Technische Universität München
TUM School of Natural Sciences

Development of mim-tRNAseq for global and quantitative profiling of cellular tRNA pools by high-throughput sequencing

Andrew Behrens

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Franz Hagn

Prüfer*innen der Dissertation:

1. Prof. Dr. Danny Nedialkova
2. Prof. Dr. Julien Gagneur
3. Prof. Dr. Maria Colomé-Tatché

Die Dissertation wurde am 11.08.20222 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 11.11.2022 angenommen.

‘‘New directions in science are launched by new tools much more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained.’’

Freeman Dyson

TABLE OF CONTENTS

Acknowledgements

The past five years have been particularly turbulent and arduous. Surprisingly, this PhD was by no means the most difficult part, and for that I owe thanks to the following people:

To my advisor and mentor, Danny Nedialkova, who has facilitated my growth as a young scientist by providing an endless supply of curiosity-driven excitement for new data. This kind of excitement is especially infectious. Her openness to new ideas, and her keenness for new knowledge creates a space for research filled with opportunity and discontent with mediocrity. My desire to always make a more beautiful plot is owed almost entirely to her.

I would like to thank my thesis advisory committee, Julien Gagneur and Maria Colomé-Tatché. So many of the great ideas that turned into features in mim-tRNAseq came from our very productive meetings together. I learned the invaluable worth of an outside perspective in a research project from the two of them.

At the intersection between great friend and supportive colleague are Kris Le Vay and Geri Rodschinka. A huge shout out to Geri for being my first friend here, and for her support in all things German, especially upon my arrival from a foreign land. To Kris for support beyond the academic kind, and for incredible meals on Sundays.

Many thanks to Hans Jörg, Maxi, and the other members of the IMPRS-LS graduate school – a truly world-class environment geared towards support and education.

A special thanks to Assa Yeroslaviz at the MPIB bioinformatics core facility, who mentored me early on in my PhD journey into computational biology. Also, to Marja Driessen and Rin Ho Kim from the MPIB NGS core facility for sequencing our libraries "on-demand" and, sometimes, at such short notice.

To the rest of Nedialkova lab for their support, especially Katrin Strasser for her technical support, without which many important mim-tRNAseq datasets would not look so great!

Old friends from afar, Tyrone and Zak. Our Friday sessions were paramount in keeping me sane and providing much-needed entertainment and laughs. Howzit Friday.

To my partner, Anni, who has had the rather unfortunate task of being with me for the final year and a half of this PhD. No easy task, to say the least – thank you! Her optimism, ambition, and dedication are exemplary. You may not have taught me about science, but you have taught me about myself.

Finally, to the Behrens clan; my loud, exaggerative, lovely family. Particularly Mom, my unwavering emotional support, always kind and selfless…

This is dedicated to my dad; my original source of curiosity and logic.
"Wish you were here" – Pink Floyd

Abstract

Transfer RNAs (tRNAs) are adapter molecules that bridge the gap between RNA and protein during translation where they supply amino acids to the growing polypeptide chain. It has been suggested that tRNA abundance and the stoichiometry of their extensive modifications is dynamically regulated in different cell contexts, and defects in abundance or modifications are linked to neurological disease and cancer. While the impact cell context-specific tRNA regulation on translation speed, co-translational folding of proteins, and proteome integrity has been suggested, technical limitations to methods for global tRNA quantitation have impeded these analyses.

Two main challenges exist for high-throughput sequencing-based technologies in generating snapshots of cellular tRNA profiles: 1) cDNA synthesis by reverse transcriptases (RTs) during sequencing library generation is impeded by prevalent Watson-Crick face nucleotide modifications and tRNA secondary structure, resulting in abortive cDNA synthesis, tRNA coverage biases, and inaccurate abundance estimation. 2) High levels of gene duplication and sequence similarity amongst eukaryotic tRNA genes complicates unambiguous, high-resolution, and accurate read alignment and transcript quantitation.

Current tRNA sequencing technologies have yet to fully overcome these two challenges, especially evidenced by the lack of robust computational pipelines and algorithms tailored to tRNA sequencing data analysis. Furthermore, a lack of detailed protocols, documentation, and community support has hindered the progress of library generation methods and computational tools alike.

In this thesis, I introduce modification-induced misincorporation tRNA sequencing (mim-tRNAseq), which overcomes challenges to tRNA sequencing and facilitates the investigation of previously intractable questions in tRNA biology. In chapter 2, the development and testing of the mim-tRNAseq method is described. mim-tRNAseq includes an optimized workflow for library generation with a highly processive RT (TGIRT) that enables readthrough of almost all common Watson-Crick face modifications, mitigating coverage and transcript abundance estimation biases. In combination, the open-source mim-tRNAseq computational package allows the customizable, user-friendly analysis of coverage, abundance, charging fractions, and modification identity and stoichiometry in a single command. Several novel algorithms and concepts in the package allow for drastically improved read alignment to clustered tRNAs while retaining transcript-level resolution for abundance and modification analysis.

By comparing to existing methods, and confirmation with experimental evidence, we show the accuracy and improved resolution of mim-tRNAseq in quantifying transcripts and modifications compared to other methods. Interestingly, we show that despite extensive tRNA transcript regulation amongst human cell types, abundances of anticodon pools remain stable.

With improved modification analysis, we also find a striking interdependence of modifications at distinct tRNA positions, giving insight into structural determinants of tRNA modifications.

Chapter 3 describes a detailed protocol for the implementation and use of mim-tRNAseq, assisting users with all steps of the protocol, and providing useful tips, alternatives, and a troubleshooting guide. I describe upgrades to the read deconvolution algorithm, which offer further improvements to accuracy and resolution of transcript-level analyses. Various outputs and quality-control metrics are described to help users in optimizing library generation and customization of the analysis steps.

The mim-tRNAseq computational package represents the first open-source, comprehensive toolkit for tRNA sequencing analysis with extensive documentation, community-support, and ongoing development. For these reasons, mim-tRNAseq is expected to be widely adopted for studying new aspects of tRNA biology, and will see future improvements that extend its functionality and accuracy further.

Zusammenfassung

Transfer-RNAs (tRNAs) sind Adaptermoleküle, die während der Translation die ein Bindeglied zwischen RNA und Protein darstellen, indem sie Aminosäuren für die wachsende Polypeptidkette bereitstellen. Es wurde vermutet, dass die Häufigkeit und die Stöchiometrie von tRNAs durch ihre umfangreichen Modifikationen in verschiedenen Zellkontexten dynamisch reguliert wird und dass Defekte in der Häufigkeit oder bei den Modifikationen mit neurologischen Erkrankungen und Krebs in Verbindung gebracht werden. Es wurden zahlreiche Hypothesen aufgestellt, um die Auswirkungen zellkontextspezifischer tRNA-Regulierung auf die Translationsgeschwindigkeit, auf die co-translationale Faltung von Proteinen und auf die Integrität des Proteoms zu erklären, allerdings haben technische Beschränkungen bei Methoden zur globalen tRNA-Quantifizierung diese Analysen erschwert.

Bei der Erstellung von Momentaufnahmen von zellulären tRNA-Profilen mit Hilfe von Hochdurchsatz-Sequenzierungstechnologien gibt es zwei große Herausforderungen: 1) Die cDNA-Synthese durch Reverse Transkriptasen (RTs) während der Generierung von Sequenzierproben wird durch die vorherrschenden Modifikationen an der Watson-Crick Basenpaarung und die tRNA-Sekundärstruktur behindert, was zu einer unvollständigen cDNA-Synthese, einer verzerrten tRNA Coverage und einer ungenauen Schätzung der Abundanz führt. 2) Das hohe Maß an Genduplikation und Sequenzähnlichkeit zwischen eukaryotischen tRNA-Genen erschwert ein eindeutiges, hochauflösendes und genaues Read-Alignment und die Quantifizierung von Transkripten.

