the optimized TruSeq sRNA protocol as well as with the standard Illumina TruSeq sRNA procedure is listed. In particular, the number of reads mapped to tRNAs with a 3'-CCA-end, with 3'-ends other than a CCA (such as partial CCA-ends or tRNAs without a CCA-end) and non-tRNA reads are given in percentage for six different species. Additionally, the percentage of reads mapped to tRNAs as well as other genomic regions (ambiguous tRNAs) are listed. In this case, the true origin of the read is indeterminable. Two replicates (rep) of MiSeq-based sequence analyses of each species were investigated for each RNA-seq method. In all investigations, LOTTE-seq shows the highest amount of CCA-containing tRNA reads. Species abbreviations: *Dictyostelium discoideum*: Ddi; *Escherichia coli*: Eco; *Geobacillus stearothermophilus*: Gst; *Homo sapiens*: Hsa; *Saccharomyces cerevisiae*: Sce; *Spinacia oleracea*: Sol.

Species	Method	Rep.	tRNAs 3'-CCA end	tRNAs other 3'-end	Ambiguous tRNAs	Other RNAs
Ddi	LOTTE-seq	1	99.23	0.10	0.34	0.33
Ddi	LOTTE-seq	2	98.77	0.26	0.39	0.57
Ddi	opt. TruSeq	1	73.32	22.95	1.07	2.65
Ddi	opt. TruSeq	2	74.93	21.03	1.01	3.03
Ddi	TruSeq	1	3.60	0.80	0.14	95.47
Ddi	TruSeq	2	3.14	1.22	0.00	95.64
Eco	LOTTE-seq	1	99.12	0.25	0.02	0.61
Eco	LOTTE-seq	2	98.04	0.35	0.03	1.58
Eco	opt. TruSeq	1	77.61	8.55	0.00	13.84
Eco	opt. TruSeq	2	79.31	9.84	0.01	10.83
Eco	TruSeq	1	5.37	0.41	0.19	94.03
Eco	TruSeq	2	5.92	0.28	0.00	93.80
Gst	LOTTE-seq	1	97.77	0.31	0.02	1.90
Gst	LOTTE-seq	2	97.60	0.22	0.04	2.13
Gst	opt. TruSeq	1	74.28	11.66	0.03	14.03
Gst	opt. TruSeq	2	72.48	13.90	0.09	13.53
Gst	TruSeq	1	4.34	0.83	0.07	94.76
Gst	TruSeq	2	1.81	0.34	0.00	97.85
Hsa	LOTTE-seq	1	55.83	2.08	37.18	4.90

Continued on next page

Table B4 – continued from previous page

Species	Method	Rep.	tRNAs 3'-CCA end	tRNAs other 3'-end	Ambiguous tRNAs	Other RNAs
Hsa	LOTTE-seq	2	60.43	2.27	33.79	3.50
Hsa	opt. TruSeq	1	28.76	3.43	22.38	45.43
Hsa	opt. TruSeq	2	29.09	2.28	26.83	41.81
Hsa	TruSeq	1	0.23	0.21	1.46	98.10
Hsa	TruSeq	2	0.92	0.11	1.91	97.07
Sce	LOTTE-seq	1	98.44	0.18	0.29	1.08
Sce	LOTTE-seq	2	97.65	0.25	1.50	0.61
Sce	opt. TruSeq	1	53.12	15.58	1.11	30.18
Sce	opt. TruSeq	2	41.40	43.45	1.06	14.09
Sce	TruSeq	1	2.11	3.72	0.51	93.66
Sce	TruSeq	2	3.03	4.59	0.61	91.76
Sol	LOTTE-seq	1	75.50	0.25	15.77	8.48
Sol	LOTTE-seq	2	81.20	0.36	8.55	9.88
Sol	opt. TruSeq	1	65.87	5.20	14.95	13.98
Sol	opt. TruSeq	2	63.16	4.96	13.09	18.79
Sol	TruSeq	1	9.09	1.15	1.97	87.80
Sol	TruSeq	2	6.45	3.23	0.00	90.32

Table B5: Tissue-specific differences regarding the number of modified tRNAs in human. Number of modified human transfer RNAs (tRNAs) in the six investigated tissues (cerebellum (C), diencephalon (D), ovary (O), skeletal muscle (S), esophagus muscularis mucosae (E), and testis (T)) are listed. The tRNA modifications are detected by accumulations base-calling errors. The same number of modified tRNAs is identified only at position 49 for each tissue. The number of modified tRNAs varies greatly at the other positions between tissues. Only seven positions are frequently modified in each tissue. Abbreviations: A – adenine; AC – anticodon; ACC – acceptor; C – cytosine; D – dihydrouridine; G – guanine; T – thymine V – variable.

Area	Position	Alteration	# O	# T	# C	# D	# E	# S
5'-ACC-stem	6	G→(C|T)	12	12	-	-	14	-
		T→A	-	-	-	-	1	-
variable region	9	A→(C|G|T)	21	22	10	8	21	5
		G→(A|T)	19	13	19	26	30	25
5'-D-stem	10	G→A	-	-	-	1	1	-
D-loop	18	T→C	-	3	-	-	-	-
	19	C→(A|T)	-	-	-	-	11	-
3'-D-stem	22	G→T	-	-	-	-	-	5
	23	A→C	-	-	6	-	-	9
	24	A→(C|T)	-	-		6	-	-
	25	C→(A)	2	2	-	-	2	-
variable region	26	G→(A|C|T)	35	66	31	14	95	9
5'-AC-stem	31	C→T	-	-	-	-	1	-
		A→G	-	-	10		-	-
AC-loop	32	C→(A|T)	6	7	6	4	10	6
	34	A→G	18	20	38	30	40	38
	37	A→(G|T)	5	2	23	23	27	23
		G→T	19	19	12	-	8	-
V-region	12e	G→C	1	-	-	-	1	1
	2e	C→(A|G|T)	10	8	4	2	12	-
5'-TΨC-stem	49	A→T	1	1	1	1	1	1
	52	G→C	-	-	-	-	1	-

Continued on next page

Table B5 – continued from previous page

Area	Position	Alteration	# O	# T	# C	# D	# E	# S
TΨC-loop	57	T→C	-	1	-	-	-	-
	58	A→(G|T)	175	171	196	119	244	115
3'-TΨC-stem	61	C→A	-	-	-	1	-	-
	65	G→T	-	1	-	13	-	-

Table B6: Differences regarding the number of modified tRNAs during the life cycle of *Dictyostelium discoideum*. Numbers of modified transfer RNAs (tRNAs) during the developmental stages (0h, 6h, 16h, 20h, and 24h after starvation) of *Dictyostelium discoideum* are listed. The modifications were identified based on the interpretation of base-calling errors. At each stage of development, the same tRNA positions are modified, but vary greatly in the number of modified tRNAs. Only individual modifications at positions 34, 37 and 68 show the same number of modified tRNAs at each investigated developmental stage. Modified tRNAs at 6h after starvation varies most from the other stages. Abbreviations: A – adenine; AC – anticodon; C – cytosine; D – dihydrouridine; G – guanine; T – thymine; V – variable.

Area	Position	Alteration	# Oh	# 6h	# 16h	# 20h	# 24h
-	9	G→(C|T)	71	47	74	36	38
D-loop	20	T→C	4	62	15	-	-
		C→(A|T)	35	2	35	35	35
-	26	G→(A|T)	248	205	248	247	248
AC-loop	32	C→T	61	61	61	61	61
	34	A→G	91	91	91	91	91
		C→T	1	1	1	1	1
	37	A→T	23	15	23	23	23
		G→(A|C|T)	53	39	53	52	47
V-region	47	T→(A|C)	42	26	26	24	23
TΨC-loop	58	A→(C|G|T)	401	400	401	401	401
3'-TΨC-stem	68	A→G	1	1	1	1	1

Table B7: Remolding events in primates in detail. All annotated transfer RNA (tRNA) isoacceptor and alloacceptor events and the primates in which they were found are listed in detail. Remolding events were grouped corresponding to their annotation origin: found in both our and the previous study [197] (common), only in the previous study (Rogers (2014)) and only in our study (novel). Our method is more sensitive and predicts more tRNA remolding events. The conventional 3-letter code is used as abbreviation for each tRNA type: Ala – alanine; Arg – arginine; Asn – asparagine; Asp – aspartic acid; Cys – cysteine; Gln – glutamine; Glu – glutamic acid; Gly – glycine; His – histidine; Ile – isoleucine; Leu – leucine; Lys – lysine; Met – methionine; Phe – phenylalanine; Pro – proline; Thr – threonine; Trp – tryptophan; Tyr – tyrosine; Ser – serine; Val – valine. Species abbreviations: human, *Homo sapiens*: Hsa; chimapanzee, *Pan troglodytes*: Ptr; gorilla, *Gorilla gorilla gorilla*: Ggo; orangutan, *Pongo abelii*: Pab; gibbon, *Nomascus leucogenys*: Nle; rhesus macaque, *Macaca mulatta*: Mmu. Nulceobase abbreviations: A – adenine; C – cytosine; G – guanine; T – thymine.

Remolding event	Species	Source
Isoacceptor		
Ala(GGC→AGC)	Ggo	common
Ala(TGC→CGC)	Ggo, Hsa, Pab,	common
Ala(AGC→CGC)	Ptr	novel
Ala(TGC→AGC)	Nle, Mmu	novel
Asn(ATT→GTT)	Ggo, Nle	common
Cys(ACA→GCA)	Ggo, Nle	common
Glu(TTC→CTC)	Pab	common
Gly(TCC→CCC)	Nle	common
Leu(CAA→CAG)	Hsa	common
Leu(CAG→TAG)	Mmu	novel
Pro(AGG→GGG)	Nle	novel
Pro(TTG→CTG)	Mmu	novel
Ser(TGA→AGA)	all	common
Thr(CGT→TGT)	Mmu, Pab, Ptr	common
Thr(AGT→CGT)	Mmu, Pab, Ptr	novel
Thr(AGT→TGT)	Mmu, Nle, Pab Ptr	novel
Val(AAC→TAC)	Ptr	novel

Continued on next page

Table B7 – continued from previous page

Remolding event	Species	Source
Val(CAC→TAC)	Nle	novel
Alloacceptor		
Ala(AGC)→Gly(TCC)	Pab	novel
Arg(TCG)→SeC(TCA)	Ggo, Hsa, Pab, Ptr	novel
Arg(GCG)→Cys(GCA)	–	Rogers (2014)
Arg(GCG)→His(GTG)	Ptr	common
Arg(CCG)→Gly(CCC)	Pab	common
Asp(GTC)→Asn(GTT)	Pab	novel
Cys(GCA)→His(GTG)	Pab	novel
Gln(TTG)→Arg(TCG)	Mmu	common
Gln(TTG)→Glu(TTC)	Pab	novel
Gln(CTG)→Glu(CTC)	Mmu	novel
Gln(TTG)→Pro(TGG)	Mmu	novel
Gln(TTG)→SeC(TCA)	Ggo	novel
Glu(TTC)→Gly(TCC)	Hsa, Mmu	common
Glu(TTC)→Lys(TTT)	Ptr, Ggo	common
Glu(CTC)→Ala(TGC)	–	Rogers (2014)
His(GTG)→Gln(TTG)	Mmu	novel
Ile(GAT)→Phe(GAA)	Nle	common
Ile(AAT)→Ser(ACT)	Ptr	novel
Leu(CAA)→Ser(CGA)	Ptr	common
Leu(CAA)→Met(CAT)	Ggo	common
Lys(CTT)→Asn(GTT)	Mmu, Pab	common
Lys(CTT)→Arg(CCT)	Mmu	novel
Met(CAT)→Thr(CGT)	Hsa	novel

Continued on next page

Table B7 – continued from previous page

Remolding event	Species	Source
Met(CAT)→Ile(TAT)	Ggo	novel
Phe(GAA)→Ser(GCA)	Ggo	novel
Phe(GAA)→Ser(GCA)	Mmu	novel
SeC(TCA)→Cys(GCA)	–	Rogers (2014)
Ser(AGA)→Cys(GCA)	Pab	common
Ser(ACT)→Ile(AAT)	Hsa	common
Thr(CGT)→Met(CAT)	Hsa	common
Thr(TGT)→Ile(TAT)	Mmu	novel
Tyr(GTA)→Cys(GCA)	all (deleted in Nle)	common
Val(TAC)→Ile(TAT)	Ptr, Ggo	common
Val(AAC)→Ala(AGC)	Hsa	common
Val(TAC)→Leu(TAA)	Ptr	common
Val(AAC)→Ile(AAT)	Nle	novel
Val(CAC)→Gly(CCC)	Ptr	common

Table B8: Remolding events in drosophilids in detail. All annotated transfer RNA (tRNA) isoacceptor and alloacceptor events and the drosophilids in which they were found are listed in detail. Remolding events were grouped corresponding to their annotation origin: found in both our and the previous studies [196] (common), only in the previous studies (Rogers (2010)) and only in our study (novel). Our method is more sensitive and predicts more tRNA remolding events. The conventional 3-letter code is used as abbreviation for each tRNA type: Ala – alanine; Arg – arginine; Asn – asparagine; Asp – aspartic acid; Cys – cysteine; Gln – glutamine; Gly – glycine; His – histidine; Ile – isoleucine; Leu – leucine; Lys – lysine; Met – methionine; Pro – proline; Thr – threonine; Tyr – tyrosine; Ser – serine. Species abbreviations: *Drosophila simulans*: Dsim; *Drosophila sechellia*: Dsec; *Drosophila melanogaster*: Dmel; *Drosophila yakuba*: Dyak; *Drosophila erecta*: Dere; *Drosophila ananassae*: Dana; *Drosophila pseudoobscura*: Dpse; *Drosophila persimilis*: Dper; *Drosophila willistoni*: Dwil; *Drosophila mojavensis*: Dmoj; *Drosophila virilis*: Dvir; *Drosophila grimshawi*: Dgri. Nulceobase abbreviations: A – adenine; C – cytosine; G – guanine; T – thymine.

Remolding event	Species	Source
Isoacceptor		
Ala(AGC→TGC)	all	novel
Arg(ACG→TCG)	Dgri	common
Arg(CCT→TCT)	Dsec	common
Arg(TCG→TCT)	Dsim, Dsec, Dmel, Dyak, Dere, Dpse, Dper, Dana	common
Arg(TCG→CCG)	Dwil	novel
Cys(GCA→ACA)	Dana, Dere	common
Cys(GCA→ACA)	–	Rogers (2010)
Gly(GCC→CCC)	Dsim, Dsec, Dmel, Dyak, Dere, Dpse, Dper, Dana, Dwil	common
Ile(AAT→GAT)	Dgri	novel
Leu(AAG→TAG)	all	novel
Lys(CTT→TTT)	Dyak	novel
Pro(AGG→CGC)	all	common
Pro(AGG→TGG)	all	common
Ser(TGA→AGA)	all	common

Continued on next page

Table B8 – continued from previous page

Remolding event	Species	Source
Alloacceptor		
Asn(GTT)→Lys(TTT)	Dana	common
Asn(GTT)→Thr(GGT)	Dper	novel
Asp(GTC)→Asn(GTT)	–	Rogers (2010)
Lys(CTT)→Met(CAT)	Dsec	common
Met(CAT)→Leu(CAA)	Dmoj	novel
Met(CAT)→Thr(CGT)	Dsim	common
Ser(GCT)→Asn(GTT)	Dana	novel
Tyr(GTA)→His(GTG)	Dsec	common

Table B9: Protein binding sites of nim-tRNAs. The relative enrichments over expected nucleotide coverage of RNA-binding protein (RBP) binding events in human nuclear-encoded intronic mitochondrial-derived transfer RNAs (nim-tRNAs) over background are listed. For the background we calculated the expected coverage of RBP binding sites per nucleotide in all human introns. The 31 RBPs with overlapping nim-tRNAs have a splicing function or other regulatory roles.