Die aktuellen tRNA-Sequenzierungstechnologien müssen diese beiden Herausforderungen noch vollständig bewältigen, was insbesondere durch den Mangel an robusten Rechenpipelines und Algorithmen für die Analyse von tRNA-Sequenzierungsdaten belegt wird. Darüber hinaus sind detaillierte Protokolle, Dokumentationen und die Unterstützung durch die Gemeinschaft mangelhaft und haben den Fortschritt der Methodenentwicklung zur Herstellung von Sequenzierproben und Analyseprogrammen gleichermaßen behindert.

In dieser Arbeit stelle ich die „modification-induced tRNA sequencing“ (mim-tRNAseq) vor, die die Herausforderungen der tRNA-Sequenzierung überwindet und die Untersuchung von bisher unlösbaren Fragen der tRNA-Biologie erleichtert. In Kapitel 2 wird die Entwicklung und Erprobung der mim-tRNAseq-Methode beschrieben: mim-tRNAseq umfasst einen optimierten Arbeitsablauf für die Generierung von Sequenzierproben mit einer hochgradig prozessiven RT (TGIRT), die das Lesen fast aller gängigen Watson-Crick- Modifikationen ermöglicht und damit die Datenverzerrungen bei der Schätzung von Coverage und Transkriptmenge vermindert. In Kombination mit unserem Open-Source-Rechenpaket ermöglicht mim-tRNAseq die anpassbare, benutzerfreundliche Analyse von Coverage, Abundanz, Menge an geladenen Aminosäuren sowie Identität und Stöchiometrie der Modifikationen in einem einzigen Befehl. Zahlreiche neue Algorithmen und Konzepte in dem Paket ermöglichen ein drastisch verbessertes Read-Alignment auf geclusterte tRNA, während zugleich die Auflösung auf Transkript-Ebene für die Analyse von Abundanz und Modifikationen beibehalten wird.

Durch Vergleiche mit bestehenden Methoden und durch labortechnische Experimente zeigen wir die Genauigkeit und verbesserte Auflösung von mim-tRNAseq bei der Quantifizierung von Transkripten und Modifikationen im Vergleich zu anderen Methoden. Interessanterweise konnten wir zeigen, dass trotz der umfangreichen Regulierung von tRNA-Transkripten in verschiedenen menschlichen Zelltypen die Häufigkeiten der Anticodon-Pools stabil bleibt. Mit einer verbesserten Modifikationsanalyse finden wir auch eine auffällige Abhängigkeit von Modifikationen an verschiedenen tRNA-Positionen, die Einblicke in strukturelle Determinanten von tRNA-Modifikationen geben.

Kapitel 3 beschreibt ein detailliertes Protokoll für die Implementierung und Verwendung von mim-tRNAseq. Es unterstützt die Benutzer bei allen Schritten des Protokolls und bietet nützliche Tipps, Alternativen und eine Anleitung zur Fehlerbehebung. Ich beschreibe Upgrades für den Read-Dekonvolution-Algorithmus, die weitere Verbesserungen bei der Genauigkeit und Auflösung von Analysen auf Transkript-Ebene bieten. Es werden verschiedene Ergebnisse und Metriken zur Qualitätskontrolle beschrieben, die den Benutzern bei der Optimierung der Erstellung von Sequenzierproben und der Anpassung der Analyseschritte helfen.

Das Analysepaket mim-tRNAseq ist das erste umfassende Open-Source-Toolkit für die tRNA-Sequenzierungsanalyse mit umfangreicher Dokumentation, Community-Unterstützung und kontinuierlicher Weiterentwicklung. Aus diesen Gründen ist zu erwarten, dass mim-tRNAseq in großem Umfang für die Untersuchung neuer Aspekte der tRNA-Biologie eingesetzt wird und in Zukunft weitere Verbesserungen erfahren wird, die die Funktionalität und Genauigkeit weiter ausbauen.

List of contributed publications

- **Andrew Behrens**, Geraldine Rodschinka, & Danny D. Nedialkova. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Molecular. Cell* **81**, 1–14 (2021). https://doi.org/10.1016/j.molcel.2021.01.028

- **Andrew Behrens** & Danny D. Nedialkova. Experimental and computational workflow for the analysis of tRNA pools from eukaryotic cells by mim-tRNAseq. *STAR Protocols.* **3**, 101579 (2022). https://doi.org/10.1016/j.xpro.2022.101579

CHAPTER 1

General introduction

Introduction

Transfer RNAs (tRNAs) are short, abundant, noncoding RNA that supply amino acids to the growing polypeptide chain during translation. In this regard, they function as the adapter between RNA and protein as they selectively base pair to codons in mRNA sequence via their anticodons. Due to structural constraints that permit tRNA accommodation in empty ribosomal A-sites during translation, mature tRNA transcripts are typically 76 – 90 nucleotides in length and conform to cloverleaf secondary structures and L-shaped tertiary structures[1]. To aid in conferring this structure, tRNA nucleotide modifications are both the most numerous and most diverse of any cellular RNA species[2]. However, these modifications also confer additional functionality to tRNAs that assist in decoding, wobble pairing, and frame maintenance during translation, and therefore also range in essentiality for normal function and cell physiology[3–6].

tRNAs are classified into 21 isotypes, based on which of the 21 amino acids (including selenocysteine) they carry to the ribosome during translation. Furthermore, due to the degeneracy of the genetic code, tRNAs of a particular isotype usually contain isoacceptors that possess different anticodons yet are charged with the same amino acid. Further still, tRNAs sharing an anticodon but differing in sequence elsewhere are known as isodecoders. In eukaryotic genomes, it is also not uncommon that multiple copies of a particular tRNA gene exist. This multiplicity results in fairly expansive and diverse tRNA gene sets in eukaryotes, ranging from 275 tRNA genes in yeast, to ~600 in human, and up to ~8600 in zebrafish[7].

Given their central role in translation, and requirement in high abundance, tRNAs were long assumed to be ubiquitously expressed. With the advent of quantitative methods such as microarrays[8,9] and high-throughput sequencing[10–14], investigators quickly learned that tRNA regulation in multicellular organisms is more complex than originally thought; differential expression amongst tRNA anticodon pools might be concordant with changing codon demand in divergent transcriptomes[15], numerous sources of evidence in multiple species also point to precise regulation at the transcript level, even among those sharing anticodons[10,11], and even transcription of identical tRNA genes has been shown to be variable within an organism[16,17].

Still, little is understood regarding the function of tRNA regulation. Isodecoders that "read" the same codon in mRNA sequence are seemingly redundant, at least at the level of translation, yet they are often differentially regulated. Does sequence variation outside of the anticodon also impact the dynamics of translation, or is this regulation a byproduct of gene duplication, evolution, silencing, and pseudogenization? More intriguing still are the mechanisms underlying this regulation, which mostly remain unclear.

The importance of tRNA expression, structure, modification, aminoacylation, and nearly all other levels of tRNA biogenesis and quality control, are underscored by the plethora of diseases and abnormal physiology that result from dysregulation of these processes (reviewed in [18–21]). Additionally, promising technologies aimed at using tRNAs as therapies for protein synthesis-related disorders reveal the utility and power that tRNAs possess as biotechnological targets in human health[22,23].

However, questions regarding fundamental tRNA biology remain unanswered, particularly those pertaining to their regulation. Advances in this regard have been hindered by technical challenges in global, accurate quantitation of tRNA pools in eukaryotes. These challenges stem from the multiplicity of tRNA genes, as well as their modifications, which disrupt reverse transcription - an essential and central step in high-throughput sequencing library generation. Optimization of tRNA sequencing methods has been an active area of research in the last decade, however many biases and shortcomings remain to be solved. This is particularly true for computational methods tailored to suit the analysis of tRNA sequencing datasets, which so far have been lacking in the field.

This chapter briefly describes the basic biology of tRNA. This is contextualized by current literature on the importance of tRNAs in human health and tRNA-based therapy, and the need for accurate methods to measure their abundance and modifications to further understand complex pathophysiology. Lastly, the status quo of global tRNA quantitation methods is explored, with focus on the complications that both biochemical and computational methods have faced so far.

tRNA structure and biogenesis

Structure

Mature tRNA transcripts consist of conserved stem-loop elements that can be depicted as the canonical cloverleaf-like secondary structure (**Figure 1A**). The acceptor stem is formed between the 5’ and 3’ end and is covalently linked to its corresponding amino acid at the conserved 3’-CCA tail. In a 5’ to 3’ direction, this stem is followed by the D-arm, the anticodon stem-loop that harbors the anticodon, the variable loop, and the T-arm. The D-arm gets its name from its conserved dihydrouridine modifications that confer tertiary structural elements to the tRNA. Similarly, the T-arm, or TΨC arm, contains conserved thymidine, pseudouridine (symbolized by Ψ), and cytosine residues[1]. The variable loop is the only part of the tRNA with some variability in length that is not conserved across all tRNAs. The stem-loops fold onto themselves forming the conserved L-shaped tertiary structure of tRNAs (**Figure 1B**), with one arm formed by the acceptor stem and the T-arm, and the other by the anticodon stem and the D-arm.

Due to this conserved secondary structure, tRNAs also have a canonical numbering system for nucleotide positions along their length, from 1 – 76 (**Figure 1A**)[24]. This facilitates easy reference to specific elements, including modification positions. For example, the anticodon is always in positions 34 – 36 and the 3'-CCA tail is at 74 – 76. Importantly, the discriminator base at position 73 is required for recognition specificity and charging of the tRNA by its correct aminoacyl tRNA synthetase (aaRS)[25].

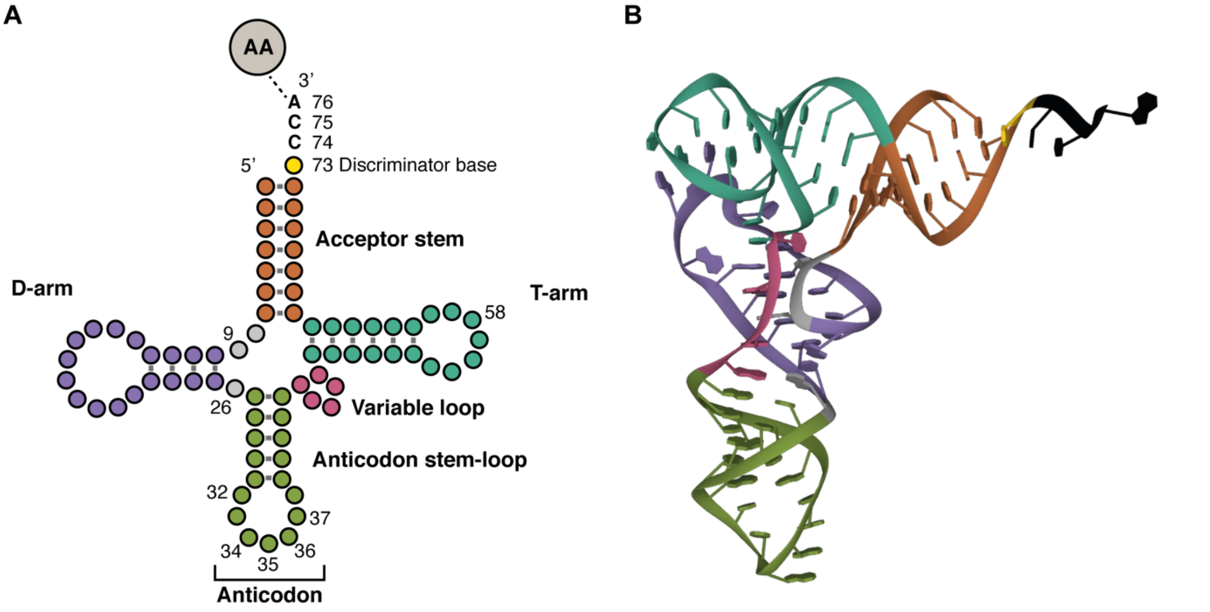


Figure 1 Basic tRNA structure.
(A) Secondary structure of tRNA showing the typical cloverleaf form. Various structural elements are indicated; from 5' to 3'; Acceptor stem, D-arm, anticodon stem-loop, variable loop, and acceptor stem. Also highlighted are the conserved canonical base positions for important elements, such as the anticodon (34 – 36), discriminator base (73), CCA tail (74 – 76), and multiple well-conserved modified nucleotide positions (e.g., 9, 32, 34, 37, 58). AA – amino acid.
(B) Tertiary molecular structure of *S. cerevisiae* tRNA-Phe (PDB; 1EHZ) showing typical L-shaped structure. Colors are consistent with structural elements in (A).
Adapted from Berg and Brandl (2021)[137].

Transcription

In eukaryotes, all tRNAs are transcribed by the 17-subunit RNA polymerase III (RNAPIII) complex to produce tRNA precursor transcripts (pre-tRNA; **Figure 2A**). Apart from tRNAs, RNAPIII is also responsible for the transcription of a limited set of other short, non-coding RNAs including ribosomal 5S RNA, the U6 small nuclear RNA (snRNA) required for the spliceosomal complex, 7SL RNA of the signal recognition particle complex (SRP), among a growing list of additional targets[26].

As with all polymerases, RNAPIII recruitment and transcription is mediated by transcription factors (TFs) and a variety of promoter elements that recruit them. Three classes of RNAPIII promoters exist, named type 1 – 3; type 1 promoters are found exclusively in 5S rRNA genes and contain an internal control region (ICR), consisting of the A- and C-box promoter elements, which recruit TFIIIA and TFIIIC[27]. Type 2 promoters are those found in tRNA genes, are also located internally, and consist of an A- and B-box that recruit TFIIIC[28,29]. Lastly, type 3 promoters, which are vertebrate-specific, contain external 5' TATA-box and proximal sequence element (PSE) promoters upstream of the transcriptional start site (TSS)[31]. The PSE is recognized and bound by the snRNA activating protein complex (SNAPc)[30,31]. Successful recruitment and binding of the TFs in all three cases leads to the recruitment of the common TFIIIB and subsequent recruitment of RNAPIII complex and transcription initiation.

tRNA gene transcription begins with the binding of TFIIIC to the internal A- and B-box elements using two domains that are separated by a flexible linker (**Figure 2A**)[32]. This flexibility affords TFIIIC the ability to bind variably spaced A- and B-boxes accounting for difference in variable region length, and the subset of tRNAs with introns. TFIIIC binding is thought to be transient, but is required for the recruitment of the multi-subunit TFIIIB upstream of the TSS, which binds far more stably and can facilitate multiple rounds of transcription reinitiation[33]. Finally, tRNA transcription is terminated at a stretch of at least five T residues, and is dependent on three subunits of RNAPIII[34,35]. Because of this affinity and stability, regulation of TFIIIB recruitment by modulation of its interaction with DNA and TFIIIC by other proteins (e.g., Dr1, RB, p107, p130, and p53) may be an important level of regulation for tRNA gene transcription (reviewed in [36,37]).