Cellline	RBP	RBP sites in introns	Intron coverage	RBP sites in nim-tRNAs	Fold enrichment
HepG2	DDX52	192756	0.0041284	1	0.2699093
HepG2	DDX6	36663	0.0007853	1	1.4190501
HepG2	DHX30	95293	0.0020409	4	2.18385956
HepG2	DROSHA	386995	0.0082885	2	0.2688750
HepG2	EIF3D	321832	0.0068929	1	0.1616578
HepG2	FASTKD2	120719	0.0025855	1	0.4309731
HepG2	G3BP1	24087	0.0005159	1	2.1599466
HepG2	HLTF	270375	0.0057908	1	0.1924240
HepG2	HNRNPA1	590724	0.0126519	1	0.0880727
HepG2	HNRNPC	880024	0.0188489	2	0.1182391
HepG2	HNRNPL	571862	0.0122479	2	0.1819552

Continued on next page

Table B9 – continued from previous page

Cellline	RBP	RBP sites in introns	Intron coverage	RBP sites in nim-tRNAs	Fold enrichment
HepG2	HNRNPU	384074	0.0082259	1	0.1354600
HepG2	QKI	294484	0.0063071	4	0.7066820
HepG2	RBM22	169577	0.0036319	1	0.3068024
HepG2	RBM5	230897	0.0049452	1	0.2253240
HepG2	UCHL5	201213	0.0043095	1	0.2585650
HepG2	ZC3H11A	326633	0.0069957	1	0.1592817
K562	CSTF2T	215118	0.0046073	1	0.2418516
K562	EXOSC5	163787	0.0035079	2	0.6352962
K562	FASTKD2	55225	0.0011828	1	0.9420848
K562	HLTF	157472	0.0033727	2	0.6607731
K562	HNRNPL	527655	0.0113011	2	0.1971994
K562	HNRNPU	706052	0.0151219	1	0.0736867
K562	ILF3	351169	0.0075212	1	0.1481526
K562	KHDRBS1	509365	0.010909	4	0.4085607
K562	NSUN2	19825	0.0004246	1	2.6242942
K562	PUS1	26788	0.0005737	1	1.9421619
K562	QKI	77069	0.0016506	1	0.6750656
K562	SAFB2	383009	0.0082031	2	0.2716732
K562	TAF15	184423	0.0039499	1	0.2821049
K562	ZC3H11A	99807	0.0021376	1	0.5212724

List of Abbreviations

Ψ	pseudouridine
A	adenine
ADAR	adenosine deaminase acting on RNA
ANG	angiogenin
ARM-seq	AlkB-facilitated RNA methylation sequencing
BDP1	TFIIB double prime 1
BLOSUM	blocks substitution matrix
bp	base pairs
BRF1	TFIIB-related factor 1
C	cytosine
cAMP	cyclic AMP
cDNA	complementary DNA
CM	covariance model
Cm	2'-O-methylcytidine
CMCT	1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene

CTNNBL1	catenin–ß–like 1
Cys	cysteine
D	dihydrouridine
D-arm	dihydrouridine arm
DM-tRNA-seq	Demethylase-thermostable group II intron RT tRNA sequencing
DNA	deoxyribonucleic acid
dT	oligo
ESA	enhanced suffix array
FC	fold change
FDR	false discovery rate
fMet	N-formylmethionyl
FN	false negative
FNR	false negative rate
FP	false positive
G	guanine
hm^{5}C	5-hydroxymethylcytidine
HSP	high-scoring segment pair
I	inosine
i^{6}A	N^6-isopentenyladenosine
indel	insertion or deletion
INM	inner nuclear membrane

k^{2}C	lysidine
kbp	kilo base pairs
KRAB-ZNF	Krüppel-associated box domain zinc finger protein
Leu	leucine
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA
LOTTE-seq	long hairpin oligonucleotide-based tRNA high-throughput sequencing
m^{1}A	1-methyladenosine
m^{1}G	1-methylguanosine
m^{1}I	1-methylinosine
m$^2{}_2$G	N^2,N^2-dimethylguanosine
m^{2}G	N^2-methylguanosine
m^{3}C	3-methylcytidine
m^{5}C	5-methylcytidine
m^{5}U	5-methyluridine
m^{6}A	N^6-methyladenosine
m^{7}G	7-methyl-guanosine
mcm^{5}s^{2}U	5-methoxycarbonylmethyl-2-thiouridine
mcm^{5}Um	5-methoxycarbonylmethyl-2'-O-methyluridine
miRNA	microRNA
mRNA	messenger RNA
mRNA-seq	poly-A-selected RNA-sequencing
ms^{2}i^{6}A	2-methylthio-N^6-isopentenyladenosine
MSA	multiple sequence alignment

mt-DNA	mitochondrial DNA
mt-genome	mitochondrial genome
mt-tRNA	mitochondrial tRNA
mtRPOL	mitochondrial RNA polymerase
$NaBH_4$	sodium borohydride
NAD	nicotinamide adenine dinucleotide
ncRNA	non-coding RNA
NGS	next-generation sequencing
nim-tRNA	nuclear-encoded intronic mitochondrial-derived tRNA
nm-tRNA	nuclear-encoded mitochondrial-derived tRNA
NP	nondeterministic polynomial time
nt	nucleotide
NUMT	nuclear mitochondrial DNA
o_2yW	peroxywybutosine
PAM	point accepted mutation
PCR	polymerase chain reaction
PhyloP	phylogenetic p-value
PNPase	polynucleotide phosphorylase
Pol III	RNA polymerase III
poly-A	polyadenylated
pre-mRNA	precursor mRNA
pre-tRNA	precursor tRNA
Q	queuosine

RBP	RNA-binding protein
rmRNA-seq	ribo-minus RNA sequencing
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RNase	ribonuclease
RNase Z	tRNA 3'-endonuclease
rRNA	ribosomal RNA
RT	read termination
s_2U	2-thiouridine
SINE	short interspersed nuclear element
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
SNP	single-nucleotide polymorphism
snRNA	small nuclear RNA
T	thymine
t_6A	N^6-threonylcarbamoyladenosine
T-arm	TΨC-arm
TBP	TATA-box binding protein
TFIIIB	transcription factor for polymerase III B
TFIIIC	transcription factor for polymerase III C
Thr	threonine
tiRNA	tRNA-derived stress induced RNA
TN	true negative
TNR	true negative rate

TP	true positive
TPR	true positive rate
tRF	tRNA-derived fragment
tRNA	transfer RNA
tsRNA	tRNA-derived small RNA
U	uracil
UTR	untranslated regions
V-loop	variable loop
vtRNA	vault RNA

List of Figures

List of Tables

Bibliography

[1] Crick, F. "On degenerate templates and the adapter hypothesis: A note for the RNA Tie Club". In: (1955).

[2] Kresge, N., R. D. Simoni, and R. L. Hill. "The Discovery of tRNA by Paul C. Zamecnik". In: *J. Biol. Chem.* 280.40 (2005), e37. *DOI*: 10.1146/annurev.biochem.74.050304.091632.

[3] Hoagland, M. B., P. C. Zamecnik, and M. L. Stephenson. "Intermediate reactions in protein biosynthesis". In: *Biochim. Biophys. Acta.* 24.1 (1957), pp. 215–216. *DOI*: 10.1016/0006-3002(57)90175-0.

[4] Lin, S.-L., J. D. Miller, and S.-Y. Ying. "Intronic MicroRNA (miRNA)". In: *J. Biomed. Biotechnol.* 2006.26818 (2006). *DOI*: 10.1155/JBB/2006/26818.

[5] Thore, S., C. Mayer, C. Sauter, S. Weeks, and D. Suck. "Crystal Structures of the Pyrococcus abyssi Sm Core and Its Complex with RNA: Common Features of RNA binding in Archaea and Eukarya". In: *J. Biol. Chem.* 278.2 (2003), pp. 1239–1247. *DOI*: 10.1074/jbc.M207685200.

[6] Perreault, J., J.-P. Perreault, and G. Boire. "Ro-Associated Y RNAs in Metazoans: Evolution and Diversification". In: *Mol. Biol. Evol.* 24.8 (2007), pp. 1678–1689. *DOI*: 10.1093/molbev/msm084.

[7] *Section 11.2, Processing of Eukaryotic mRNA*. New York: W. H. Freeman, 2000.

[8] Crick, F. "Central Dogma of Molecular Biology". In: *Nature* 227 (1970), pp. 561–563. *DOI*: 10.1038/227561a0.

[9] *Codon wheel*. https://upload.wikimedia.org/wikipedia/commons/7/70/Aminoacids_table.svg. Accessed: 2019-07-7.

[10] Crick, F. "Codon–anticodon pairing: The wobble hypothesis". In: *J. Mol. Biol.* 19.2 (1966), pp. 548–555. *DOI*: 10.1016/s0022-2836(66)80022-0.

[11] Kirchner, S. and Z. Ignatova. "Emerging roles of tRNA in adaptive translation, signalling dynamics and disease". In: *Nat. Rev. Genet.* 2.16 (2015), pp. 98–112. *DOI*: 10.1038/nrg3861.

[12] Chan, P. and T. Lowe. "GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes". In: *Nucl. Acids Res.* 44.Database issue (2016), pp. D184–D189. *DOI*: 10.1093/nar/gkv1309.

[13] Goodenbour, J. M. and T. Pan. "Diversity of tRNA genes in eukaryotes". In: *Nucl. Acids Res.* 34.21 (2006), pp. 6137–6146. *DOI*: 10.1093/nar/gkl725.

[14] Velandia-Huerto, C. A., S. J. Berkemer, A. Hoffmann, N. Retzlaff, L. C. Romero Marroquín, M. Hernández-Rosales, P. F. Stadler, and C. I. Bermúdez-Santana. "Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies". In: *BMC Genom.* 17.1 (2016), p. 617. *DOI*: 10.1186/s12864-016-2927-4.

[15] Berkemer, S. J., A. Hoffmann, C. R. A. Murray, and P. F. Stadler. "SMORE: Synteny Modulator of Repetitive Elements". In: *Life* 7.4 (2017), p. 42. *DOI*: 10.3390/life7040042.

[16] Hoffmann, A., J. Fallmann, E. Vilardo, M. Mörl, P. F. Stadler, and F. Amman. "Accurate mapping of tRNA reads". In: *Bioinf.* 34.7 (2018), pp. 1116–1124. *DOI*: 10.1093/bioinformatics/btx756.

[17] Erber, L., A. Hoffmann, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "LOTTE-seq (Long hairpin oligonucleotide based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing". In: *RNA Biol.* 0 (2019), pp. 1–10. *DOI*: 10.1080/15476286.2019.1664250.

[18] Erber, L., A. Hoffmann, J. Fallmann, H. Betat, S. Prohaska, P. F. Stadler, and M. Mörl. "*Dictyostelium discoideum*: Unusual occurrence of two active CCA-adding enzymes". In: *IJMS* accepted (2020).

[19] Hoffmann, A., L. Erber, H. Betat, P. F. Stadler, M. Mörl, and J. Fallmann. "Changes of the tRNA modification pattern during the development of *Dictyostelium discoideum*". In: *RNA Biol.* under review (2020).

[20] Hoffmann, A., C. Lorenz, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "Temperature Dependence of Bacterial tRNA Modifications". In: In preparation (2020).

[21] Hoser, S., A. Hoffmann, A. Meindl, M. Gamper, S. Bernhart, L. Müller, M. Misslinger, M. Hoelzl, K. Perfler, K. Singer, M. Ploner, H. Lindner, H. Schaal, P. F. Stadler, and A. Hüttenhofer. "Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing". In: *Genom Biol.* accepted (2020).

[22] Coordinators, N. R. "Database resources of the National Center for Biotechnology Information". In: *Nucl. Acids Res.* 46.D1 (2017), pp. D8–D13. *DOI*: 10.1093/nar/gkx1095.

[23] Jühling, F., J. Pütz, M. Bernt, A. Donath, M. Middendorf, C. Florentz, and P. F. Stadler. "Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements". In: *Nucleic Acids Res.* 40.7 (2011), pp. 2833–2845. *DOI*: 10.1093/nar/gkr1131.

[24] Telonis, A. G., Y. Kirino, and I. Rigoutsos. "Mitochondrial tRNA-lookalikes in nuclear chromosomes: Could they be functional?" In: *RNA Biol.* 12.4 (2015), pp. 375–380. *DOI*: 10.1080/15476286.2015.1017239.

[25] Kumar, P., J. Anaya, S. B. Mudunuri, and A. Dutta. "Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets". In: *BMC Biol.* 12.1 (2014), p. 78. *DOI*: 10.1186/s12915-014-0078-0.

[26] Shi, H. and P. B. Moore. "The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited". In: *RNA* 6.8 (2000), pp. 1091–1105. *DOI*: 10.1017/s1355838200000364.

[27] Lorenz, C., C. E. Lünse, and M. Mörl. "tRNA Modifications: Impact on Structure and Thermal Adaptation". In: *Biomolecules* 7.2 (2017), p. 35. *DOI*: 10.3390/biom7020035.

[28] Root-Bernstein, R., Y. Kim, A. Sanjay, and Z. F. Burton. "tRNA evolution from the proto-tRNA minihelix world". In: *Transcription* 7.5 (2016), pp. 153–163. *DOI*: 10.1080/21541264.2016.1235527.

[29] Giegé, R., F. Jühling, J. Pütz, F. P. Stadler, C. Sauter, and C. Florentz. "Structure of transfer RNAs: similarity and variability". In: *Wiley Interdiscip. Rev.: RNA* 3.1 (2012), pp. 37–61. *DOI*: 10.1002/wrna.103.

[30] Sprinzl, M. and K. S. Vassilenko. "tRNAdb 2009: Compilation of tRNA sequences and sequences of tRNA genes". In: *Nucleic Acids Res.* 33.Database issue (2005), pp. D139–D140. *DOI*: 10.1093/nar/gkn772.

[31] Crothers, D. M., T. Seno, and G. Söll. "Is there a discriminator site in transfer RNA?" In: *Proc. Natl. Acad. Sci. U.S.A.* 69.10 (1972), pp. 3063–3067. *DOI*: 10.1073/pnas.69.10.3063.

[32] Williams, J. B., L. Cooley, and D. Söll. "Enzymatic addition of guanylate to histidine transfer RNA". In: *RNA Processing Part B: Specific Methods*. Vol. 181. Methods in Enzymology. Academic Press, 1990, pp. 451–462. *DOI*: 10.1016/0076-6879(90)81143-i.

[33] Gu, W., J. E. Jackman, A. J. Lohan, M. W. Gray, and E. M. Phizicky. "tRNAHis maturation: An essential yeast protein catalyzes addition of a guanine nucleotide to the 5' end of tRNAHis". In: *Genes Dev.* 17.23 (2003), pp. 2889–2901. *DOI*: 10.1101/gad.1148603.

[34] Rosen, A. E., B. S. Brooks, E. Guth, C. S. Francklyn, and K. Musier-Forsyth. "Evolutionary conservation of a functionally important backbone phosphate group critical for aminoacylation of histidine tRNAs". In: *RNA* 12.7 (2006), pp. 1315–1322. *DOI*: 10.1261/rna.78606.

[35] Kim, S. H., F. L. Suddath, G. J. Quigley, A. McPherson, J. L. Sussman, A. H. J. Wang, N. C. Seeman, and A. Rich. "Three-Dimensional Tertiary Structure of Yeast Phenylalanine Transfer RNA". In: *Science* 185.4149 (1974), pp. 435–440. *DOI*: 10.1126/science.185.4149.435.

[36] Itoh, Y., S. Sekine, S. Suetsugu, and S. Yokoyama. "Tertiary structure of bacterial selenocysteine tRNA". In: *Nucleic Acids Res.* 41.13 (2013), pp. 6729–6738. *DOI*: 10.1093/nar/gkt321.

[37] Kuhn, C.-D. "RNA versatility governs tRNA function". In: *BioEssays* 38.5 (2016), pp. 465–473. *DOI*: 10.1002/bies.201500190.

[38] Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young. "Sequence and organization of the human mitochondrial genome". In: *Nature* 290.5806 (1981), pp. 457–465. *DOI*: 10.1038/290457a0.

[39] Chomyn, A., P. Mariottini, M. W. J. Cleeter, C. I. Ragan, A. Matsuno-Yagi, Y. Hatefi, R. F. Doolittle, and G. Attardi. "Six unidentified reading frames of human mitochondrial DNA encode components of the respiratory-chain NADH dehydrogenase". In: *Nature* 314.6012 (1986), pp. 592–597. *DOI*: 10.1038/314592a0.

[40] Taanman, J.-W. "The mitochondrial genome: structure, transcription, translation and replication". In: *Biochim. Biophys. Acta, Bioenerg.* 1410.2 (1999), pp. 103–123. *DOI*: 10.1016/s0005-2728(98)00161-3.

[41] Huot, J. L., L. Enkler, C. Megel, L. Karim, D. Laporte, H. D. Becker, A.-M. Duchêne, M. Sissler, and L. Marchaél-Drouard. "Idiosyncrasies in decoding mitochondrial genomes". In: *Biochim.* 100 (2014), pp. 95–106. *DOI*: 10.1016/j.biochi.2014.01.004.

[42] Lithgow, T. and A. Schneider. "Evolution of macromolecular import pathways in mitochondria, hydrogenosomes and mitosomes". In: *Philos. Trans. R. Soc. London, Ser. B* 365.1541 (2010), pp. 799–817. *DOI*: 10.1098/rstb.2009.0167.

[43] Bullerwell, C. E. and M. W. Gray. "In Vitro Characterization of a tRNA Editing Activity in the Mitochondria of Spizellomyces punctatus, a Chytridiomycete Fungus". In: *J. Biol. Chem.* 280.4 (2005), pp. 2463–2470. *DOI*: 10.1074/jbc.M411273200.

[44] Arcari, P. and G. G. Brownlee. "The nucleotide sequence of a small (3S) seryl-tRNA (anticodon GCU) from beef heart mitochondria". In: *Nucleic Acids Res.* 8.22 (1980), pp. 5207–5212. *DOI*: 10.1093/nar/8.22.5207.