Processing, modification and aminoacylation

Pre-tRNA transcripts require nuclear processing to function as mature tRNA in translation (**Figure 2A**). La protein binds the poly(U) tract at pre-tRNA 3' ends to aid in correct pre-tRNA folding and to protect it from 3' exonuclease activity. Meanwhile, the 5' leader sequence is removed by the conserved RNase P ribozyme, and 3' trailer sequences are cleaved by RNase Z following the discriminator base at position 73. 3' end maturation is completed by the CCA adding enzyme, which post-transcriptionally and without template, adds the CCA tail to eukaryotic tRNA ends. This enzyme is also responsible for the repair of truncated CCA ends. It has been hypothesized that global tRNA 3' truncation, which also globally reduces aminoacylated tRNA proportions, might serve as a rapid and efficient way to globally inhibit translation in stress conditions[38]. However, this remains an open question. Intron-containing tRNAs are recognized by the tRNA splicing endonuclease (TSEN) complex for intron removal.

tRNAs are the most modified RNA in any cell, with on average ~12 of their 76 nucleotides receiving chemical modifications. Over 100 different types of modifications have been detected in tRNAs so far, while human cytosolic tRNAs contain at least 39 of these[21]. Modifications in the body of the tRNA often regulate their structural stability[2]. However, modifications at position 34, the first nucleotide of the anticodon that pairs with the third position of the codon, is often modified with a variety of modifications that can enable or limit wobble base pairing of tRNAs to non-cognate codons (**Figure 1A**)[3]. Therefore, these modifications are important regulators controlling the stringency as well as the flexibility of tRNAs to accurately decode more than one codon without error. Position 37 outside of the anticodon is also often highly modified and is known to regulate decoding. These modifications can stabilize the pairing of tRNA position 36 with the first nucleotide of the codon, thereby reducing frameshifting and increasing fidelity during translation[4,39].

tRNA charging with amino acids is carried out by aminoacyl tRNA synthetases (aaRS). Generally, one aaRS is specific for one amino acid, and recognizes one group of isoacceptors. There are exceptions, however, for example SerRS serylates both serine and selocysteine tRNAs[40]. This recognition is dependent on identity elements, which often include the anticodon and discriminator base at position 73, as well as other single nucleotides and nucleotide pairs in some cases[41]. Importantly, the ribosome does not select tRNAs based on the amino acid they are charged with. This permits incorrectly charged tRNAs to misincorporate amino acids into the nascent polypeptide, which can severely impact protein function and proteostasis in general.

Mechanisms of tRNA regulation

Since early discoveries using microarrays[8,9], and the subsequent development of high-throughput sequencing-based technologies that allow quantitation of cellular tRNA pools[10–12,14], the conundrum of context-specific tRNA regulation in multicellular eukaryotes has puzzled investigators. In general, there is some contention in the field regarding the function and requirement (if any) for the specific regulation of individual tRNA transcripts in different cell contexts or stress.

Gingold et al. suggested that proliferating and differentiating cell types have distinct transcriptomes with distinct synonymous codon usages, and that the expressed tRNA anticodon pools within these cells are matched to meet their unique cellular codon demands[15]. Contrary to this, it has been shown that despite divergent protein-coding gene expression between mouse brain and liver tissue, relative codon demand remains stable. Furthermore, tRNA gene expression in these same tissues measured by RNAPIII ChIP-seq, generating maps of occupancy for the RNA polymerase, also revealed stable anticodon pools despite drastically different transcriptional programs in these cells[16].

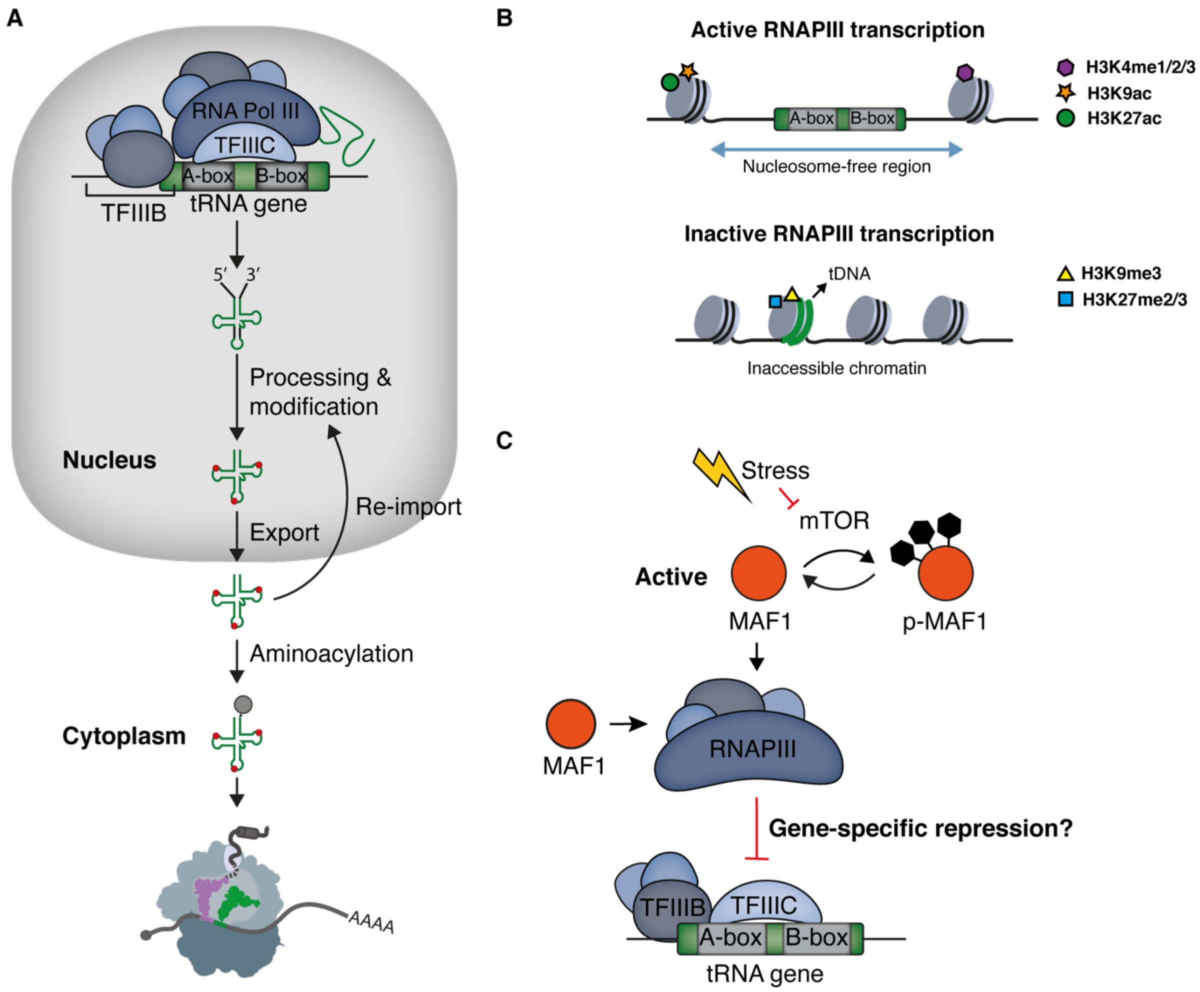


Figure 2 tRNA transcription, maturation, and regulation.
(A) Brief overview of tRNA biogenesis and maturation. Transcription occurs in the nucleus via recruitment of the transcription factors TFIIIC and TFIIIB, resulting in RNA polymerase III (RNAPIII) recruitment and tRNA transcription. tRNA leader and trailer sequence removal, and in some organisms, intron removal occurs in the nucleus, along with the addition of some modifications. tRNAs are exported for aminoacylation to occur, and may be reimported for final addition of specific modifications. Functional, mature, and charged tRNAs can participate in translation by binding with codons in mRNA sequence once loaded into an empty ribosomal A-site.
(B) tRNA transcriptional regulation is partly regulated by chromatin dynamics, accessibility, and histone modification. Actively transcribed tRNAs have been associated with nucleosome-free regions (NFRs), and specific histone 3 lysine modifications which also mark RNA polymerase II coding genes for active transcription (top panel). Conversely, inactive tRNAs are associated with condensed heterochromatin (both facultative and constitutive), and are marked with known repressive histone marks (lower panel)[43].
(C) MAF1 acts as the only known global RNAPIII-specific repressor. Under normal growth, and in certain cell types, MAF1 is phosphorylated by mTOR at one or more of its three main phospho-sites, leading to its inability to interact with, and repress RNAPIII. Under stress conditions, nutrient deprivation, and in certain differentiated cell types, mTOR is inactivated leading to MAF1 dephosphorylation and activation such that it is able to repress RNAPIII recruitment to tRNA genes.

Since a minority of mRNA transcripts from coding genes show tissue- or cell-type specific expression in most cases[42], global estimates of synonymous codon usage might remain stable when considering all transcripts in a sample. Perhaps optimized codon demand in a subset of context-specific transcripts impacts their translation more drastically than other housekeeping or ubiquitously expressed genes by ensuring their codon usage more closely matches the availability of tRNAs. At the gene and transcript level, however, tRNA differential regulation is well-established. Mechanisms of this regulation are an area of continued interest, while the function of this regulation - especially in light of stable anticodon expression - is far less clear.

Mechanisms of RNAPIII regulation at tRNA genes

Growing evidence shows that epigenetic modifications influence chromatin state and RNAPIII transcription similarly to that seen for RNAPII transcription at protein-coding genes (**Figure 2B**; reviewed in [43]). For example, expressed tRNAs are found in nucleosome-free regions (NFRs) that are flanked up- and downstream by nucleosomes[44,45] and they exhibit histone modifications associated with active transcription and open chromatin, such as H3K4me1/2/3, H3K9ac and H3K27ac (**Figure 2B**). By contrast, inactive tRNAs are found in heterochromatic regions marked by repressive histone modifications such as H3K27me2/3 and H3K9me3 (**Figure 2B**)[46,47]. Changing chromatin states across development, differentiation, or under different stress conditions surely contributes to the regulation of tRNA genes, yet the factors regulating these states remain unknown.

What is less clear is the extent that sequence-dependent mechanisms contribute to tRNA transcriptional regulation. The A- and B-box promoters overlap with strongly conserved structural elements in the D- and T-arms[28], and show little variation in sequence even between sets of tRNAs that are constitutively bound by RNAPIII and those that are lowly occupied or unoccupied[46,48]. Furthermore, although the TATA-box-binding protein (TBP) required by all eukaryotic RNAPs is a subunit of TFIIIB, the presence of a TATA-like element in tRNA upstream regions seems to be specific to only some eukaryotes including plants[49], insects[50], and the fission yeast *Schizosaccharomyces pombe*[51,52], while animal sequences are heterogenous with no discernible motifs found in these regions so far. Despite this, even tRNA gene copies that produce identical transcripts have shown variable RNAPIII occupancy, possibly implicating divergent upstream sequence and TFIIIB affinity underlying this regulation[53]. Perhaps the use of newer technologies, such as neural convolutional networks (CNNs) combined with high-resolution occupancy maps in multiple cell or tissue types[54] can accurately model and predict upstream motifs underlying the regulation of tRNA genes.

The general RNAPIII repressor, MAF1

So far, the protein MAF1 remains the most well-characterized general repressor of RNAPIII (**Figure 2C**)[55]. MAF1 is highly conserved in eukaryotes from yeast to humans, and its activity as a RNAPIII repressor is dependent on phosphorylation status[36]. Only dephosphorylated MAF1 is able to bind RNAPIII and repress transcription[56]. In human, the mammalian target of rapamycin (mTOR) kinase is responsible for the phosphorylation of MAF1 at three main phosphorylation sites during normal growth conditions[57]. Phosphorylated MAF1 is inactive and destabilized, while under stress conditions or nutrient deprivation, MAF1 is dephosphorylated and is able to inhibit transcription (**Figure 2C;** reviewed in [36]).

The structural basis of MAF1 interaction with RNAPIII has recently suggested the possible mechanism underlying MAF1 repression[58,59]. Active MAF1 binds the RNAPIII subunit RPC1 at the clamp domain, and overlaps with the interaction domain of the TFIIIB subunit, BRF1. This explains why MAF1 and TFIIIB binding to RNAPIII is mutually exclusive, and proposes that TFIIIB recruitment of RNAPIII to tRNA loci is inhibited by MAF1 occupancy[59,60].

Despite advances in understanding MAF1-mediated repression, the differences in responsiveness of certain tRNA subsets to MAF1 remains elusive. In yeast, a so-called “housekeeping” set of tRNAs exists, which are apparently less sensitive to MAF1 repression and environmental signals, and proposes more complex mechanisms underlying regulation[61]. This suggests a combinatorial effect of chromatin dynamics, MAF1-mediated silencing, and potential, unknown sequence-dependent mechanisms underlying the heterogenous regulation at different tRNA loci.

tRNA in health, disease, and therapeutics

A multitude of pathologies and diseases are associated with tRNA-related defects in eukaryotes, underpinning the need for accurate methods to globally analyze tRNA repertoires. Understanding the nature and mechanism of complex tRNA synthesis and turnover is essential in understanding their role in complex disease phenotypes, and in guiding novel treatments and therapies. tRNA-related defects often present in a surprisingly tissue-specific manner; for example, defects in tRNA expression, modifications, aminoacylation, and processing have often been linked to abnormal phenotypes and development in neural cell types, and even to neuropsychiatric disease (reviewed in [19]), underscoring the increased susceptibility of specific tissues and cell types to tRNA-related defects[62]. The exact mechanism of this phenomenon remains unclear and an active area of research in the field, particularly in the context of treatment and tRNA-based therapies.

tRNA-related pathology research has implicated nearly all major steps of tRNA synthesis, maturation, and turnover in the cause of disease, emphasizing the requirement for tight control over each of these processes and their essentiality in multicellular eukaryotes. Briefly reviewed below are some examples of such abnormalities in tRNA abundance and modifications, with the focus on highlighting various mechanisms that tRNA-related abnormalities result in complex and context-specific phenotypes. Novel therapeutic avenues exploiting tRNAs are also briefly discussed.

Perturbed tRNA regulation and expression

Despite the relatively stable anticodon pools between different cell types, tRNA isodecoders show surprising context-specific expression, implicating their role outside of functional redundancy with other members of a given isodecoder family[8,10–12]. In some cases, the limited expression of only one or a few isodecoders in a specific cellular context prevents compensation of potentially deleterious tRNA dysregulation by other isodecoders. The abundance of a highly CNS-specific tRNA-Arg-UCU in mouse, *n-Tr20* (tRNA-Arg-TCT-4–1), is significantly reduced in a specific genetic background with a SNP in this gene (**Figure 3A**)[63]. This isodecoder is one of five for this particular family but accounts for ~60% of the expression of tRNA-Arg-UCU in mouse brain tissue. The decreased abundance of *n-Tr20* in the mutant background, combined with a loss-of-function GTPBP2 mutation, causes significant ribosome stalling at AGA codons in mouse brain tissue that cannot be resolved. This leads to eIF2α phosphorylation, the activation of the integrated stress response (ISR), and a global decrease in translation resulting in significant neurodegeneration and death within two months after birth (**Figure 3A**)[63,64].

Surprisingly, transgenic expression of any of the other isodecoders rescues the phenotype of *n-Tr20* loss, suggesting that the abundance of individual tRNA isodecoders and their sequence outside of the anticodon is not as important as the overall abundance of tRNAs with the same anticodon[64]. This is in line with the reproducible observation that anticodon pools are stably expressed despite significant isodecoder regulation.