[45] Helm, M., H. Brulé, D. Friede, R. Giegé, D. Pütz, and C. Florentz. "Search for characteristic structural features of mammalian mitochondrial tRNAs". In: *RNA* 6.10 (2000), pp. 1356–1379. *DOI*: 10.1017/s1355838200001047.

[46] Pütz, J., B. Dupuis, M. Sissler, and C. Florentz. "Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures". In: *RNA* 13.8 (2007), pp. 1184–1190. *DOI*: 10.1261/rna.588407.

[47] Jühling, F., J. Pütz, C. Florentz, and P. F. Stadler. "Armless mitochondrial tRNAs in Enoplea (Nematoda)". In: *RNA Biology* 9.9 (2012), pp. 1161–1166. *DOI*: 10.4161/rna.21630.

[48] Wolstenholme, D. R., J. L. Macfarlane, R. Okimoto, D. O. Clary, and J. A. Wahleithner. "Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms". In: *Proc. Natl. Acad. Sci. U.S.A.* 84.5 (1987), pp. 1324–1328. *DOI*: 10.1073/pnas.84.5.1324.

[49] Ernsting, B. R., D. D. Edwards, K. J. Aldred, J. S. Fites, and C. R. Neff. "Mitochondrial genome sequence of Unionicola foili (Acari: Unionicolidae): a unique gene order with implications for phylogenetic inference". In: *Experimental and Applied Acarology* 49.4 (2009), p. 305. *DOI*: 10.1007/s10493-009-9263-1.

[50] Gray, M. W. "The endosymbiont hypothesis revisited". In: *Int. Rev. Cyt.* 141 (1992), pp. 233–357. *DOI*: 10.1016/s0074-7696(08)62068-9.

[51] Bestwick, M. L. and G. S. Shadel. "Accessorizing the human mitochondrial transcription machinery". In: *Trends Biochem. Sci.* 38.6 (2013), pp. 283–291. *DOI*: 10.1016/j.tibs.2013.03.006.

[52] Epler, J. L., L. R. Shugart, and W. E. Barnett. "N-formylmethionyl transfer ribonucleic acid in mitochondria from Neurospora". In: *J. Am. Chem. Soc.* 18.9 (1970), pp. 3575–3575. *DOI*: 10.1021/bi00820a011.

[53] Dayama, G., S. B. Emery, J. M. Kidd, and R. E. Mills. "The genomic landscape of polymorphic human nuclear mitochondrial insertions". In: *Nucleic Acids Res.* 42.20 (2014), pp. 12640–12649. *DOI*: 10.1093/nar/gku1038.

[54] Ricchetti, M., C. Fairhead, and B. Dujon. "Mitochondrial DNA repairs double-strand breaks in yeast chromosomes". In: *Nature* 100.6757 (1999), pp. 96–100. *DOI*: 10.1038/47076.

[55] Richly, E. and D. Leister. "NUMTs in Sequenced Eukaryotic Genomes". In: *Mol. Biol. Evol.* 21.6 (2004), pp. 1081–1084. *DOI*: 10.1093/molbev/msh110.

[56] Hazkani-Covo, E., R. M. Zeller, and W. Martin. "Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes". In: *PLoS Genetics* 6.2 (2010), pp. 1–11. *DOI*: 10.1371/journal.pgen.1000834.

[57] Telonis, A. G., P. Loher, Y. Kirino, and I. Rigoutsos. "Nuclear and mitochondrial tRNA-lookalikes in the human genome". In: *Front. Genet.* 5 (2014), p. 344. *DOI*: 10.3389/fgene.2014.00344.

[58] Hoser, S. "Mitochondrial-derived tRNA genes: novel regulators of gene expression?" MA thesis. Leopold-Franzes-University Innsbruck, 2018.

[59] Li, S., Z. Xu, and J. Sheng. "tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA". In: *Genes (Basel)* 9.5 (2018), p. 246. *DOI*: 10.3390/genes9050246.

[60] Fu, H., J. Feng, Q. Liu, F. Sun, Y. Tie, J. Zhu, R. Xing, Z. Sun, and X. Zheng. "Stress induces tRNA cleavage by angiogenin in mammalian cells". In: *FEBS Letters* 583.2 (2009), pp. 437–442. *DOI*: 10.1016/j.febslet.2008.12.043.

[61] Yamasaki, S., P. Ivanov, G.-f. Hu, and P. Anderson. "Angiogenin cleaves tRNA and promotes stress-induced translational repression". In: *J. Cell Biol.* 185.1 (2009), pp. 35–42. *DOI*: 10.1083/jcb.200811106.

[62] Hsieh, L.-C., S.-I. Lin, A. C.-C. Shih, J.-W. Chen, W.-Y. Lin, C.-Y. Tseng, W.-H. Li, and T.-J. Chiou. "Uncovering Small RNA-Mediated Responses to Phosphate Deficiency in Arabidopsis by Deep Sequencing". In: *Plant Physiol.* 151.4 (2009), pp. 2120–2132. *DOI*: 10.1104/pp.109.147280.

[63] Thompson, D. M., C. Lu, P. J. Green, and R. Parker. "tRNA cleavage is a conserved response to oxidative stress in eukaryotes". In: *RNA* 14.10 (2008), pp. 2095–2103. *DOI*: 10.1261/rna.1232808.

[64] Honda, S., P. Loher, M. Shigematsu, J. P. Palazzo, R. Suzuki, I. Imoto, I. Rigoutsos, and Y. Kirino. "Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers". In: *Proc. Natl. Acad. Sci. U.S.A* 112.29 (2015), E3816–E3825. *DOI*: 10.1016/j.juro.2016.01.019.

[65] Lee, Y. S., Y. Shibata, A. Malhotra, and A. Dutta. "A novel class of small RNAs: tRNA-derived RNA fragments (tRFs)". In: *Genes Dev.* 23.22 (2009), pp. 2639–2649. *DOI*: 10.1101/gad.1837609.

[66] Goodarzi, H., X. Liu, H. C. Nguyen, S. Zhang, L. Fish, and S. F. Tavazoie. "Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement". In: *Cell* 161.4 (2015), pp. 790–802. *DOI*: 10.1016/j.cell.2015.02.053.

[67] Li, Z., C. Ender, G. Meister, P. S. Moore, Y. Chang, and B. John. "Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs". In: *Nucleic Acids Res.* 40.14 (2012), pp. 6787–6799. *DOI*: 10.1093/nar/gks307.

[68] Zhang, S., L. Sun, and F. Kragler. "The phloem-delivered RNA pool contains small noncoding RNAs and interferes with translation". In: *Plant Physiol.* 150.1 (2009), pp. 378–387. *DOI*: 10.1104/pp.108.134767.

[69] Jochl, C., M. Rederstorff, J. Hertel, P. Stadler, I. Hofacker, M. Schrettl, H. Haas, and A. Huttenhofer. "Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis". In: *Nucleic Acids Res.* 36.8 (2008), pp. 2677–2689. *DOI*: 10.1093/nar/gkn123.

[70] Haiser, H. J., F. V. Karginov, G. J. Hannon, and M. A. Elliot. "Developmentally regulated cleavage of tRNAs in the bacterium *Streptomyces coelicolor*". In: *Nucleic Acids Res.* 36.3 (2008), pp. 732–741. *DOI*: 10.1093/nar/gkm1096.

[71] Kumar, P., S. B. Mudunuri, J. Anaya, and A. Dutta. "tRFdb: a database for transfer RNA fragments". In: *Nucleic Acids Res.* 43.D1 (2014), pp. D141–D145. *DOI*: 10.1093/nar/gku1138.

[72] Thompson, D. M. and R. Parker. "The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in *Saccharomyces cerevisiae*". In: *J. Cell Biol.* 185.1 (2009), pp. 43–50. *DOI*: 10.1083/jcb.200811119.

[73] Mishima, E. et al. "Conformational Change in Transfer RNA Is an Early Indicator of Acute Cellular Damage". In: *J. Am. Soc. Nephrol.* 25.10 (2014), pp. 2316–2326. *DOI*: 10.1681/ASN.2013091001.

[74] Hussain, S., A. A. Sajini, S. Blanco, S. Dietmann, P. Lombard, Y. Sugimoto, M. Paramor, J. G. Gleeson, D. T. Odom, J. Ule, and M. Frye. "N-Mediated Cytosine-5 Methylation of Vault Noncoding RNA Determines Its Processing into Regulatory Small RNAs". In: *Cell Reports* 4.2 (2013), pp. 255–261. *DOI*: 10.1016/j.celrep.2013.06.029.

[75] Ivanov, P., M. M. Emara, J. Villen, S. P. Gygi, and P. Anderson. "Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation". In: *Mol. Cell* 43.4 (2011), pp. 613–623. *DOI*: 10.1016/j.molcel.2011.06.022.

[76] Kim, H. K., G. Fuchs, S. Wang, W. Wei, Y. Zhang, H. Park, B. Roy-Chaudhuri, P. Li, J. Xu, K. Chu, F. Zhang, M. S. Chua, S. So, Q. C. Zhang, P. Sarnow, and M. A. Kay. "A transfer-RNA-derived small RNA regulates ribosome biogenesis". In: *Nature* 552.7683 (2017), pp. 57–62. *DOI*: 10.1038/nature25005.

[77] Schorn, A. J., M. J. Gutbrod, C. LeBlanc, and R. Martienssen. "LTR-Retrotransposon Control by tRNA-Derived Small RNAs". In: *Cell* 170.1 (2017), 61–71.e11. *DOI*: 10.1016/j.cell.2017.06.013.

[78] Mei, Y., J. Yong, H. Liu, Y. Shi, J. Meinkoth, G. Dreyfuss, and X. Yang. "tRNA Binds to Cytochrome c and Inhibits Caspase Activation". In: *Mol. Cell* 37.5 (2010), pp. 668–678. *DOI*: 10.1016/j.molcel.2010.01.023.

[79] Martens-Uzunova, E. S., M. Olvedy, and G. Jenster. "Beyond microRNA - Novel RNAs derived from small non-coding RNA and their implication in cancer". In: *Cancer Lett.* 340.2 (2013), pp. 201–211. *DOI*: 10.1016/j.canlet.2012.11.058.

[80] Blanco, S. et al. "Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders". In: *EMBO J* 33.18 (2014), pp. 2020–2039. *DOI*: 10.15252/embj.201489282.

[81] Phizicky, E. M. and A. K. Hopper. "tRNA biology charges to the front". In: *Genes Dev.* 24.17 (2010), pp. 1832–1860. *DOI*: 10.1101/gad.1956510.

[82] Thompson, M., R. A. Haeusler, P. D. Good, and D. R. Engelke. "Nucleolar clustering of dispersed tRNA genes". In: *Science* 302.5649 (2003), pp. 1399–1401. *DOI*: 10.1126/science.1089814.

[83] Dieci, G., G. Fiorino, M. Castelnuovo, M. Teichmann, and A. Pagano. "The expanding RNA polymerase III transcriptome". In: *Trends Genet.* 23.12 (2007), pp. 614–622. *DOI*: 10.1016/j.tig.2007.09.001.

[84] Gouge, J., N. Guthertz, K. Kramm, O. Dergai, G. Abascal-Palacios, K. Satia, P. Cousin, N. Hernandez, D. Grohmann, and A. Vannini. "Molecular mechanisms of Bdp1 in TFIIIB assembly and RNA polymerase III transcription initiation". In: *Nat. Commun.* 1.8 (2017), p. 130. *DOI*: 10.1038/s41467-017-00126-1.

[85] Zhang, G., R. Lukoszek, B. Mueller-Roeber, and Z. Ignatova. "Different sequence signatures in the upstream regions of plant and animal tRNA genes shape distinct modes of regulation". In: *Nucl. Acids Res.* 39.8 (2010), pp. 3331–3339. *DOI*: 10.1093/nar/gkq1257.

[86] Kutter, C., G. D. Brown, Â. Gonçalves, M. D. Wilson, S. Watt, A. Brazma, R. J. White, and D. T. Odom. "Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes". In: *Nat. Gen.* 43 (2011), pp. 948–955. *DOI*: 10.1038/ng.906.

[87] Dong, H., L. Nilsson, and C. G. Kurland. "Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* at Different Growth Rates". In: *J. Mol. Biol.* 260.5 (1996), pp. 649–663. *DOI*: 10.1006/jmbi.1996.0428.

[88] Ojala, D., J. Montoya, and G. Attardi. "tRNA punctuation model of RNA processing in human mitochondria". In: *Nature* 5806.290 (1981), pp. 470–474. *DOI*: 10.1038/290470a0.

[89] Tiranti, V., A. Savoia, F. Forti, M.-F. D'Apolito, M. Centra, M. Rocchi, and M. Zeviani. "Identification of the Gene Encoding the Human Mitochondrial RNA Polymerase (h-

mtRPOL) by Cyberscreening of the Expressed Sequence Tags Database". In: *Hum. Mol. Genet.* 6.4 (1997), pp. 615–625. *DOI*: 10.1093/hmg/6.4.615.

[90] Woese, C. R., O. Kandler, and M. L. Wheelis. "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya". In: *Proc. Natl. Acad. Sci. U.S.A.* 87.12 (1990), pp. 4576–4579. *DOI*: 10.1073/pnas.87.12.4576.

[91] Walker, S. C. and D. R. Engelke. "Ribonuclease P: the evolution of an ancient RNA enzyme". In: *Crit. Rev. Biochem. Mol. Biol.* 41.2 (2006), pp. 77–102. *DOI*: 10.1080/10409230600602634.

[92] Li, Y. and S. Altman. "In search of RNase P RNA from microbial genomes". In: *RNA* 10.10 (2004), pp. 1533–1540. *DOI*: 10.1073/pnas.0801906105.

[93] Randau, L. and D. Söll. "Transfer RNA genes in pieces". In: *EMBO reports* 9.7 (2008), pp. 623–628. *DOI*: 10.1038/embor.2008.101.

[94] McCorkle, G. M. and S. Altman. "Large deletion mutants of Escherichia coli tRNATyr1". In: *J. Mol. Biol.* 155.2 (1982), pp. 83–103. *DOI*: 10.1016/0022-2836(82)90438-7.

[95] Apirion, D. and A. Miczak. "RNA processing in prokaryotic cells". In: *J. Mol. Biol.* 15 (1993), pp. 113–120. *DOI*: 10.1002/bies.950150207.

[96] Ray, B. and D. Apirion. "RNAase P is dependent on RNAase E action in processing monomeric RNA precursors that accumulate in an RNAase E-mutant of *Escherichia coli*". In: *J. Mol. Biol.* 149 (1981), pp. 599–617. *DOI*: 10.1016/0022-2836(81)90349.

[97] Mörl, M. and A. Marchfelder. "The final cut. The importance of tRNA 3'-processing". In: *EMBO reports* 2.1 (2001), pp. 17–20. *DOI*: 10.1093/embo-reports/kve006.

[98] Deutscher, M. P. "tRNA processing nucleases". In: *Söll, U.L. RajBhandary (Eds.)* 15 (1995), pp. 51–65.

[99] Cudny, H. and M. P. Deutscher. "Apparent involvement of ribonuclease D in the 3' processing of tRNA precursors". In: *Proc. Natl. Acad. Sci. U.S.A.* 77.37 (1980), p. 841. *DOI*: 10.1073/pnas.77.2.837.

[100] Asha, P. K., R. T. Blouin, R. Zaniewski, and M. P. Deutscher. "Ribonuclease BN: Identification and partial characterization of a new tRNA processing enzyme". In: *Proc. Natl. Acad. Sci. U.S.A.* 80 (1983), pp. 3301–3304. *DOI*: 10.1073/pnas.80.11.3301.

[101] Seidman, J. G., F. J. Schmidt, K. Foss, and W. H. McClain. "A mutant of *Escherichia coli* defective in removing 3' terminal nucleotides from some transfer RNA precursor molecules". In: *Cell* 5.4 (1975), pp. 389–400. *DOI*: 10.1016/0092-8674(75)90058-6.

[102] Schierling, K., S. Rösch, R. Rupprecht, S. Schiffer, and A. Marchfelder. "tRNA 3' End Maturation in Archaea has Eukaryotic Features: the RNase Z from *Haloferax volcanii*". In: *J. Mol. Biol.* 316.4 (2002), pp. 895–902. *DOI*: 10.1006/jmbi.2001.5395.

[103] Mohan, A., S. Whyte, X. Wang, M. Nashimoto, and L. Levinger. "The 3' end CCA of mature tRNA is an antideterminant for eukaryotic 3'-tRNase". In: *RNA* 5 (1999), pp. 245–256. *DOI*: 10.1017/s1355838299981256.

[104] Ceballos, M. and A. Vioque. "tRNase Z". In: *Protein Pept. Lett.* 14.2 (2007), pp. 137–145. *DOI*: 10.2174/092986607779816050.

[105] Yue, D., A. M. Weiner, and N. Maizels. "The CCA-adding enzyme has a single active site". In: *J. Biol. Chem.* 273 (1998), pp. 29693–29700. *DOI*: 10.1074/jbc.273.45.29693.

[106] Schürer, H., S. Schiffer, A. Marchfelder, and M. Mörl. "This Is the End: Processing, Editing and Repair at the tRNA 3'-Terminus". In: *Biol. Chem.* 382 (1987), pp. 1147–1156. *DOI*: 10.1515/BC.2001.144.

[107] Zhu, L. and M. P. Deutscher. "tRNA nucleotidyltransferase is not essential for *Escherichia coli* viability". In: *EMBO J.* 6 (1987), pp. 2473–2477. *DOI*: 10.1002/j.1460-2075.1987.tb02528.x.

[108] Martin, N. C. and A. K. Hopper. "How single genes provide tRNA processing enzymes to mitochondria, nuclei and the cytosol". In: *Biochimie* 76.12 (1994), pp. 1161–1167. *DOI*: 10.1016/0300-9084(94)90045-0.

[109] Sprinzl, M. and F. Cramer. "The -C-C-A end of tRNA and its role in protein biosynthesis". In: *Prog. Nucleic Acid Res.* 22.1 (1979), pp. 1–69.

[110] Tocchini-Valentini, G. D., P. Fruscoloni, and G. P. Tocchini-Valentini. "Processing of multiple-intron-containing pretRNA". In: *Proc. Natl. Acad. Sci. U.S.A.* 106.48 (2009), pp. 20246–20251. *DOI*: 10.1073/pnas.0911658106.

[111] Marck, C. and H. Grosjean. "Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications". In: *RNA* 9.12 (2003), pp. 1516–1531. *DOI*: 10.1261/rna.5132503.

[112] Hopper, A. K. "Transfer RNA Post-Transcriptional Processing, Turnover, and Subcellular Dynamics in the Yeast *Saccharomyces cerevisiae*". In: *Genetics* 194.1 (2013), pp. 43–67. *DOI*: 10.1534/genetics.112.147470.

[113] Johnson, P. F. and A. K. JohnHopper. "The yeast tRNA^{Tyr} gene intron is essential for correct modification of its tRNA product". In: *Nature* 302 (1983), pp. 681–687. *DOI*: 10.1038/302681a0.

[114] Abelson, J., C. R. Trotta, and H. Li. "tRNA Splicing". In: *J. Biol. Chem.* 273.21 (1998), pp. 12685–12688. *DOI*: 10.1074/jbc.273.21.12685.

[115] Kjems, J. and R. A. Garrett. "Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium desulfurococcus mobilis". In: *Cell* 54.5 (1988), pp. 693–703. *DOI*: 10.1016/s0092-8674(88)80014-x.

[116] Li, H., C. R. Trotta, and J. Abelson. "Crystal Structure and Evolution of a Transfer RNA Splicing Enzyme". In: *Science* 280.5361 (1998), pp. 279–284. *DOI*: 10.1126/science.280.5361.279.

[117] Popow, J., A. Schleiffer, and J. Martinez. "Diversity and roles of (t)RNA ligases". In: *Cell. Mol. Life Sci.* 69.16 (2012), pp. 2657–2670. *DOI*: 10.1007/s00018-012-0944-2.

[118] Cech, T. *The RNA world: Edited by R F Gesteland and J F Atkins*. Cold Spring Harbor Press, NY, 1993, pp. 239–270. *DOI*: 10.1016/0307-4412(95)90205-8.

[119] Machnicka, M. A., A. Olchowik, H. Grosjean, and J. M. Bujnicki. "Distribution and frequencies of post-transcriptional modifications in tRNAs". In: *RNA Biol.* 11.12 (2014), pp. 1619–1629. *DOI*: 10.4161/15476286.2014.992273.

[120] Suzuki, T. and T. Suzuki. "A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs". In: *Nucl. Acids Res.* 42.11 (2014), pp. 7346–7357. *DOI*: 10.1093/nar/gku390.

[121] Bai, Y., D. T. Fox, J. A. Lacy, S. G. Van Lanen, and D. Iwata-Reuyl. "Hypermodification of tRNA in Thermophilic archaea. Cloning, overexpression, and characterization of tRNA-guanine transglycosylase from Methanococcus jannaschii". In: *J. Biol. Chem.* 275.37 (2000), pp. 28731–28738. *DOI*: 10.1074/jbc.M002174200.

[122] Jackman, J. E. and J. D. Alfonzo. "Transfer RNA modifications: nature's combinatorial chemistry playground". In: *Wiley Interdiscip. Rev. RNA* 4 (2013), pp. 35–48. *DOI*: 10.1002/wrna.1144.

[123] Phizicky, E. M. and J. D. Alfonzo. "Do all modifications benefit all tRNAs?" In: *FEBS Letters* 584.2 (2010), pp. 265–271. *DOI*: 10.1016/j.febslet.2009.11.049.

[124] Duechler, M., G. Leszczyńska, E. Sochacka, and B. Nawrot. "Nucleoside modifications in the regulation of gene expression: focus on tRNA". In: *Cell. Mol. Life Sci.* 73.16 (2016), pp. 3075–3095. *DOI*: 10.1007/s00018-016-2217-y.

[125] Herschlag, D. "RNA Chaperones and the RNA Folding Problem". In: *J. Biol. Chem.* 270.36 (1995), pp. 20871–20874. *DOI*: 10.1074/jbc.270.36.20871.

[126] Urbonavičius, J., J. Armengaud, and H. Grosjean. "Identity Elements Required for Enzymatic Formation of N2,N2-dimethylguanosine from N2-monomethylated Derivative and its Possible Role in Avoiding Alternative Conformations in Archaeal tRNA". In: *J. Mol. Biol.* 357.2 (2006), pp. 387–399. *DOI*: 10.1016/j.jmb.2005.12.087.

[127] Voigts-Hoffman, F., 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 tRNA(Lys)". In: *J. Am. Chem. Soc.* 129 (2007), pp. 13382–13383. *DOI*: 10.1021/ja075520+.

[128] Giegé, R., M. Sissler, and C. Florentz. "Universal rules and idiosyncratic features in tRNA identity". In: *Nucleic Acids Res.* 26 (1998), pp. 5017–5035. *DOI*: 10.1093/nar/26.22.5017.

[129] Sakurai, M., Y. Watanabe, K. Watanabe, and T. Ohtsuki. "A protein extension to shorten RNA: elongated elongation factor-Tu recognizes the D-arm of T-armless tRNAs in nematode mitochondria". In: *Biochem. J.* 399.2 (2006), pp. 249–256. *DOI*: 10.1042/BJ20060781.

[130] Motorin, Y. and M. Helm. "tRNA stabilization by modified nucleotides". In: *Biochem.* 49.24 (2010), pp. 24934–4944. *DOI*: 10.1021/bi100408z.

[131] Charette, M. and M. W. Gray. "Pseudouridine in RNA: What, Where, How, and Why". In: *IUBMB Life* 49.5 (2000), pp. 341–351. *DOI*: 10.1080/152165400410182.

[132] Yu, F., Y. Tanaka, K. Yamashita, T. Suzuki, A. Nakamura, N. Hirano, T. Suzuki, M. Yao, and I. Tanaka. "Molecular basis of dihydrouridine formation on tRNA". In: *Proc. Natl. Acad. Sci. U.S.A.* 108.49 (2011), pp. 19593–19598. *DOI*: 10.1073/pnas.1112352108.

[133] Dalluge, J. J., T. Hamamoto, K. Horikoshi, R. Y. Morita, K. O. Stetter, and J. A. McCloskey. "Posttranscriptional modification of tRNA in psychrophilic bacteria". In: *J. Bacteriol.* 179.6 (1997), pp. 1918–923. *DOI*: 10.1128/jb.179.6.1918-1923.1997.

[134] Kinghorn, S. M., C. P. O'Byrne, I. R. Booth, and I. Stansfield. "Physiological analysis of the role of truB in *Escherichia coli*: a role for tRNA modification in extreme temperature resistance". In: *Microbiol.* 148.11 (2002), pp. 3511–3520. *DOI*: 10.1099/00221287-148-11-3511.

[135] Durant, P. C. and D. R. Davis. "Stabilization of the anticodon stem-loop of tRNALys,3 by an A+-C base-pair and by pseudouridine11Edited by I. Tinoco". In: *J. Mol. Biol.* 285.1 (1999), pp. 115–131. *DOI*: 10.1006/jmbi.1998.2297.

[136] Horie, N., M. Hara-Yokoyama, S. Yokoyama, K. Watanabe, Y. Kuchino, S. Nishimura, and T. Miyazawa. "Two tRNA1Ile species from an extreme thermophile, Thermus thermophilus HB8: effect of 2-thiolation of ribothymidine on the thermostability of tRNA". In: *Biochem.* 24.21 (1985), pp. 5711–5715. *DOI*: 10.1021/bi00342a004.

[137] Watanabe, K., S. Yokoyama, F. Hansske, H. Kasai, and T. Miyazawa. "CD and NMR studies on the conformational thermostability of 2-thioribothymidine found in the TΨC loop of thermophile tRNA". In: *Biochem. Biophys. Res. Commun.* 91.2 (1979), pp. 671–677. *DOI*: 10.1016/0006-291x(79)91574-2.

[138] Dalluge, J. J., T. Hashizume, A. E. Sopchik, J. A. McCloskey, and D. R. Davis. "Conformational Flexibility in RNA: The Role of Dihydrouridine". In: *Nucleic Acids Res.* 24.6 (1996), pp. 1073–1079. *DOI*: 10.1093/nar/24.6.1073.

[139] Hori, H. "Methylated nucleosides in tRNA and tRNA methyltransferases". In: *Front. Genet.* 5 (2014), p. 144. *DOI*: 10.3389/fgene.2014.00144.

[140] Manickam, N., K. Joshi, M. J. Bhatt, and P. J. Farabaugh. "Effects of tRNA modification on translational accuracy depend on intrinsic codon–anticodon strength". In: *Nucleic Acids Res.* 44.4 (Dec. 2015), pp. 1871–1881. *DOI*: 10.1093/nar/gkv1506.

[141] Agris, P. F., F. A. Vendeix, and W. D. Graham. "tRNA's Wobble Decoding of the Genome: 40 Years of Modification". In: *J. Mol. Biol.* 366.1 (2007), pp. 1–13. *DOI*: 10.1016/j.jmb.2006.11.046.

[142] Meier, F., B. Suter, H. Grosjean, G. Keith, and E. Kubli. "Queuosine modification of the wobble base in tRNAHis influences *in vivo* decoding properties". In: *EMBO J.* 4.3 (1985), pp. 823–827. *DOI*: 10.1002/j.1460-2075.1985.tb03704.x.

[143] Begley, U., M. Dyavaiah, A. Patil, J. P. Rooney, D. DiRenzo, C. M. Young, D. S. Conklin, R. S. Zitomer, and T. J. Begley. "Trm9-Catalyzed tRNA Modifications Link Translation to the DNA Damage Response". In: *Mol. Cell* 28.5 (2007), pp. 860–870. *DOI*: 10.1016/j.molcel.2007.09.021.

[144] Urbonavičius, J., G. Stahl, J. M. Durand, S. N. Ben Salem, Q. Qian, P. J. Farabaugh, and G. R. Björk. "Transfer RNA modifications that alter +1 frameshifting in general fail to affect -1 frameshifting". In: *RNA* 9.6 (2003), pp. 760–768. *DOI*: 10.1261/rna.5210803.

[145] Grosjean, H., J. Edqvist, K. B. Stråby, and R. Giegé. "Enzymatic Formation of Modified Nucleosides in tRNA: Dependence on tRNA Architecture". In: *J. Mol. Biol.* 255.1 (1996), pp. 67–85. *DOI*: 10.1006/jmbi.1996.0007.

[146] Cabello-Villegas, J., I. Tworowska, and E. P. Nikonowicz. "Metal Ion Stabilization of the U-Turn of the A37 N6-Dimethylallyl-Modified Anticodon Stem-Loop of *Escherichia coli* tRNAPhe". In: *Biochem.* 43.1 (2004), pp. 55–66. *DOI*: 10.1021/bi0353676.

[147] Bjork, G., P. Wikstrom, and A. Bystrom. "Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine". In: *Science* 244.4907 (1989), pp. 986–989. *DOI*: 10.1126/science.2471265.

[148] Maehigashi, T., J. A. Dunkle, S. J. Miles, and C. M. Dunham. "Structural insights into +1 frameshifting promoted by expanded or modification-deficient anticodon stem loops". In: *Proc. Natl. Acad. Sci. U.S.A.* 111.35 (2014), pp. 12740–12745. *DOI*: 10.1073/pnas.1409436111.

[149] Kierzek, E. and R. Kierzek. "The thermodynamic stability of RNA duplexes and hairpins containing N6-alkyladenosines and 2-methylthio-N6-alkyladenosines". In: *Nucleic Acids Res.* 31.15 (2003), pp. 4472–4480. *DOI*: 10.1093/nar/gkg633.

[150] Endres, L., U. Begley, R. Clark, C. Gu, A. Dziergowska, A. Małkiewicz, J. A. Melendez, P. C. Dedon, and T. J. Begley. "Alkbh Regulates Selenocysteine-Protein Expression to Protect against Reactive Oxygen Species Damage". In: *PLoS ONE* 10.7 (2015), pp. 1–23. *DOI*: 10.1371/journal.pone.0131335.

[151] Ishitani, R., S. Yokoyama, and O. Nureki. "Structure, dynamics, and function of RNA modification enzymes". In: *Curr. Opin. Struct. Biol.* 18 (2008), pp. 330–339. *DOI*: 10.1016/j.sbi.2008.05.003.

[152] Motorin, Y. and M. Helm. "RNA nucleotide methylation". In: *Wiley Interdiscip. Rev.: RNA* 2 (2011), pp. 611–631. *DOI*: 10.1002/wrna.79.

[153] Dominissini, D., 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. Dore, N. Amariglio, G. Rechavi, and C. He. "The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA". In: *Nature* 530 (2016), pp. 441–446. *DOI*: 10.1038/nature16998.

[154] Desrosiers, R., K. Friderici, and F. Rottman. "Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells". In: *Proc. Natl. Acad. Sci. U.S.A.* 71.10 (1974), pp. 3971–3975. *DOI*: 10.1073/pnas.71.10.3971.

[155] Perry, R. P., D. E. Kelley, K. Friderici, and F. Rottman. "The methylated constituents of L cell messenger RNA: Evidence for an unusual cluster at the 5' terminus". In: *Cell* 4.4 (1975), pp. 387–394. *DOI*: 10.1016/0092-8674(75)90159-2.

[156] Hussain, S., J. Aleksic, S. Blanco, S. Dietmann, and M. Frye. "Characterizing 5-methylcytosine in the mammalian epitranscriptome". In: *Genome Biol.* 14.11 (2013), p. 215. *DOI*: 10.1186/gb4143.

[157] Squires, J. E., 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". In: *Nucl. Acids Res.* 40.11 (2012), pp. 5023–5033. *DOI*: 10.1093/nar/gks144.

[158] Dubin, D. T. and R. H. Taylor. "The methylation state of poly A-containing-messenger RNA from cultured hamster cells". In: *Nucleic Acids Res.* 2.10 (1975), pp. 1653–1668. *DOI*: 10.1093/nar/2.10.1653.

[159] Delatte, B. et al. "Transcriptome-wide distribution and function of RNA hydroxymethylcytosine". In: *Science* 351.6270 (2016), pp. 282–285. *DOI*: 10.1126/science.aac5253.

[160] Meyer, K. D., 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". In: *Cell* 149.7 (2012), pp. 1635–1646. *DOI*: 10.1016/j.cell.2012.05.003.

[161] Fustin, J.-M., M. Doi, Y. Yamaguchi, H. Hida, S. Nishimura, M. Yoshida, T. Isagawa, M. S. Morioka, H. Kakeya, I. Manabe, and H. Okamura. "RNA-Methylation-Dependent RNA Processing Controls the Speed of the Circadian Clock". In: *Cell* 155.4 (2013), pp. 793–806. *DOI*: 10.1016/j.cell.2013.10.026.

[162] Zhou, J., J. Wan, X. Gao, X. Zhang, S. R. Jaffrey, and S. B. Qian. "Dynamic m(6)A mRNA methylation directs translational control of heat shock response". In: *Nature* 526.7574 (2015), pp. 591–594. *DOI*: 10.1038/nature15377.

[163] Choi Junhong abd Leong, K.-W., H. Demirci, J. Chen, A. Petrov, A. Prabhakar, S. E. O'Leary, D. Dominissini, G. Rechavi, S. M. Soltis, M. Ehrenberg, and J. D. Puglisi. "N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics". In: *Nat. Struct. Mol. Biol* 23 (2016), pp. 110–115. *DOI*: 10.1038/nsmb.3148.

[164] Roundtree, I. A. and C. He. "RNA epigenetics–chemical messages for posttranscriptional gene regulation". In: *Curr. Opin. Chem. Biol.* 30 (2016), pp. 46–51. *DOI*: 10.1016/j.cbpa.2015.10.024.

[165] Berulava, T., S. Rahmann, K. Rademacher, L. Klein-Hitpass, and B. Horsthemke. "N6-Adenosine Methylation in MiRNAs". In: *PLoS ONE* 10.2 (2015), pp. 1–13. *DOI*: 10.1371/journal.pone.0118438.