This result highlights that tRNA regulation acts to fine-tune translation in a tissue-specific manner. Even when tRNA abundance is not perturbed, silent SNPs (sSNPs) that cause synonymous codon changes in mRNA may result in the use of a rare isoacceptor and have knock-on effects on translation and protein structure acquisition. Indeed, this phenomenon has been linked to aberrant protein cotranslational folding and function in a tRNA-concentration dependent manner, relevant to cystic fibrosis in humans[65].

Conversely, changes to the abundance or function of anticodon pools can have drastic effects on codon-biased translation[66]. This phenomenon has been noticed numerous times in various cancers[67,68]. For example, overexpression of tRNA-Glu-UUC and tRNA-Arg-CCG leads to augmented translation of transcripts enriched in their cognate codons, driving pro-metastatic programs in breast cancer[68]. Furthermore, a meta-analysis of tRNA expression profiles in The Cancer Genome Atlas across 31 cancer types found that, in general, dysregulated tRNA expression is widespread in cancer[67]. This was also complemented by increased expression of tRNA aminoacyl synthetases and modification enzymes, indicating that multiple steps of tRNA biosynthesis are upregulated in cancers, driving the biased translation of oncogenic transcripts.

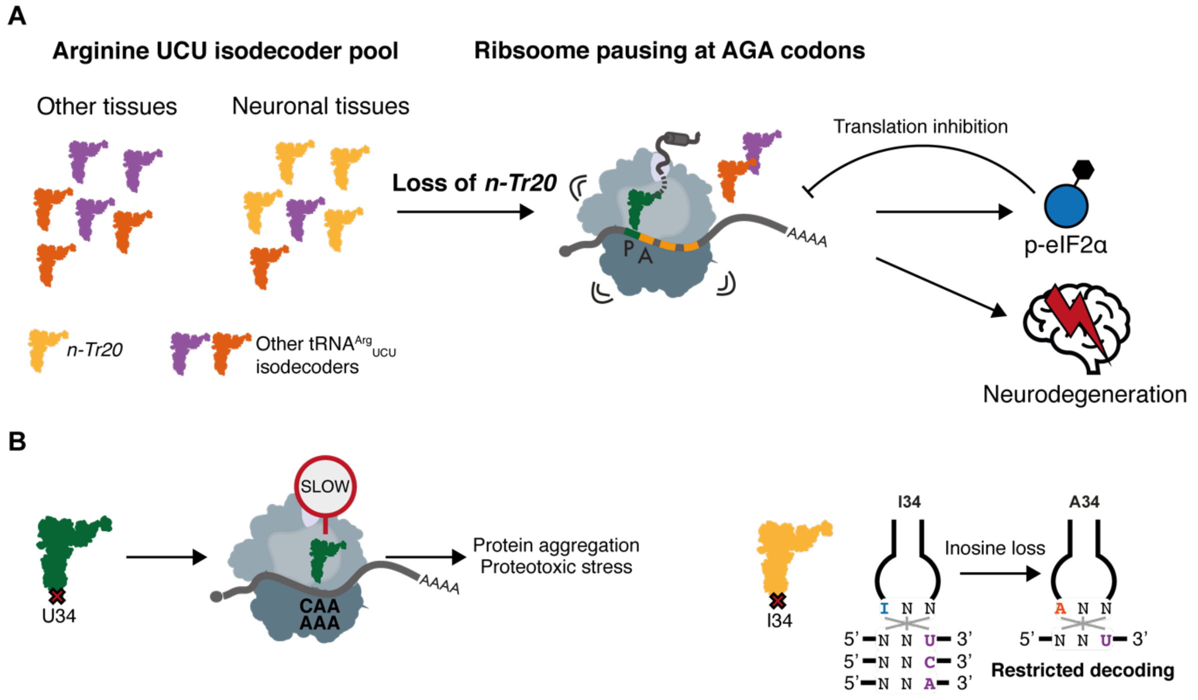


Figure 3 Abnormalities in tRNA abundance or modification status affect translation and normal physiology. (A) Mutations in tRNA genes can lead to lowered expression, for example a mutation in *n-Tr20* in B6J mice. This isodecoder is highly neuronal-specific and constitutes ~60% of all $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ expression in these tissues. Loss of this tRNA in a GTPBP2 loss-of-function background results in significant ribosome pauses at AGA codons, activation of the integrated stress response (ISR) via eIF2α phosphorylation, and neurodegeneration. Adapted from Kapur et al. (2020)[64].
(B) Defective tRNA modifications at position 34 of the anticodon can alter the decoding capacity of some tRNAs. Show are examples of uridine 34 loss (left) affecting CAA and AAA decoding resulting in toxic protein aggregation[73–75], and how inosine 34 (I34) allows wobble decoding for C and A nucleotides in the first codon position (right). Adapted from Nedialkova et al. (2015) [75] and Suzuki (2021)[21].

Although there are some studies investigating the ramifications on health and disease of mutations in tRNA genes, mutations in these sequences are generally under very strong purifying selection. Interestingly, the 5' flanking region of ~20 bp upstream of tRNA genes show the highest rates of sequence variation[69]. Little is known about the sequence dependence of TFIIIB binding to upstream regions, and although some evidence suggests that mutations in upstream regions can affect human tRNA gene expression in HeLa cells[70], direct testing of this hypothesis has not been addressed. Perhaps variability in upstream regions can affect transcription, however the mechanism and potential effects of this remain unknown.

tRNA modification abnormalities

Nucleoside modifications in mature tRNA are required to confer multiple functions and structural properties to these essential molecules, from providing stability[2], to enabling wobble pairing and modulating the decoding capacities during anticodon-codon base pairing[3], and regulating reading frame maintenance of the ribosome during translation[39,71]. Therefore, it is not surprising that some modification defects result in serious phenotypic manifestations with diverse mechanisms, often referred to collectively as "modopathies"[21,72].

Modification defects resulting in impediments to translation have been extensively studied and documented for cytosolic tRNAs. For example, loss of wobble position 34 modifications such as 5-methoxycarbonylmethyluridine (mcm^5U34) and 5-methoxycarbonylmethyl-2-thiouridine (mcm^5s^2U34) have been shown to impair codon-specific translation, resulting in protein aggregation and proteotoxic stress (**Figure 3B**)[73–75]. Deficiencies in mcm^5U34 are also linked to the occurrence of familial dysautonomia[76], a genetic disorder affecting survival and development of neurons in the autonomic and sensory nervous system. Deficiency of conserved wybutosine (yW) derivatives at tRNA position 37, specifically in eukaryotic tRNA-Phe has been found in a variety of cancers[77–79]. Hypomodified tRNA-Phe induces -1 frameshifting in ribosomes causing mRNA nonsense-mediated decay (NMD) of transcripts[80], potentially underlying the mechanism of disease. In eukaryotes, all eight tRNAs with an adenosine at the wobble position 34 are deaminated to produce inosine[81]. This is crucial to ensure wobble pairing with U, C or A in the third position of the codon, and removal of I34 modifications is not viable, highlighting the essentiality of inosine for correct decoding (**Figure 3B**)[5,6].

Some of the most notable tRNA modopathies studied so far result from defects in mitochondrial tRNA modifications. A set of 22 tRNA species are encoded entirely within the mitochondrial genome (mt-tRNAs), and are essential for the decoding and translation of the 13 mitochondrial-encoded oxidative phosphorylation (OXPHOS) complex polypeptides, essential for cellular ATP production. Therefore, defects in mitochondrial protein synthesis can result in severe phenotypes, especially in high-energy-consuming cell types and tissues.