[166] Lovejoy, A. F., D. P. Riordan, and P. O. Brown. "Transcriptome-Wide Mapping of Pseudouridines: Pseudouridine Synthases Modify Specific mRNAs in *S. cerevisiae*". In: *PLoS ONE* 9.10 (2014), pp. 1–15. *DOI*: 10.1371/journal.pone.0110799.

[167] Schwartz, S., D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst, B. X. León-Ricardo, J. M. Engreitz, M. Guttman, R. Satija, E. S. Lander, G. Fink, and A. Regev. "Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA". In: *Cell* 159.1 (2014), pp. 148–162. *DOI*: 10.1016/j.cell.2014.08.028.

[168] Zhao, B. S. and C. He. "Pseudouridine in a new era of RNA modifications". In: *Cell Res.* 25 (2014), pp. 153–154. *DOI*: 10.1038/cr.2014.143.

[169] Carlile, T. M., 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". In: *Nature* 515 (2014), pp. 143–146. *DOI*: 10.1038/nature13802.

[170] Sakurai, M., T. Yano, H. Kawabata, H. Ueda, and T. Suzuki. "Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome". In: *Nat. Chem. Biol.* 6 (2010), pp. 733–740. *DOI*: 10.1038/nchembio.434.

[171] Savva, Y. A., L. E. Rieder, and R. A. Reenan. "The ADAR protein family". In: *Genome Biol.* 13.12 (2012), p. 252. *DOI*: 10.1186/gb-2012-13-12-252.

[172] Solomon, O., L. Bazak, E. Y. Levanon, N. Amariglio, R. Unger, G. Rechavi, and E. Eyal. "Characterizing of functional human coding RNA editing from evolutionary, structural, and dynamic perspectives". In: *Proteins Struct. Funct. Bioinf.* 82.11 (2014), pp. 3117–3131. *DOI*: 10.1002/prot.24672.

[173] Vesely, C., S. Tauber, F. J. Sedlazeck, M. Tajaddod, A. von Haeseler, and M. F. Jantsch. "ADAR2 induces reproducible changes in sequence and abundance of mature microRNAs in the mouse brain". In: *Nucl. Acids Res.* 42.19 (2014), pp. 12155–12168. *DOI*: 10.1093/nar/gku844.

[174] Marvin, B. and M. Inada. *Co-transcriptional mRNA Processing in Eukaryotes. In: Bell E. (eds) Molecular Life Sciences*. Springer, NY, 2014.

[175] Haeusler, R. A., M. Pratt-Hyatt, P. D. Good, T. A. Gipson, and D. R. Engelke. "Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes". In: *Genes Dev.* 22 (2008), pp. 2204–2214. *DOI*: 10.1101/gad.1675908.

[176] Jarrous, N., J. S. Wolenski, D. Wesolowski, C. Lee, and S. Altman. "Localization in the nucleolus and coiled bodies of protein subunits of the ribonucleoprotein ribonuclease P". In: *J. Cell. Biol.* 146 (1999), pp. 559–572. *DOI*: 10.1083/jcb.146.3.559.

[177] Paushkin, S., M. Patel, B. S. Furia, S. W. Peltz, and C. R. Trotta. "Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation". In: *Mol. Biol. Cell* 117 (2004), pp. 311–321. *DOI*: 10.1016/s0092-8674(04)00342-3.

[178] Robertis, E. M. D. "Nucleocytoplasmic segregation of proteins and RNAs". In: *Cell* 32.4 (1983), pp. 1021–1025. *DOI*: 10.1016/0092-8674(83)90285-4.

[179] Mori, T., C. Ogasawara, T. Inada, M. Englert, H. Beier, M. Takezawa, T. Endo, and T. Yoshihisa. "Dual Functions of Yeast tRNA Ligase in the Unfolded Protein Response: Unconventional Cytoplasmic Splicing of HAC1 Pre-mRNA Is Not Sufficient to Release Translational Attenuation". In: *Mol. Biol. Cell* 21.21 (2010), pp. 3722–3734. *DOI*: 10.1091/mbc.E10-08-0693.

[180] Yoshihisa, T. "Handling tRNA introns, archaeal way and eukaryotic way". In: *Front. Genet.* 5 (2014), p. 213. *DOI*: 10.3389/fgene.2014.00213.

[181] Anderson, J., L. Phan, and A. G. Hinnebusch. "The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of *Saccharomyces*

cerevisiae". In: *Proc. Natl. Acad. Sci. U.S.A.* 97.10 (2000), pp. 5173–5178. *DOI*: 10.1073/pnas.090102597.

[182] Hopper, A. K. and E. M. Phizicky. "tRNA transfers to the limelight". In: *Genes Dev.* 17.2 (2003), pp. 162–180. *DOI*: 10.1101/gad.1049103.

[183] Lee, C., G. Kramer, D. E. Graham, and D. R. Appling. "Yeast Mitochondrial Initiator tRNA Is Methylated at Guanosine 37 by the Trm5-encoded tRNA (Guanine-N1-)-methyltransferase". In: *J. Biol. Chem.* 282.38 (2007), pp. 27744–27753. *DOI*: 10.1074/jbc.M704572200.

[184] Melton D. A. De Robertis E. M., C. R. "Order and intracellular location of the events involved in the maturation of a spliced tRNA". In: *Nature* 284.5752 (1980), pp. 143–148. *DOI*: 10.1038/284143a0.

[185] Schneider, A. "Mitochondrial tRNA Import and Its Consequences for Mitochondrial Translation". In: *Annu. Rev. Biochem.* 80.1 (2011), pp. 1033–1053. *DOI*: 10.1146/annurev-biochem-060109-092838.

[186] Eigen, M., B. F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A. Dress, and A. von Haeseler. "How old is the genetic code? Statistical geometry of tRNA provides an answer". In: *Science* 244.4905 (1989), pp. 673–679. *DOI*: 10.1126/science.2497522.

[187] Eigen, M. and R. Winkler-Oswatitsch. "Transfer-RNA, an early gene?" In: *Naturwiss.* 68 (1981), pp. 282–292. *DOI*: 10.1007/BF01047470.

[188] Sun, F. J., S. Fleurdépine, C. Bousquet-Antonelli, G. Caetano-Anollés, and J. Deragon. "Common evolutionary trends for SINE RNA structures". In: *Trends Genet.* 23.1 (2007), pp. 26–33. *DOI*: 10.1016/j.tig.2006.11.005.

[189] Rozhdestvensky, T. S., A. M. Kopylov, J. Brosius, and A. Hüttenhofer. "Neuronal BC1 RNA structure: Evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure". In: *RNA* 7.5 (2001), pp. 722–730. *DOI*: 10.1017/s1355838201002485.

[190] Lacoangeli, A., T. S. Rozhdestvensky, N. Dolzhanskaya, B. Tournier, J. Schütt, J. Brosius, R. B. Denman, E. W. Khandjian, S. Kindler, and H. Tiedge. "On BC1 RNA and the fragile X mental retardation protein". In: *Proc. Natl. Acad. Sci. U.S.A.* 105.2 (2008), pp. 734–739. *DOI*: 10.1073/pnas.0710991105.

[191] Nishihara, H., A. F. A. Smit, and N. Okada. "Functional noncoding sequences derived from SINEs in the mammalian genome". In: *Genome Res.* 16.7 (2006), pp. 864–874. *DOI*: 10.1101/gr.5255506.

[192] Weber, M. J. "Mammalian small nucleolar RNAs are mobile genetic elements". In: *PLoS Genet.* 2 (2006), e205. *DOI*: 10.1371/journal.pgen.0020205.

[193] Michaud, M., V. Cognar, A.-M. Duchêne, and L. Maréchal-Drouard. "A global picture of tRNA genes in plant genomes". In: *Plant J.* 66 (2011), pp. 80–93. *DOI*: 10.1111/j.1365-313X.2011.04490.x.

[194] Bermúdez-Santana, C., C. Stephan-Otto Attolini, T. Kirsten, J. Engelhardt, S. J. Prohaska, S. Steigele, and P. F. Stadler. "Genomic Organization of Eukaryotic tRNAs". In: *BMC Genom.* 11 (2010), p. 270. *DOI*: 10.1186/1471-2164-11-270.

[195] Wang, P. P. and I. Ruvinsky. "Family size and turnover rates among several classes of small non-protein-coding RNA genes in *Caenorhabditis* nematodes". In: *Genome Biol. Evol.* 4 (2012), pp. 565–574. *DOI*: 10.1093/gbe/evs034.

[196] Rogers, H. H. and S. Griffiths-Jones. "tRNA anticodon shifts in eukaryotic genomes". In: *RNA* 20 (2014), pp. 269–281. *DOI*: 10.1261/rna.041681.113.

[197] Rogers, H. H., C. M. Bergman, and S. Griffiths-Jones. "The evolution of tRNA genes in *Drosophila*". In: *Genome Biol. Evol.* 2 (2010), pp. 467–477. *DOI*: 10.1093/gbe/evq034.

[198] Hertel, J. and P. F. Stadler. "The Expansion of Animal MicroRNA Families Revisited". In: *Life* 5 (2015), pp. 905–920. *DOI*: 10.3390/life5010905.

[199] Cantatore, P., M. N. Gadaleta, M. Roberti, C. Saccone, and A. C. Wilson. "Duplication and remoulding of tRNA genes during the evolutionary rearrangement of mitochondrial genomes". In: *Nature* 329 (1987), pp. 853–855. *DOI*: 10.1038/329853a0.

[200] Rawlings, T. A., T. M. Collins, and R. Bieler. "Changing identities: tRNA duplication and remolding within animal mitochondrial genomes". In: *Proc. Natl. Acad. Sci. U.S.A.* 100 (2003), pp. 15700–15705. *DOI*: 10.1073/pnas.2535036100.

[201] Sahyoun, A. H., M. Hölzer, F. Jühling, C. Höner zu Siederdissen, M. Al-Arab, K. Tout, M. Marz, M. Middendorf, P. F. Stadler, and M. Bernt. "Towards a Comprehensive Picture of Alloacceptor tRNA Remolding in Metazoan Mitochondrial Genomes". In: *Nucleic Acids Res.* 43 (2015), pp. 8044–8056. *DOI*: 10.1093/nar/gkv746.

[202] Yona, A. H., Z. Bloom-Ackermann, I. Frumkin, V. Hanson-Smith, Y. Charpak-Amikam, Q. Feng, J. D. Boeke, O. Dahan, and Y. Pilpel. "tRNA genes rapidly change in evolution to meet novel translational demands". In: *eLife* 2 (2013), e01339. *DOI*: 10.7554/eLife.01339.

[203] Fitch, W. M. "Distinguishing Homologous from Analogous Proteins". In: *Syst. Biol.* 19.2 (1970), pp. 99–113. *DOI*: 10.2307/2412448.

[204] Liao, D. "Concerted Evolution: Molecular Mechanisms and Biological Implications". In: *Am. J. Hum. Genet.* 64 (1999), pp. 24–30. *DOI*: 10.1086/302221.

[205] Teshima, K. M. and H. Innan. "The Effect of Gene Conversion on the Divergence Between Duplicated Genes". In: *Genetics* 166 (2004), pp. 1553–1560. *DOI*: 10.1534/genetics.166.3.1553.

[206] Amstutz, H., P. Munz, W. D. Heyer, U. Leupoid, and J. Kohli. "Concerted evolution of tRNA genes: intergenic conversion among three unlinked serine tRNA genes in *S. pombe*". In: *Cell* 40 (1985), pp. 879–886. *DOI*: 10.1016/0092-8674(85)90347-2.

[207] Naidoo, K., E. Steenkamp, M. P. Coetzee, M. J. Wingfield, and B. D. Wingfield. "Concerted evolution in the ribosomal RNA cistron". In: *PLoS One* 8 (2013), e59355. *DOI*: 10.1371/journal.pone.0059355.

[208] Scienski, K., J. C. F. Fay, and G. C. Conant. "Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex". In: *Genome Biol. Evol.* 7 (2015), pp. 3249–3258. *DOI*: 10.1093/gbe/evv216.

[209] Wang, Z., M. Gerstein, and M. Snyder. "RNA sequencing: advances, challenges and opportunities". In: *Nat. Rev. Genet.* 10 (2009), pp. 57–63. *DOI*: 10.1038/nrg2484.

[210] Ozsolak, F. and P. M. Milos. "RNA sequencing: advances, challenges and opportunities". In: *Nat. Rev. Genet.* 12 (2011), pp. 87–98. *DOI*: 10.1038/nrg2934.

[211] Wilhelm, B. T. and J.-R. Landry. "RNA-Seq–quantitative measurement of expression through massively parallel RNA-sequencing". In: *Methods* 48.3 (2009), pp. 249–257. *DOI*: 10.1016/j.ymeth.2009.03.016.

[212] Zhao, W., X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou. "Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling". In: *BMC Genom.* 15.419 (2014). *DOI*: 10.1186/1471-2164-15-419.

[213] Cui, P., Q. Lin, F. Ding, C. Xin, W. Gong, L. Zhang, J. Geng, B. Zhang, X. Yu, J. Yang, S. Hu, and J. Yu. "A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing". In: *Genomics* 96.5 (2010), pp. 259–265. *DOI*: 10.1016/j.ygeno.2010.07.010.

[214] Chaitankar, V., G. Karakülah, R. Ratnapriya, F. O. Giuste, M. J. Brooks, and A. Swaroop. "Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research". In: *Prog. Retin. Eye Res.* 55 (2016), pp. 1–31. *DOI*: 10.1016/j.preteyeres.2016.06.001.

[215] Westermann, A. J., S. A. Gorski, and J. Vogel. "Dual RNA-seq of pathogen and host". In: *Nat. Rev. Microbiol.* 10 (2012), pp. 618–630. *DOI*: 10.1038/nrmicro2852.

[216] Shigematsu, M., S. Honda, P. Loher, A. G. Telonis, I. Rigoutsos, and Y. Kirino. "YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs". In: *Nucl. Acids Res.* 45.9 (2017), e70–e70. *DOI*: 10.1093/nar/gkx005.

[217] Pang, Y. L. J., R. Abo, S. S. Levine, and P. C. Dedon. "Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number". In: *Nucl. Acids Res.* 42.22 (2014), e170–e170. *DOI*: 10.1093/nar/gku945.

[218] Katz, Y., E. T. Wang, E. M. Airoldi, and C. B. Burge. "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nat. Methods* 7.12 (2010), pp. 1009–1015. *DOI*: 10.1038/nmeth.1528.

[219] Garber, M., M. G. Grabherr, M. Guttman, and C. Trapnell. "Computational methods for transcriptome annotation and quantification using RNA-seq". In: *Nat. Methods* 8 (2011), pp. 469–477. *DOI*: 10.1038/nmeth.1613.

[220] Metzker, M. L. "Sequencing technologies – the next generation". In: *Nat. Rev. Genet.* 11 (2010), pp. 31–46. *DOI*: 10.1038/nrg2626.

[221] B. Mayer, ed. *Methods and Protocols*. Humana Press, 2011, pp. 199–217. *DOI*: 10.1007/978-1-61779-027-0.

[222] Del Fabbro, C., S. Scalabrin, M. Morgante, and F. M. Giorgi. "An extensive evaluation of read trimming effects on Illumina NGS data analysis". In: *PLoS One* 8.12 (2013), pp. 31–46. *DOI*: 10.1371/journal.pone.0085024.

[223] Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. "A survey of best practices for RNA-seq data analysis". In: *Genome Biol.* 17.13 (2016). *DOI*: 10.1186/s13059-016-0881-8.

[224] Levenshtein, V. I. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710.

[225] Kemena, C. and C. Notredame. "Upcoming challenges for multiple sequence alignment methods in the high-throughput era". In: *Bioinform.* 25.19 (2009), pp. 2455–2465. *DOI*: 10.1093/bioinformatics/btp452.

[226] Needleman, S. B. and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *J. Mol. Biol.* 48.3 (1970), pp. 443–453. *DOI*: 10.1016/0022-2836(70)90057-4.

[227] Smith, T. and M. Waterman. "Identification of Common Molecular Subsequences". In: *J. Mol. Biol.* 147 (1981), pp. 195–197. *DOI*: 10.1016/0022-2836(81)90087-5.

[228] Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "Basic local alignment search tool". In: *J. Mol. Biol.* 215.3 (1990), pp. 403–410. *DOI*: 10.1016/S0022-2836(05)80360-2.

[229] Oehmen, C. S. and D. J. Baxter. "ScalaBLAST 2.0: rapid and robust BLAST calculations on multiprocessor systems". In: *Bioinform.* 29.6 (2013), pp. 797–798. *DOI*: 0.1093/bioinformatics/btt013.

[230] Feng, D.-F. and R. F. Doolittle. "Progressive sequence alignment as a prerequisite correct phylogenetic trees". In: *J. Mol. Evol.* 25.4 (1987), pp. 351–360. *DOI*: 10.1007/BF02603120.

[231] Thompson, J. D., D. G. Higgins, and T. J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucl. Acids Res.* 22.22 (Nov. 1994), pp. 4673–4680. *DOI*: 10.1093/nar/22.22.4673.

[232] Notredame, C., D. G. Higgins, and J. Heringa. "T-coffee: a novel method for fast and accurate multiple sequence alignment". In: *J. Mol. Biol.* 302.1 (2000), pp. 205–217. *DOI*: 10.1006/jmbi.2000.4042.

[233] Gollery, M. "Bioinformatics: Sequence and Genome Analysis, 2nd ed. David W. Mount. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004, 692 pp., $75.00, paperback. ISBN 0-87969-712-1." In: 2005. *DOI*: 10.1373/clinchem.2005.053850.

[234] Daugelaite, J., A. O. Driscoll, and R. D. Sleator. "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics". In: *ISRN Biomath.* 2013.615630 (2013). *DOI*: 10.1155/2013/615630.

[235] Pearson, W. R. "Selecting the Right Similarity-Scoring Matrix". In: *Curr. Protoc. Bioinformatics* 43.3 (2013), pp. 3.5.1–3.5.9. *DOI*: 10.1002/0471250953.bi0305s43.

[236] Karlin, S. and S. F. Altschul. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes". In: *Proc. Natl. Acad. Sci. U.S.A.* 87.6 (1990), pp. 2264–2268. *DOI*: 10.1073/pnas.87.6.2264.

[237] Dayhoff, M. O. and R. M. Schwartz. "Chapter 22: A model of evolutionary change in proteins". In: *Atlas of Protein Sequence and Structure*. 1978. *DOI*: doi.org/10.2307/2412074.

[238] Li, R., Y. Li, K. Kristiansen, and J. Wang. "SOAP: short oligonucleotide alignment program". In: *Bioinform.* 24.5 (2008), pp. 713–714. *DOI*: 10.1093/bioinformatics/btn025.

[239] Mercer, T. R., S. Neph, M. E. Dinger, J. Crawford, M. A. Smith, A.-M. J. Shearwood, E. Haugen, C. P. Bracken, O. Rackham, J. A. Stamatoyannopoulos, A. Filipovska, and J. S. Mattick. "The Human Mitochondrial Transcriptome". In: *Cell* 146.4 (2011), pp. 645–658. *DOI*: 10.1016/j.cell.2011.06.051.

[240] Clark, W. C., M. E. Evans, D. Dominissini, G. Zheng, and T. Pan. "tRNA base methylation identification and quantification via high-throughput sequencing". In: *RNA* 22.11 (2016), pp. 1771–1784. *DOI*: 10.1261/rna.056531.116.

[241] Hauenschild, R., 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". In: *Nucl. Acids Res.* 43.20 (2015), pp. 9950–9964. *DOI*: 10.1093/nar/gkv895.

[242] Cozen, A. E., E. Quartley, A. D. Holmes, E. Hrabeta-Robinson, E. M. Phizicky, and T. M. Lowe. "ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments". In: *Nat. Methods* 12 (2015), pp. 879–884. *DOI*: 10.1038/nmeth.3508.

[243] Langmead, B. and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nat. Methods* 9 (2012), pp. 357–359. *DOI*: 10.1038/nmeth.1923.

[244] Li, H. and R. Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinform.* 25.14 (2009), pp. 1754–1760. *DOI*: 10.1093/btp324.

[245] Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biol.* 14.R36 (2013). *DOI*: 10.1186/gb-2013-14-4-r36.

[246] Hoffmann, S., C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller. "Fast Mapping of Short Sequences with Mismatches, Insertions

and Deletions Using Index Structures". In: *PLoS Computational Biology* 5.9 (2009), pp. 1–10. *DOI*: 10.1371/journal.pcbi.1000502.

[247] Otto, C., P. F. Stadler, and S. Hoffmann. "Lacking alignments? The next-generation sequencing mapper segemehl revisited". In: *Bioinform.* 30.13 (2014), pp. 1837–1843. *DOI*: 10.1093/bioinformatics/btu146.

[248] Bentley, J. L. and R. Sedgewick. "Fast Algorithms for Sorting and Searching Strings". In: *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. 1997, pp. 360–369.

[249] Manber, U. and G. Myers. "Suffix Arrays: A New Method for On-Line String Searches". In: *SIAM J. Comput.* 22.5 (1993), pp. 935–948. *DOI*: 10.1137/0222058.

[250] Weiner, P. "Linear pattern matching algorithms". In: *14th Annual Symposium on Switching and Automata Theory (swat 1973)*. 1973, pp. 1–11. *DOI*: 10.1109/SWAT.1973.13.

[251] Abouelhoda, M. I., S. Kurtz, and E. Ohlebusch. "Replacing suffix trees with enhanced suffix arrays". In: *J. Discrete Algoritms* 2.1 (2004), pp. 53–86. *DOI*: 10.1016/S1570-8667(03)00065-0.

[252] Myers, G. "A Fast Bit-vector Algorithm for Approximate String Matching Based on Dynamic Programming". In: *J. ACM* 46.3 (1999), pp. 395–415. *DOI*: 10.1145/316542.316550.

[253] Hoffmann, S., C. Otto, G. Doose, A. Tanzer, D. Langenberger, S. Christ, M. Kunz, L. M. Holdt, D. Teupser, J. Hackermüller, and P. F. Stadler. "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection". In: *Genome Biol.* 15.2 (2014), R34. *DOI*: 10.1186/gb-2014-15-2-r34.

[254] Lowe, T. M. and P. P. Chan. "tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes". In: *Nucl. Acids Res.* 44.W1 (2016), W54–W57. *DOI*: 10.1093/nar/gkw413.

[255] Lowe, T. M. and S. R. Eddy. "tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence". In: *Nucl. Acids Res.* 25.5 (1997), pp. 955–64. *DOI*: 10.1093/nar/25.5.955.

[256] Nawrocki, E. P. and S. R. Eddy. "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinf.* 29.22 (2013), pp. 2933–2935. *DOI*: 10.1093/bioinformatics/btt509.

[257] Sakakibara, Y., M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. "Stochastic context-free grammers for tRNA modeling". In: *Nucl. Acids Res.* 22.23 (1994), pp. 5112–5120. *DOI*: 10.1093/nar/22.23.5112.

[258] Eddy, S. R. and R. Durbin. "RNA sequence analysis using covariance models". In: *Nucl. Acids Res.* 22.11 (1994), pp. 2079–2088. *DOI*: 10.1093/nar/22.11.2079.

[259] Laslett, D. and B. Canbäck. "ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences". In: *Bioinformatics* 24.2 (Nov. 2007), pp. 172–175. *DOI*: 10.1093/bioinformatics/btm573.

[260] Schwartz, S. and Y. Motorin. "Next-generation sequencing technologies for detection of modified nucleotides in RNAs". In: *RNA Biol.* 14.9 (2017), pp. 1124–1137. *DOI*: 10.1080/15476286.2016.1251543.

[261] Ryvkin, P., Y. Y. Leung, I. M. Silverman, M. Childress, O. Valladares, I. Dragomir, B. D. Gregory, and L.-S. W. Wang. "HAMR: high-throughput annotation of modified ribonucleotides". In: *RNA* 19 (2013), pp. 1684–1692. *DOI*: 10.1007/978-1-4939-8808-2_4.

[262] Motorin, Y., S. Muller, I. Behm-Ansmant, and C. Branlant. "Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods". In: *Methods Enzymol.* 425 (2007), pp. 21–53. *DOI*: 2007;425:21-53.

[263] Findeiß, S., D. Langenberger, P. F. Stadler, and S. Hoffmann. "Traces of Post-Transcriptional RNA Modifications in Deep Sequencing Data". In: *Biol. Chem.* 392 (2011), pp. 305–313. *DOI*: 10.1515/BC.2011.043.

[264] Behm-Ansmant, I., M. Helm, and Y. Motorin. "Use of Specific Chemical Reagents for Detection of Modified Nucleotides in RNA". In: *J. Nucleic Acids* 408053 (2011). *DOI*: 10.4061/2011/408053.

[265] Zheng, G., Y. Qin, W. C. Clark, Q. Dai, C. Yi, C. He, A. M. Lambowitz, and T. Pan. "Efficient and quantitative high-throughput tRNA sequencing". In: *Nat. Methods* 12 (2015), pp. 835–837. *DOI*: 10.1038/nmeth.3478.

[266] Hoffmann, S., P. F. Stadler, and K. Strimmer. "A simple data-adaptive probabilistic variant calling model". In: *Algorithms Mol. Biol.* 10 (2015), p. 10. *DOI*: 10.1186/s13015-015-0037-5.

[267] DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. "A framework for variation discovery and genotyping using next-generation dna sequencing data". In: *Nat. Genetics* 43 (2011), pp. 491–498. *DOI*: 10.1038/ng.806.

[268] Li, H. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (2011), pp. 2987–2993. *DOI*: 10.1093/bioinformatics/btr509.

[269] McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome Res.* 20.9 (2010), pp. 1297–1303. *DOI*: 10.1101/gr.107524.110.

[270] Bayes, T. and N. Price. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S". In: *Philos. Trans. Royal Soc.* 53 (1763), pp. 370–418. *DOI*: 10.1098/rstl.1763.0053.

[271] Li, H., J. Ruan, and R. Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome Res.* 18.11 (2008), pp. 1851–1858. *DOI*: 10.1101/gr.078212.108.

[272] Li, R., Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. "SNP detection for massively parallel whole-genome resequencing". In: *Genome Res.* 19.6 (2009), pp. 1124–1132. *DOI*: 10.1101/gr.088013.108.

[273] Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. "Genotype and SNP calling from next-generation sequencing data". In: *Nat. Rev. Genet.* 12.6 (2011), pp. 443–451. *DOI*: 10.1038/nrg2986.

[274] Silberberg, G. and M. Öhman. "The edited transcriptome: novel high throughput approaches to detect nucleotide deamination". In: *Curr. Opin. Genet. Dev.* 21.4 (2011), pp. 401–406. *DOI*: 10.1016/j.gde.2011.04.009.

[275] Thomassin, H., E. J. Oakeley, and T. Grange. "Identification of 5-Methylcytosine in Complex Genomes". In: *Methods* 19.3 (1999), pp. 465–475. *DOI*: 10.1006/meth.1999.0883.

[276] Stanley, J. and S. Vassilenko. "A different approach to RNA sequencing". In: *Nature* 274 (1978), pp. 87–89. *DOI*: 10.1038/274087a0.

[277] Squires, J. E. and T. Preiss. "Function and detection of 5-methylcytosine in eukaryotic RNA". In: *Epigenom.* 2.5 (2010), pp. 709–715. *DOI*: 10.2217/epi.10.47.

[278] Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. "Human–Mouse Alignments with BLASTZ". In: *Genome Res.* 13.1 (2003), pp. 103–107. *DOI*: 10.1101/gr.809403.

[279] Capra, J. A., M. Stolzer, D. Durand, and K. S. Pollard. "How old is my gene?" In: *Trends Genet.* 29 (2013), pp. 659–668. *DOI*: 10.1016/j.tig.2013.07.001.

[280] Holland, P. W. "Evolution of homeobox genes". In: *Wiley Interdiscip. Rev. Dev. Biol.* 2 (2013), pp. 31–45. *DOI*: 10.1002/wdev.78.

[281] Hiller, M., B. T. Schaar, V. B. Indjeian, D. M. Kingsley, L. R. Hagey, and G. Bejerano. "A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species". In: *Cell. Rep.* 2 (2012), pp. 817–823. *DOI*: 10.1016/j.celrep.2012.08.032.

[282] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. "A genomic perspective on protein families". In: *Science* 278 (1997), pp. 631–637. *DOI*: 10.1126/science.278.5338.631.

[283] Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin. "Computational methods for Gene Orthology inference". In: *Briefings Bioinf.* 12 (2011), pp. 379–391. *DOI*: 10.1093/bib/bbr030.

[284] Dalquen, D. A., A. M. Altenhoff, G. H. Gonnet, and C. Dessimoz. "The Impact of Gene Duplication, Insertion, Deletion, Lateral Gene Transfer and Sequencing Error on Orthology Inference: A Simulation Study". In: *PLoS ONE* 8 (2013), e56925. *DOI*: 10.1371/journal.pone.0056925.

[285] Altenhoff, A. M. and C. Dessimoz. "Phylogenetic and functional assessment of orthologs inference projects and methods". In: *PLoS Comput Biol.* 5 (2009), e1000262. *DOI*: 10.1371/journal.pcbi.1000262.

[286] Nei, M. and A. P. Rooney. "Concerted and Birth-and-Death Evolution of Multigene Families". In: *Annu. Rev. Genet.* 39 (2005), pp. 121–152. *DOI*: 10.1146/annurev.genet.39.073003.112240.

[287] Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. "Aligning multiple genomic sequences with the threaded blockset aligner". In: *Genome Res.* 14 (2004), pp. 708–715. *DOI*: 10.1101/gr.1933104.

[288] P. Bonizzoni, V. Brattka, and B. Löwe, eds. *Aligning and Labeling Genomes under the Duplication-Loss Model*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 97–107. *DOI*: 10.1007/978-3-642-39053-1.

[289] Benzaid, B., R. Dondi, and N. El-Mabrouk. "Duplication-Loss Genome Alignment: Complexity and Algorithm". In: *Language and Automata Theory and Applications, (LATA)*. Ed. by C. Dediu A.-H. amd Martin-Vide and B. Truthe. Vol. 7810. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 116–127. *DOI*: 10.1007/978-3-642-00982-2.

[290] Tremblay-Savard, O., B. Benzaid, B. F. Lang, and N. El-Mabrouk. "Evolution of tRNA Repertoires in Bacillus Inferredf with `OrthoAlign`". In: *Mol. Biol. Evol.* 32 (2015), pp. 1643–1656. *DOI*: 10.1093/molbev/msv029.

[291] Hellmuth, M., M. Hernandez-Rosales, K. T. Huber, V. Moulton, P. F. Stadler, and N. Wieseke. "Orthology Relations, Symbolic Ultrametrics, and Cographs". In: *J. Math. Biol.* 66 (2013), pp. 399–420. *DOI*: 10.1007/s00285-012-0525-x.

[292] Hellmuth, M., N. Wieseke, M. Lechner, H.-P. Lenhof, M. Middendorf, and P. F. Stadler. "Phylogenetics from Paralogs". In: *Proc. Natl. Acad. Sci. USA* 112 (2015). 10.1073/pnas.1412770112, pp. 2058–2063.

[293] Lafond, M. and N. El-Mabrouk. "Orthology and paralogy constraints: satisfiability and consistency". In: *BMC Genom.* 15 S6 (2014), S12. *DOI*: 10.1186/1471-2164-15-S6-S12.

[294] Lafond, M., M. Semeria, K. M. Swenson, E. Tannier, and N. El-Mabrouk. "Gene tree correction guided by orthology". In: *BMC Bioinform.* 14 S15 (2013), S5. *DOI*: 10.1186/1471-2105-14-S15-S5.

[295] Lafond, M., R. Dondi, and N. El-Mabrouk. "The link between orthology relations and gene trees: a correction perspective". In: *Alg. Mol. Biol.* 11 (2016), p. 4. *DOI*: 10.1186/s13015-016-0067-7.

[296] Hernandez-Rosales, M., M. Hellmuth, N. Wieseke, K. T. Huber, V. Moulton, and P. F. Stadler. "From Event-Labeled Gene Trees to Species Trees". In: *BMC Bioinform.* 13.Suppl. 19 (2012), S6. *DOI*: 10.1186/1471-2105-13-S19-S6.

[297] Corneil, D. G., H. Lerchs, and L. Steward Burlingham. "Complement reducible graphs". In: *Discr. Appl. Math.* 3 (1981), pp. 163–174. *DOI*: 10.1016/0166-218X(81)90013.

[298] Liu, Y., J. Wang, J. Guo, and J. Chen. "Complexity and parameterized algorithms for Cograph Editing". In: *Theor. Comp. Sci.* 461 (2012), pp. 45–54. *DOI*: 10.1016/j.tcs.2011.11.040.

[299] Farris, J. S. "Phylogenetic analysis under Dollo's law". In: *Syst. Zoology* 26 (1977), pp. 77–88. *DOI*: 10.2307/2412867.

[300] Sprinzl, M., C. Horn, M. Brown, A. Loudovitch, and S. Steinberg. "Compilation of tRNA sequences and sequences of tRNA genes". In: *Nucleic Acids Res.* 26.1 (1998), pp. 148–153. *DOI*: 10.1093/nar/gki012.

[301] Jühling, F., M. Mörl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Pütz. "tRNAdb 2009: compilation of tRNA sequences and tRNA genes". In: *Nucleic Acids Res.* 37.suppl_1 (2009), p. D159. *DOI*: 10.1093/nar/gkn772.

[302] Bushnell, B. *BBMap/BBTools*. https://github.com/BioInfoTools/BBMap. [Online; accessed 13-02-2018]. 2016.

[303] Martin, M. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet J.* 17.1 (2011), pp. 10–12. *DOI*: 10.14806/ej.17.1.200.