Indeed, tRNA modification defects that result in perturbed function have been shown to result in complex phenotypic manifestations (reviewed in [82]). Two of the most well-studied pathogenic mutations in mt-tRNAs, namely 3243A>G (within MT-TL1, encoding mt-tRNA-Leu-UUR) and 8344A>G (within MT-TK, encoding mt-tRNA-Lys), which together account for ~85% of all mt-tRNA-related mitochondrial disease, both present with disrupted taurine modification levels at the wobble position 34 of the tRNA[82,83]. In the case of the 3243A>G mutation, 5-taurinomethyluridine (τm^5U) hypomodification at position 34 significantly affects this tRNAs ability to decode UUG codons[84]. Ribosome profiling experiments show reduced translation speed and mitoribosome accumulation at these codons[85]. Consequently, biosynthesis of a subunit of the respiratory chain complex I, ND6, is reduced in patients with the mutation[86]. This gene has the highest usage of UUG codons amongst all 13 mitochondrially encoded proteins. About 80% of patients with mitochondrial encephalomyopathy, lactic acidosis and stroke- like episodes (MELAS), a severe maternally-inherited nervous and muscular disorder, present with this particular mt-tRNA-Leu-UUR 3243A>G mutation[87].

Similarly, 8344A>G in mt-tRNA^{Lys} results in decreased 5-taurinomethyl-2-thiouridine (τm^5s^2U) modification at the wobble position[84]. However, the loss of this modification enables superwobbling and potential misreading of asparagine codons AAC and AAU by this lysine tRNA, and subsequent amino acid misincorporation[88]. Additionally, defective aminoacylation and altered abundance of the mature tRNA, along with modification defects, have all been shown to contribute to the strong translational defects of mitochondrial transcripts and manifest as myoclonus epilepsy, ragged-red fibers (MERRF) syndrome in some patients with the mutation[89,90].

These examples illustrate the diverse mechanisms underlying tRNA modification deficiency-linked disorders. The importance and pervasiveness of tRNA modifications means that epitranscriptomic investigation in these molecules is of increasing interest for investigators seeking to understand cause, effect and mechanism of tRNA modification perturbations. High-throughput next-generation sequencing has already been applied in numerous ground-breaking ways to globally assess modification status and abundance in RNA molecules[91–93], however various technical challenges have so far limited sensitivity and accuracy of these methods[94]. Nevertheless, by overcoming these challenges it is foreseeable that generating personalized “epi-tRNAomes” may become feasible, and could act as new tools for patient-specific diagnosis and treatment in human health and disease[95].

tRNAs in therapeutics

Leveraging tRNAs for therapy has generated significant biotechnological interest recently[22]. Of particular interest has been the repurposing of suppressor tRNAs (sup-tRNAs) as treatment for inherited diseases, such as β-thalassemia, muscular dystrophy, Rett syndrome, and cystic fibrosis (**Figure 4**)[96–99]. Sup-tRNAs arise naturally in some species by mutations in tRNA anticodons, allowing them recognize a stop codon instead[100]. This can facilitate nonsense mutation readthrough that would otherwise be detrimental. Instead, an amino acid is incorporated at the premature stop codon mitigating translation termination by preventing release factor (RF) binding. This concept is now being utilized for the design and delivery of custom sup-tRNAs targeted to overcome nonsense-induced diseases in humans, which collectively account for up to 11% of inherited disease (**Figure 4**)[101].

Indeed, one study that characterized a library of all possible tRNAs capable of suppressing premature termination codons, called anticodon-edited tRNAs (ACE-tRNAs)[97], has been licensed by Tevard Biosciences in its aim to design therapeutic agents for Dravet syndrome, among others. Dravet syndrome, a rare form of epilepsy, is sometimes caused by nonsense mutations in the sodium channel gene *SCN1A* causing premature stops[102], implicating sup-tRNAs as prime candidates for treatment.

However, most-often the disease is caused by heterozygous loss-of-function mutations in *SCN1A.* For these patients, Tevard is working on using a combination of three “enhancer” tRNAs, specifically tailored to the codon usage of *SCN1A,* which are apparently able to double protein production from the functional copy of the gene[22]. This technology highlights a way in which tRNAs might be exogenously administered to exploit the codon optimality of genes involved in disease, and thereby increase protein expression from functional alleles.

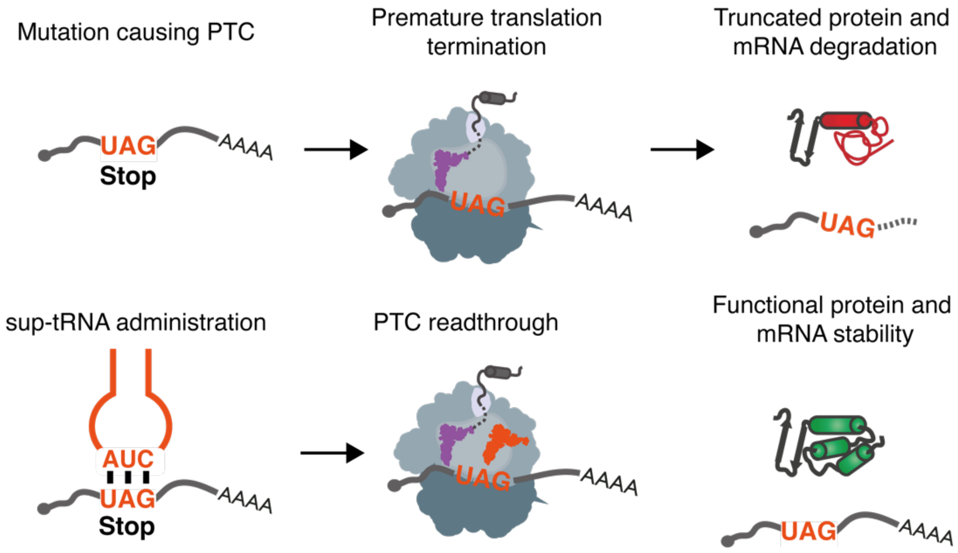

Mutation causing PTC
Premature translation termination
Truncated protein and mRNA degradation
UAG
Stop
AAAA
UAG
AAAA
UAG
sup-tRNA administration
PTC readthrough
Functional protein and mRNA stability
AUC
UAG
Stop
AAAA
UAG
AAAA
UAG
AAAA

Figure 4 Schematic representation of suppressor tRNA (sup-tRNA) activity in premature termination codon (PTC) readthrough as therapeutic targets for disease. Mutations that cause PTCs in coding transcripts result in premature termination to translation, production of non-functional, truncated protein products and can result in mRNA degradation through the nonsense-mediated decay (NMD) pathway (top). Nonsense and PTC-induced diseases account for ~11% of inherited human disease. Leveraging suppressor tRNAs can mitigate nonsense codon termination of translation and relieve disease symptoms (bottom). Anticodons of functional, aminoacylated tRNAs are edited to read and decode stop codons, for example UAG. This leads to PTC readthrough, amino acid incorporation, and functional protein products.

One concern with sup-tRNA therapies is the off-target readthrough of native stop codons. However, ribosome dynamics, RNA binding protein recognition, and genomic context at native stop codons seem to discourage readthrough by sup-tRNAs compared to nonsense stops upstream[97]. In other words, native stop codons, having been evolved for the specific task of terminating translation, might be much better targets to catalyze termination than errors that lead to missense stops upstream. Other challenges to these therapies include a suitable delivery system, however, recombinant adeno-associated virus (rAAV) systems have already shown promise in this regard[103]. Efficiently returning protein abundance to normal levels without overexpressing them to toxic quantities also remains a concern. Despite these challenges, tRNA-based therapies have seen significant recent academic and industry interest and have a growing compendium of research that bolsters their efficacy in the treatment of genetic disorders.