[304] Andrews, S. *FastQC*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc. [Online; accessed 12-10-2017]. 2010.

[305] Quinlan, A. R. and I. M. Hall. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". In: *Bioinform.* 26.6 (2010), pp. 841–842. *DOI*: 10.1093/bioinformatics/btq033.

[306] Edgar, R. C. "Search and clustering orders of magnitude faster than BLAST". In: *Bioinformatics* 26.19 (2010), p. 2460. *DOI*: 10.1093/bioinformatics/btq461.

[307] Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *JSTOR* 57.q (1995), pp. 289–300. *DOI*: 10.1111/j.2517-6161.1995.tb02031.x.

[308] Sun, S., M. Hood, L. Scott, Q. Peng, S. Mukherjee, J. Tung, and X. Zhou. "Differential expression analysis for RNAseq using Poisson mixed models". In: *Nucl. Acids Res.* 45.11 (2017), e106. *DOI*: 10.1093/nar/gkx204.

[309] Benjamini, Y. and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *J. R. Stat. Soc.* 57.1 (1995), pp. 289–300. *DOI*: 10.1111/j.2517-6161.1995.tb02031.x.

[310] Läuter, H. "Cook, R. D., S. Weisberg: Residuals and influence in regression. Chapman and Hall, New York — London 1982. VIII, 229 pp., £ 12,-". In: *Biom. J.* 27.1 (1985), pp. 80–80. *DOI*: 10.1002/bimj.4710270110.

[311] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.

[312] Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. "The Human Genome Browser at UCSC". In: *Genome Res.* 12.6 (2002), pp. 996–1006. *DOI*: 10.1101/gr.229102.

[313] Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses". In: *Curr. Protoc. Bioinformatics* 54.1 (2016), pp. 1.30.1–1.30.33. *DOI*: 10.1002/cpbi.5.

[314] Zerbino, D. R. et al. "Ensembl 2018". In: *Nucl. Acids Res.* 46.D1 (2017), pp. D754–D761. *DOI*: 10.1093/nar/gkx1098.

[315] Li, W. *RNASeqReadSimulator*. github.com/davidliwei/RNASeqReadSimulator. [Online; accessed 27-02-2018]. 2014.

[316] Tserovski, L., V. Marchand, R. Hauenschild, F. Blanloeil-Oillo, M. Helm, and Y. Motorin. "High-throughput sequencing for 1-methyladenosine (m1A) mapping in RNA". In: *Methods* 107 (2016), pp. 110–121. *DOI*: 10.1016/j.gpb.2018.03.003.

[317] Robinson, J. T. R., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. "Integrative Genomics Viewer". In: *Nat. Biotechnol.* 29 (2011), pp. 24–26. *DOI*: 10.1038/nbt.1754.

[318] Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". In: *Briefings Bioinf.* 14 (2013), pp. 178–192. *DOI*: 10.1093/bib/bbs017.

[319] Holtgrewe, M., A.-K. Emde, D. Weese, and K. Reinert. "A novel and well-defined benchmarking method for second generation read mapping". In: *BMC Bioinf.* 12.1 (2011), p. 210. *DOI*: 10.1186/1471-2105-12-210.

[320] Wildschutte, J. H., A. Baron, N. M. Diroff, and J. M. Kidd. "Discovery and characterization of Alu repeat sequences via precise local read assembly". In: *Nucleic Acids Res.* (2015), gkv1089. *DOI*: 10.1093/nar/gkv1089.

[321] Kahles, A., J. Behr, and G. Rätsch. "MMR: a tool for read multi-mapper resolution". In: *Bioinformatics* (2015), btv624. *DOI*: 10.1093/bioinformatics/btv624.

[322] Hashimoto, T., M. J. De Hoon, S. M. Grimmond, C. O. Daub, Y. Hayashizaki, and G. J. Faulkner. "Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite". In: *Bioinformatics* 25.19 (2009), pp. 2613–2614. *DOI*: 10.1093/bioinformatics/btp438.

[323] Břinda, K., V. Boeva, and G. Kucherov. "Dynamic read mapping and online consensus calling for better variant detection". In: *arXiv preprint arXiv:1605.09070* (2016).

[324] Consortium, T. E. P. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489.7414 (2012), pp. 57–74. *DOI*: 10.1186/s13058-016-0718.

[325] Sloan, C. A., E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. "ENCODE data at the ENCODE portal". In: *Nucleic Acids Res.* 44.Database issue (2016), pp. D726–D732. *DOI*: 10.1093/nar/gkv1160.

[326] Balakrishnan, R., J. Park, K. Karra, B. C. Hitz, G. Binkley, E. L. Hong, J. Sullivan, G. Micklem, and J. Michael Cherry. "YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit". In: *Database* 2012 (2012). *DOI*: 10.1093/database/bar062.

[327] Ernst, F. G. M., L. Erber, J. Sammler, F. Jühling, H. Betat, and M. Mörl. "Cold adaptation of tRNA nucleotidyltransferases: A tradeoff in activity, stability and fidelity". In: *RNA Biol.* 15.1 (2018), pp. 144–155. *DOI*: 10.1080/15476286.2017.1391445.

[328] Cathala, G., J.-F. Savouret, B. Mendez, B. L. West, M. Karin, J. A. Martial, and J. D. Baxter. "A Method for Isolation of Intact, Translationally Active Ribonucleic Acid". In: *DNA* 2.4 (1983), pp. 329–335. *DOI*: 10.1089/dna.1983.2.329.

[329] Humana press., 2006. *DOI*: 10.1385/1597451444.

[330] Pavon-Eternod, M., S. Gomes, R. Geslain, Q. Dai, M. R. Rosner, and T. Pan. "tRNA over-expression in breast cancer and functional consequences". In: *Nucl. Acids Res.* 37.21 (2009), pp. 7268–7280. *DOI*: 10.1093/nar/gkp787.

[331] Goodarzi, H., H. C. Nguyen, S. Zhang, B. D. Dill, H. Molina, and S. F. Tavazoie. "Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression". In: *Cell* 165.6 (2016), pp. 1416–1427. *DOI*: 10.1016/j.cell.2016.05.046.

[332] Plotkin, J. B., H. Robins, and A. J. Levine. "Tissue-specific codon usage and the expression of human genes". In: *Proc. Natl. Acad. Sci. USA* 101.34 (2004), pp. 12588–12591. *DOI*: 10.1073/pnas.0404957101.

[333] Kietrys, A. M., W. A. Velema, and E. T. Kool. "Fingerprints of Modified RNA Bases from Deep Sequencing Profiles". In: *J. Am. Chem. Soc.* 139.47 (2017), pp. 17074–17081. *DOI*: 10.1021/jacs.7b07914.

[334] Zhou, Y., J. M. Goodenbour, L. A. Godley, A. Wickrema, and T. Pan. "High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma". In: *Biochem. Biophys. Res. Commun.* 385.2 (2009), pp. 160–164. *DOI*: 10.1016/j.bbrc.2009.05.031.

[335] Helm, M. and Y. Motorin. "Detecting RNA modifications in the epitranscriptome: predict and validate". In: *Nat. Rev. Genet* 18 (2017), pp. 275–291. *DOI*: 10.1038/nrg.2016.169.

[336] Hanada, T. et al. "CLP1 links tRNA metabolism to progressive motor-neuron loss". In: *Nature* 495 (2013), pp. 474–480. *DOI*: 10.1016/j.cub.2016.01.045.

[337] Clarke, C. J. et al. "The Initiator Methionine tRNA Drives Secretion of Type II Collagen from Stromal Fibro s to Promote Tumor Growth and Angiogenesis". In: *Curr. Biol.* 26.6 (2016), pp. 755–765. *DOI*: 10.1016/j.cub.2016.01.045.

[338] Orioli, A. "tRNA biology in the omics era: Stress signalling dynamics and cancer progression". In: *BioEssays* 39.3 (2017), p. 1600158. *DOI*: 10.1002/bies.201600158.

[339] Björk, G. R. ., J. U. Ericson, C. E. D. Gustafsson, T. G. Hagervall, Y. H. Jönsson, and P. M. Wikström. "Transfer RNA Modification". In: *Annu. Rev. Biochem.* 56.1 (1987), pp. 263–285. *DOI*: 10.1146/annurev.bi.56.070187.001403.

[340] Pan, T. "Modifications and functional genomics of human transfer RNA". In: *Cell Res.* 28.10 (2018), pp. 395–404. *DOI*: 10.1038/s41422-018-0013-y.

[341] Wittig, B. and S. Wittig. "Reverse transcription of tRNA". In: *Nucl. Acids Res.* 5 (1978), pp. 1165–1178. *DOI*: 10.1093/nar/5.4.1165.

[342] Verma, I. M. "6 Reverse Transcriptase". In: ed. by P. D. Boyer. Vol. 14. The Enzymes. Academic Press, 1981, pp. 87–103. *DOI*: 10.1016/S1874-6047(08)60332-7.

[343] Topisirovic, I. and N. Sonenberg. "Distinctive tRNA Repertoires in Proliferating versus Differentiating Cells". In: *Cell* 158.6 (2014), pp. 1238–1239. *DOI*: 10.1016/j.cell.2014.08.031.

[344] Parisien, M., X. Wang, and T. Pan. "Diversity of human tRNA genes from the 1000-genomes project". In: *RNA Biol.* 10.12 (2013), pp. 1853–1867. *DOI*: 10.4161/rna.27361.

[345] Geslain, R. and T. Pan. "Functional Analysis of Human tRNA Isodecoders". In: *J. Mol. Biol.* 396.3 (2010), pp. 821–831. *DOI*: 10.1016/j.jmb.2009.12.018.

[346] Michaud, M., V. Cognat, A.-M. Duchêne, and L. Maréchal-Drouard. "A global picture of tRNA genes in plant genomes". In: *Plant J.* 66.1 (2011), pp. 80–93. *DOI*: 10.1111/j.1365-313X.2011.04490.x.

[347] Grosjean, H. *DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution: Nucleic Acids Are Not Boring Long Polymers of Only Four Types of Nucleotides: A Guided Tour*. Austin, TX, USA, 2009.

[348] Brandmayr, C., M. Wagner, T. Brückl, D. G. Globisch, D. Pearson, A. C. Kneuttinger, V. Reiter, A. Hienzsch, S. Koch, I. T. Thoma, P. Thumbs, S. Michalakis, M. Müller, M. Biel, and T. Carell. "Isotope-Based Analysis of Modified tRNA Nucleosides Correlates Modification Density with Translational Efficiency". In: *Angew. Chem. Int. Ed.* 51.30 (2012), pp. 11162–11165. *DOI*: 10.1002/anie.201203769.

[349] Wang, Y., C. Pang, X. Li, Z. Hu, Z. Lv, B. Zheng, and P. Chen. "Identification of tRNA nucleoside modification genes critical for stress response and development in rice and *Arabidopsis*". In: *BMC Plant. Biol.* 17.1 (2017), p. 261. *DOI*: 10.1186/s12870-017-1206-0.

[350] Cambridge, UK: Cambridge University Press, 2001.

[351] Chisholm, R. L. and R. A. Firtel. "Insights into morphogenesis from a simple developmental system". In: *Nat. Rev. Mol. Cell Biol.* 5 (2004), pp. 531–541. *DOI*: 10.1038/nrm1427.

[352] Schwartz, M. H., H. Wang, J. N. Pan, W. C. Clark, S. Cui, M. J. Eckwahl, D. W. Pan, M. Parisien, S. M. Owens, B. L. Cheng, K. Martinez, J. Xu, E. B. Chang, T. Pan, and A. M. Eren. "Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis". In: *Nat. Commun.* 9.5353 (2018). *DOI*: 10.1038/s41467-018-07675-z.

[353] Lorenz, C. "Analysen zur Temperaturabhängigkeit posttranskriptioneller Modifikationen in bakteriellen tRNAs". dissertation. University Leipzig, 2019.

[354] Michael I Love, W. H. and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12 (2015), p. 550. *DOI*: 10.1186/s13059-014-0550-8.

[355] Jiang, H. and W. H. Wong. "Statistical inferences for isoform expression in RNA-Seq". In: *Bioinf.* 25.8 (2009), pp. 1026–1032. *DOI*: 10.1093/bioinformatics/btp113.

[356] Ver Hoef, J. M. and P. L. Boveng. "Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?" In: *Ecology* 88.11 (2007), pp. 2766–2772. *DOI*: 10.1890/07-0043.1.

[357] Huang, H.-C., Y. Niu, and L.-X. Qin. "Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software". In: *Cancer Inf.* 14.Suppl. 1 (2015), pp. 57–67. *DOI*: 10.4137/CIN.S21631.

[358] Ebhardt, H. A., 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". In: *Nucl. Acids Res.* 37.8 (2009), pp. 2461–2470. *DOI*: 10.1093/nar/gkp093.

[359] Iida, K., H. Jin, and J. K. Zhu. "Bioinformatics analysis suggests base modifications of tRNAs and miRNAs Arabidopsis thaliana". In: *BMC Genom.* 10 (2009), p. 155. *DOI*: 10.1186/1471-2164-10-155.

[360] Li, H. "Toward better understanding of artifacts in variant calling from high-coverage samples". In: *Bioinform.* 30.20 (2014), pp. 2843–2851. *DOI*: 10.1093/bioinformatics/btu356.

[361] Richardson, R. B., D. S. Allan, and Y. Le. "Greater organ involution in highly proliferative tissues associated with the early onset and acceleration of ageing in humans". In: *Exp. Gerontol.* 55 (2014), pp. 80–91. *DOI*: 10.1016/j.exger.2014.03.015.

[362] Dittmar, K. A., J. M. Goodenbour, and T. Pan. "Tissue-specific differences in human transfer RNA expression". In: *PLoS Genet* 2.12 (2006), e221. *DOI*: 10.1371/journal.pgen.0020221.

[363] Maraia, R. J. and A. G. Arimbasseri. "Factors That Shape Eukaryotic tRNAomes: Processing, Modification and Anticodon–Codon Use". In: *Biomol.* 7.1 (2017), p. 26. *DOI*: 10.3390/biom7010026.

[364] Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel. "An evolutionarily conserved mechanism for controlling the efficiency of protein translation". In: *Cell* 141.2 (2010), pp. 344–354. *DOI*: 10.1016/j.cell.2010.03.031.

[365] Marchand, V., L. Ayadi, F. G. M. Ernst, J. Hertler, V. Bourguignon-Igel, A. Galvanin, A. Kotter, M. Helm, D. L. J. Lafontaine, and Y. Motorin. "AlkAniline-Seq: Profiling of m7G and m3C RNA Modifications at Single Nucleotide Resolution". In: *Angew. Chem. Int. Ed.* 57.51 (2018), pp. 16785–16790. *DOI*: 10.1002/anie.201810946.

[366] Berkemer, S. J. "Towards Dynamic Programming on Generalized Data Structures and Applications of Dynamic Programming in Bioinformatics. Across-frequency in convolutive blind source separation". dissertation. University Leipzig, 2019.

[367] Drosophila 12 Genomes Consortium. "Evolution of genes and genomes on the *Drosophila* phylogeny". In: *Nature* 450 (2007), pp. 203–218. *DOI*: 10.1038/nature06341.

[368] Powers, J. G., V. J. Weigman, J. Shu, J. M. Pufky, D. Cox, and P. Hurban. "Efficient and accurate whole genome assembly and methylome profiling of *E. coli*". In: *BMC Genom.* 14 (2013), p. 675. *DOI*: 10.1186/1471-2164-14-675.

[369] Lopes, R. R., A. C. Kessler, C. Polycarpo, and J. D. Alfonzo. "Cutting, dicing, healing and sealing: the molecular surgery of tRNA". In: *Wiley Interdiscip. Rev. RNA* 6 (2015), pp. 337–349. *DOI*: 10.1002/wrna.1279.

[370] Lalaouna, D., M. C. Carrier, S. Semsey, J. S. Brouard, J. Wang, J. T. Wade, and E. Massé. "A 3' external transcribed spacer in a tRNA transcript acts as a sponge for small RNAs to prevent transcriptional noise". In: *Mol. Cell* 58 (2015), pp. 393–405. *DOI*: 10.1016/j.molcel.2015.03.013.

[371] Van Bortle, K. and V. G. Corces. "tDNA insulators and the emerging role of TFIIIC in genome organization". In: *Transcription* 3 (2012), pp. 277–284. *DOI*: 10.4161/trns.21579.

[372] Gstir, R., S. Schafferer, M. Scheideler, M. Misslinger, M. Griehl, N. Daschil, C. Humpel, G. J. Obermair, C. Schmuckermair, J. Striessnig, B. E. Flucher, and A. Hüttenhofer. "Generation of a neuro-specific microarray reveals novel differentially expressed noncoding RNAs in mouse models for neurodegenerative diseases". In: *RNA* 20.12 (2014), pp. 1929–1943. *DOI*: 10.1261/rna.047225.114.

[373] Tsuji, J., M. C. Frith, K. Tomii, and P. Horton. "Mammalian NUMT insertion is non-random". In: *Nucl. Acids Res.* 40.18 (2012), pp. 9073–9088. *DOI*: 10.1093/nar/gks424.

[374] Calabrese, F. M., D. Simone, and M. Attimonelli. "Primates and mouse NumtS in the UCSC Genome Browser". In: *BMC Bioinform.* 13.S15 (2012). *DOI*: 10.1186/1471-2105-13-S4-S15.

[375] Schimmel, P. "The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis". In: *Nat. Rev. Mol. Cell Biol.* 19 (2018), pp. 45–58. *DOI*: 10.1038/nrm.2017.77.

[376] Paul, D., A. N. Sinha, A. Ray, M. Lal, S. Nayak, A. Sharma, B. Mehani, D. Mukherjee, S. V. Laddhaa, A. Suri, C. Sarkar, and A. Mukhopadhyay. "A-to-I editing in human miRNAs is enriched in seed sequence, influenced by sequence contexts and significantly hypoedited in glioblastoma multiforme". In: *Sci. Rep.* 7.2466 (2017). *DOI*: 10.1038/s41598-017-02397-6.

[377] Sarin, L. P. and S. A. Leidel. "Modify or die? - RNA modification defects in metazoans". In: *RNA Biol.* 11.12 (2014), pp. 1555–1567. *DOI*: 10.4161/15476286.2014.992279.

[378] Schachner, E., H.-J. Aschhoff, and H. Kersten. "Specific changes in lactate levels, lactate dehydrogenase patterns and cytochrome b559 in Dictyostelium discoideum caused by queuine". In: *Eur. J. Biochem* 139.3 (1984), pp. 481–487. *DOI*: 10.1111/j.1432-1033.1984.tb08031.x.

[379] Carlile, T. M., M. F. Rojas-Duran, and W. V. Gilbert. "Transcriptome-Wide Identification of Pseudouridine Modifications Using Pseudo-seq". In: *Curr. Proto. Mol. Biol.* 112.1 (2015), pp. 4.25.1–4.25.24. *DOI*: 10.1002/0471142727.mb0425s112.

[380] Treangen, T. J. and S. L. Salzberg. "Repetitive DNA and next-generation sequencing: computational challenges and solutions". In: *Nature Rev. Gen.* 13 (2012), pp. 36–46. *DOI*: 10.1038/nrg3117.

[381] Storvall, H., D. Ramsköld, and R. Sandberg. "Efficient and Comprehensive Representation of Uniqueness for Next-Generation Sequencing by Minimum Unique Length Analyses". In: *PLoS ONE* 8 (2013), e53822. *DOI*: 10.1371/journal.pone.0053822.

[382] Eirín-López, J. M., L. Rebordinos, A. P. Rooney, and J. Rozas. "The Birth- and- Death Evolution of Multigene Families Revisited". In: *Genome Dyn.* 7 (2012), pp. 170–196. *DOI*: 10.1159/000337119.

[383] Lopez, J., J. C. Stephens, and S. J. O'Brien. "The long and short of nuclear mitochondrial DNA (Numt) lineages". In: *Trends Ecol. Evol.* 12.3 (1997), p. 114. *DOI*: 10.1016/S0169-5347(97)84925-7.

[384] Martin, E. R., D. D. Kinnamon, M. A. Schmidt, E. H. Powell, S. Zuchner, and R. W. Morris. "SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies". In: *Bioinform.* 26.22 (2010), pp. 2803–2810. *DOI*: 10.1093/bioinformatics/btq526.

[385] Hauenschild, R., S. Werner, L. Tserovski, A. Hildebrandt, Y. Motorin, and M. Helm. "CoverageAnalyzer (CAn): A Tool for Inspection of Modification Signatures in RNA Sequencing Profiles". In: *Biomol.* 6.4 (2016). *DOI*: 10.3390/biom6040042.

[386] Gilbert, W. "Origin of life: The RNA world". In: *Nature* 319.2 (1986), p. 618. *DOI*: 10.1038/319618a0.

[387] Karki, R., D. Pandya, R. C. Elston, and C. Ferlini. "Defining mutation and polymorphism in the era of personal genomics". In: *BMC Med. Genom.* 8 (2015), p. 37. *DOI*: 10.1186/s12920-015-0115-z.

[388] Abbott, J. A., C. S. Francklyn, and S. M. Robey-Bond. "Transfer RNA and human disease". In: *Front. Genet.* 5 (2014), p. 158. *DOI*: 10.3389/fgene.2014.00158.

[389] Belostotsky, R., Y. Frishberg, and N. Entelis. "Human mitochondrial tRNA quality control in health and disease". In: *RNA Biol.* 9.1 (2012), pp. 33–39. *DOI*: 10.4161/rna.9.1.18009.

[390] Feijão, P. "Reconstruction of ancestral gene orders using intermediate genomes". In: *BMC Bioinform.* 16 S14 (2015), S3. *DOI*: 10.1186/1471-2105-16-S14-S3.

[391] Hu, F., Y. Lin, and J. Tang. "MLGO: phylogeny reconstruction and ancestral inference from gene-order data". In: *BMC Bioinform.* 15 (2014), p. 354. *DOI*: 10.1186/s12859-014-0354-6.

[392] Braga, M. D. V. and J. Stoye. "Sorting Linear Genomes with Rearrangements and Indels". In: *IEEE/ACM Trans. Comp. Biol. Bioinf.* 12 (2015), pp. 500–506. *DOI*: 10.1109/TCBB.2014.2329297.

[393] Fried, C., W. Hordijk, S. J. Prohaska, C. R. Stadler, and P. F. Stadler. "The Footprint Sorting Problem". In: *J. Chem. Inf. Comput. Sci.* 44 (2004), pp. 332–338. *DOI*: 10.1021/ci030411+.

[394] Eger, S. "Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics". In: *Information Sci.* 237 (2013), pp. 287–304. *DOI*: 10.1016/j.ins.2013.02.031.

[395] Lokshtanov, D., F. Mancini, and C. Papadopoulos. "Characterizing and Computing Minimal Cograph Completions". In: *Discrete Appl. Math.* 158 (2010), pp. 755–764. *DOI*: 10.1016/j.dam.2009.01.016.

[396] Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos. "Gene conversion: mechanisms, evolution and human disease". In: *Nat. Rev. Genet.* 8 (2007), pp. 762–775. *DOI*: 10.1038/nrg2193.

[397] Lui, L. and T. Lowe. "Small nucleolar RNAs and RNA-guided post-transcriptional modification". In: *Essays Biochem.* 54 (2013), pp. 53–77. *DOI*: 10.1042/bse0540053.

[398] Kondrak, G. "A New Algorithm for the Alignment of Phonetic Sequences". In: *Proceedings of NAACL 2000 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 288–295.

[399] Torres, A. G. T., D. Piñaeyro, N. Rodríguez-Escribáand Camacho, O. Reina, A. Saint-Láger, L. Filonava, E. Batlle, and L. Ribas de Pouplana. "Inosine Modifications in Human tRNAs Are Incorporated at the Precursor tRNA Level". In: *Nucl. Acids Res.* 43.10 (2015), pp. 5145–5157. *DOI*: 10.1093/nar/gkv277.

[400] Bernt, M., D. Merkle, K. Rasch, G. Fritzsch, M. Perseke, D. Bernhard, M. Schlegel, P. F. Stadler, and M. Middendorf. "CREx: Inferring Genomic Rearrangements Based on Common Intervals". In: *Bioinform.* 23 (2007), pp. 2957–2958. *DOI*: 10.1093/bioinformatics/btm468.

[401] Vinayak, M. and C. Pathak. "Queuosine modification of tRNA: its divergent role in cellular machinery". In: *Biosci. Rep.* 30.2 (2010), pp. 135–148. *DOI*: 10.1042/BSR20090057.

[402] Noma, A., Y. Kirino, Y. Ikeuchi, and T. Suzuki. "Biosynthesis of wybutosine, a hypermodified nucleoside in eukaryotic phenylalanine tRNA". In: *EMBO J.* 25.10 (2006), pp. 2142–2154. *DOI*: 10.1038/sj.emboj.7601105.

[403] Englert, M., A. Latz, D. Becker, O. Gimple, H. Beier, and K. Akama. "Plant pre-tRNA splicing enzymes are targeted to multiple cellular compartments". In: *Biochimie* 89.11 (2007), pp. 1351–1365. *DOI*: 10.1016/j.biochi.2007.06.014.

[404] Park, M. Y., G. Wu, A. Gonzalez-Sulser, H. Vaucheret, and R. S. Poethig. "Nuclear processing and export of microRNAs in Arabidopsis". In: *Proc. Natl. Acad. Sci. U.S.A.* 102.10 (2005), pp. 3691–3696. *DOI*: 10.1073/pnas.0405570102.

[405] Chen, Q., M. Yan, Z. Cao, X. Li, Y. Zhang, J. Shi, G.-h. Feng, H. Peng, X. Zhang, Y. Zhang, J. Qian, E. Duan, Q. Zhai, and Q. Zhou. "Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder". In: *Science* 351.6271 (2016), pp. 397–400. *DOI*: 10.1126/science.aad7977.

[406] Yuan, M.-L., D.-D. Wei, B.-J. Wang, W. Dou, and J.-J. Wang. "The complete mitochondrial genome of the citrus red mite *Panonychus citri*". In: *BMC Genom.* 11.1 (2010), p. 597. *DOI*: 10.1186/1471-2164-11-597.

[407] Greer, C. L., C. L. Peebles, P. Gegenheimer, and J. Abelson. "Mechanism of action of a yeast RNA ligase in tRNA splicing". In: *Cell* 32.2 (1983), pp. 537–546. *DOI*: 10.1016/0092-8674(83)90473-7.

[408] Alva, V., S.-Z. Nam, J. Söding, and A. N. Lupas. "The MPI Bioinformatics Toolkit as an Integrative Platform for Advanced Protein Sequence and Structure Analysis". In: *Nucleic Acids Res.* 44.Web Server issue (2016), W410–W415. *DOI*: 10.1093/nar/gkw348.

[409] Ghodsi, M., B. Liu, and M. Pap. "DNACLUST". In: *BMC Bioinf.* 12 (2011), p. 271. *DOI*: 10.1186/1471-2105-12-271.

[410] Essig, K., N. Kronbeck, J. C. Guimaraes, C. Lohs, A. Schlundt, A. Hoffmann, G. Behrens, S. Brenner, J. K. andC Lopez-Rodriguez, J. Jemielity, H. Holtmann, K. Reiche, J. Hackermüller, M. Sattler, M. Zavolan, and V. Heissmeyer. "Roquin targets mRNAs in a 3'-UTR-specific manner by different modes of regulation". In: *Nat. Commun.* 9.1 (2018), pp. 2041–1723. *DOI*: 10.1038/s41467-018-06184-3.

[411] Behrens, G., R. Winzen, N. Rehage, A. Dörrie, M. Barsch, A. Hoffmann, J. Hackermüller, C. Tiedje, V. Heissmeyer, and H. Holtmann. "A translational silencing function of MCPIP1/Regnase-1 specified by the target site context". In: *Nucl. Acids Res.* 46.8 (2018), pp. 4256–4270. *DOI*: 10.1093/nar/gky106.

[412] Rehage, N., E. Davydova, C. Conrad, G. Behrens, A. Maiser, J. E. Stehklein, S. Brenner, J. Klein, A. Jeridiand, A. Hoffmann, E. Lee, U. Dianzani, R. Willemsen, R. Feederle, K. Reiche, J. Hackermüller, H. Leonhardt, S. Sharma, D. Niessing, and V. Heissmeyer. "Binding of NUFIP2 to Roquin promotes recognition and regulation of ICOS mRNA". In: *Nat. Commun.* 9.1 (2018), p. 299. *DOI*: 10.1038/s41467-017-02582-1.

Curriculum Scientiae

Personal Information

Name Anne Hoffmann

Birth July 14, 1987

Education

10/2015-10/2019 PhD student
University of Leipzig, Department Bioinformatics

10/2011–09/2014 Master Student
Bioinformatics at Martin Luther University of Halle-Wittenberg
• Thesis: RNA secondary structure determinants of Roquin-RNA interactions
• At: Helmholtz Centre for Environmental Research – UFZ Leipzig

10/2008–09/2011 Bachelor Student
Bioinformatics at Martin Luther University of Halle-Wittenberg
• Thesis: Molecular genetic characterization of phosphorylable amino acids of histone demethylase LSD1 in *Drosophila melanogaster*

Awards

2016 Ferchau Engineering Advancement Award
• Award for outstanding study achievements, Germany

Working Experience

since 10/2019	Research assistant University of Leipzig, Medical Research Center
2013–2015	Student assistant Martin Luther University of Halle-Wittenberg, Department of Developmental Genetics
2010–2014	Student assistant University Hospital Halle, Institute of Medical Epidemiology, Biometry and Informatics
2011–2013	Teaching assistant Martin Luther University of Halle-Wittenberg, Institute of Mathematics • Lecture: Linear Algebra • Lecture: Geometry

IT-Knowledge

Operating systems:	MacOs, Linux, Windows
Programming:	Perl, Python, C#, R, Bash/Shell
Markup language:	Latex

Languages

German:	native speaker
English:	fluent

Publications

[376] C. A. Velandia-Huerto, S. J. Berkemer, A. Hoffmann, N. Retzlaff, L. C. Romero Marroquín, M. Hernández-Rosales, P. F. Stadler, and C. I. Bermúdez-Santana. "Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies". In: *BMC Genom.* 17.1 (2016), p. 617. *DOI*: 10.1186/s12864-016-2927-4.

[377] S. J. Berkemer, A. Hoffmann, C. R. A. Murray, and P. F. Stadler. "SMORE: Synteny Modulator of Repetitive Elements". In: *Life* 7.4 (2017), p. 42. *DOI*: 10.3390/life7040042.

[394] G. Behrens, R. Winzen, N. Rehage, A. Dörrie, M. Barsch, A. Hoffmann, J. Hackermüller, C. Tiedje, V. Heissmeyer, and H. Holtmann. "A translational silencing function of MCPIP1/Regnase-1 specified by the target site context". In: *Nucl. Acids Res.* 46.8 (2018), pp. 4256–4270. *DOI*: 10.1093/nar/gky106.

[396] K. Essig, N. Kronbeck, J. C. Guimaraes, C. Lohs, A. Schlundt, A. Hoffmann, G. Behrens, S. Brenner, J. K. andC Lopez-Rodriguez, J. Jemielity, H. Holtmann, K. Reiche, J. Hackermüller, M. Sattler, M. Zavolan, and V. Heissmeyer. "Roquin targets mRNAs in a 3'-UTR-specific manner by different modes of regulation". In: *Nat. Commun.* 9.1 (2018), pp. 2041–1723. *DOI*: 10.1038/s41467-018-06184-3.

[397] A. Hoffmann, J. Fallmann, E. Vilardo, M. Mörl, P. F. Stadler, and F. Amman. "Accurate mapping of tRNA reads". In: *Bioinf.* 34.7 (2018), pp. 1116–1124. *DOI*: 10.1093/bioinformatics/btx756.

[402] N. Rehage, E. Davydova, C. Conrad, G. Behrens, A. Maiser, J. E. Stehklein, S. Brenner, J. Klein, A. Jeridiand, A. Hoffmann, E. Lee, U. Dianzani, R. Willemsen, R. Feederle, K. Reiche, J. Hackermüller, H. Leonhardt, S. Sharma, D. Niessing, and V. Heissmeyer. "Binding of NUFIP2 to Roquin promotes recognition and regulation of ICOS mRNA". In: *Nat. Commun.* 9.1 (2018), p. 299. *DOI*: 10.1038/s41467-017-02582-1.

[406] L. Erber, A. Hoffmann, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "LOTTE-seq (Long hairpin oligonucleotide based tRNA high-throughput sequencing): Specific selection of tRNAs with 3'-CCA end for high-throughput sequencing". In: *RNA Biol.* 0 (2019), pp. 1–10. *DOI*: 10.1080/15476286.2019.1664250.

[408] L. Erber, A. Hoffmann, J. Fallmann, H. Betat, S. Prohaska, P. F. Stadler, and M. Mörl. "*Dictyostelium discoideum*: Unusual occurrence of two active CCA-adding enzymes". In: *IJMS* accepted (2020).

[409] A. Hoffmann, L. Erber, H. Betat, P. F. Stadler, M. Mörl, and J. Fallmann. "Changes of the tRNA modification pattern during the development of *Dictyostelium discoideum*". In: *RNA Biol.* under review (2020).

[410] A. Hoffmann, C. Lorenz, J. Fallmann, H. Betat, P. F. Stadler, and M. Mörl. "Temperature Dependence of Bacterial tRNA Modifications". In: In preparation (2020).

[411] S. Hoser, A. Hoffmann, A. Meindl, M. Gamper, S. Bernhart, L. Müller, M. Misslinger, M. Hoelzl, K. Perfler, K. Singer, M. Ploner, H. Lindner, H. Schaal, P. F. Stadler, and A. Hüttenhofer. "Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing". In: *Genom Biol.* accepted (2020).

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 09.06.2020

(Ort, Datum)

(Unterschrift)