In a recent example, investigators found that disease-causing variants of glycyl-tRNA synthetases (GlyRS) cause the enzyme to sequester tRNA-Gly but not release it, leading to translation defects and Charcot-Marie-Tooth peripheral neuropathy[104]. Accordingly, ribosome pausing at cognate codons and activation of the integrated stress response was noted in motor neurons in mice. Strikingly, transgenic overexpression of tRNA-Gly rescued protein synthesis in flies and mice, and attenuated neuropathic phenotypes[104].

tRNA isodecoder expression is dynamic and cell-context specific, and therefore tRNA-based therapies, such as those described above, can have variable efficacies depending on these expression profiles. A high-resolution, accurate method for generating snapshots of the tRNA landscape in an affected sample could certainly guide the choice for therapeutic tRNA targets. Isodecoders encoding the correct amino acid and expressed to the desired level to enable nonsense readthrough as engineered sup-tRNAs could be identified, or tRNAs with pathogenically low expression or availability that would benefit from exogenous tRNA expression or administration could be determined.

Methods for tRNA transcriptome and modification profiling

Understanding the regulation, expression, aminoacylation and modification status of tRNAs has proven extremely useful in discovering the basic biology and contribution of each of these factors to normal tRNA function and the causal links to abnormal tRNA-dependent physiology. As such, developing methods enabling investigators to globally query tRNA abundance and modification status have become a focus in the field in the last decade. However, due to their structure, heavily modified status, and high levels of redundancy and duplication, the process of developing methods for quantifying the levels of individual tRNAs in cells has been fraught with technical challenges.

Initially, before next-generation sequencing (NGS) was widely accessible and affordable, hybridization-based microarray approaches were used to quantify cellular tRNA pools in bacteria and human tissues and cell lines[8,9,105]. This method relies the hybridization of two sets of fluorescently labeled tRNA extracts from different samples to a set of distinct DNA probes, allowing relative tRNA quantification between samples. This method poses several problems for accurate quantitation; firstly, a difference of at least eight nucleotides is needed to prevent cross-hybridization of similar tRNAs to a probe[8]. In eukaryotes, extensive tRNA gene sequence similarity means that tRNAs can differ by only one nucleotide, even between isoacceptors for a specific amino acid[7], severely limiting the accuracy and resolution of hybridization-based approaches for quantitation of unique tRNA transcripts. Secondly, hybridization efficiency is hampered further by pervasive Watson-Crick face nucleotide modifications present in all tRNA transcripts.

Because of these limitations, focus was shifted to improving the efficiency of tRNA quantitation by high-throughput sequencing (tRNA-seq). On the one hand, tRNA-seq is well suited to tRNA quantitation; it overcomes the necessity for nonspecific probes and semi-quantitative nature of hybridization-based techniques. Furthermore, read length produced by NGS approaches is limited by its chemistry to ~200nt, and while potentially limiting for other longer RNA species (e.g., protein-coding transcripts or long-noncoding RNA [lncRNA]), mature tRNA are only 70 – 90nt long. At least theoretically, this means full tRNA transcripts can be sequenced as individual reads by this approach, negating the requirement for assembling transcripts from multiple reads computationally prior to quantitation.

Technical challenges to current tRNA-seq methods

Developments in tRNA-seq methods have faced several challenges specific to the characteristics of tRNA molecules and their organization in eukaryotic genomes. Stable secondary structure and Watson-Crick modifications pose physical barriers that block RT (**Figure 5A**), resulting in short, prematurely terminated cDNA molecules with coverage bias at the 3' end of tRNAs where RT is primed (**Figure 5B**; right). This is especially problematic considering different tRNA transcripts have different distributions and stoichiometries of modifications, resulting in differences in RT efficiency and, therefore, differences in quantitation accuracy among them. While, computationally, extensive tRNA sequence similarity and multicopy gene families in eukaryotes result in significant read alignment ambiguity when attempting to map sequence reads back to tRNA transcripts. These challenges and potential solutions are discussed in detail below.

Reverse transcription and efficient cDNA synthesis

Some significant advances have overcome impediments to generating cDNA libraries from tRNA to varying degrees. The enzymatic removal of some tRNA modifications in demethylase-thermostable group II intron RT tRNA sequencing (DM-tRNA-seq)[10], and AlkB-facilitated RNA methylation sequencing (ARM-seq)[14] has been used to prevent premature termination of RT. *E. coli* AlkB is a dealkylating enzyme, which has been shown to remove some of the most common and conserved tRNA base methylations, such as N^1-methyladenosine (m^{1}A), often found at position 58 in many eukaryotic tRNAs, and N^3-methylcytosine (m^{3}C) found at position 32 of five tRNA isoacceptors[106–108]. Some of the most conserved and prevalent Wastson-Crick face modifications in eukaryotic tRNAs are shown in **Figure 5A.** An engineered version of AlkB with an amino acid mutation, AlkB D135S, able to additionally remove N^1-methylguanosine (m^{1}G), has been used in combination with wild-type AlkB to increase the removal of RT-impeding modifications[10]. More recently, another AlkB mutant D135S/L118V has shown extended activity in removing N^2,N^2-dimethylguanosine ($m^2{}_2G$)[109].

However, modification removal poses several concerns; although improvements have been made via engineered forms of AlkB, there is still significant variation in demethylase efficiency amongst modification types and amongst different tRNA transcripts[10,109]. This certainly leads to RT biases for transcripts with less Watson-Crick face modifications, or those that are better substrates for AlkB and its engineered forms. Furthermore, without a comparative approach requiring multiple sequencing libraries per sample (i.e., with and without enzyme treatment) and complex data analysis of sequence datasets[110], removal of such modifications prohibits their analysis and quantitation in endogenous tRNA pools.

Other approaches to relieving blocks to RT include tRNA fragmentation by partial alkaline hydrolysis in Hydro-tRNAseq[12,111,112]. This method aims to generate multiple smaller RNA fragments from tRNA transcripts prior to cDNA synthesis to negate effects of tRNA secondary structure on RT efficiency. In an alternative approach, YAMAT-seq[13] and QuantM-seq[11] attempt to improve RT adapter ligation efficiency and recovery of full-length cDNA transcripts from intact tRNA transcripts. By exploiting the conserved single-stranded 3'-CCA ends of functional, mature tRNA, these methods utilize double-stranded adapters with overhangs that are splint-ligated to tRNA 5' and 3' ends simultaneously. These methods do not try to account for modification-induced stops to RT, and therefore either result in libraries of short reads from prematurely aborted cDNA synthesis, or an enrichment of lowly modified transcripts where premature stops to RT are less likely.

The most promising development in tRNA library generation, specifically in optimizing the RT reaction, has been in the utilization of a thermostable group II intron RT (TGIRT) for highly processive cDNA synthesis at elevated temperatures (**Figure 5B**; left)[113–115]. Group II introns are a class of bacterial retroelement that propagate through a host genome by a method known as retrohoming[116]. This requires accurate synthesis of cDNA from the mobile intron element, which is typically highly structured and >2 kb long, requiring high fidelity and processivity from the RT.

Additionally, unlike more common retroviral RTs, these enzymes were found to have template-switching abilities, allowing them to switch from RNA-DNA hybrid primers to new RNA templates, facilitating linking target RNA to an adapter amenable for sequencing without the need for potentially biased ligation reactions[113]. These features led to the use of TGIRT for generation of full-length cDNA from highly structured and modified tRNA with reduced propensity for stops during RT, subsequently shown somewhat effective in methods such as TGIRT-seq and DM-tRNAseq[10,113,114]. Despite this, the efficiency of the RT reaction with

TGIRT has still been shown to be relatively low and highly variable on modified, endogenous tRNA pools[10].

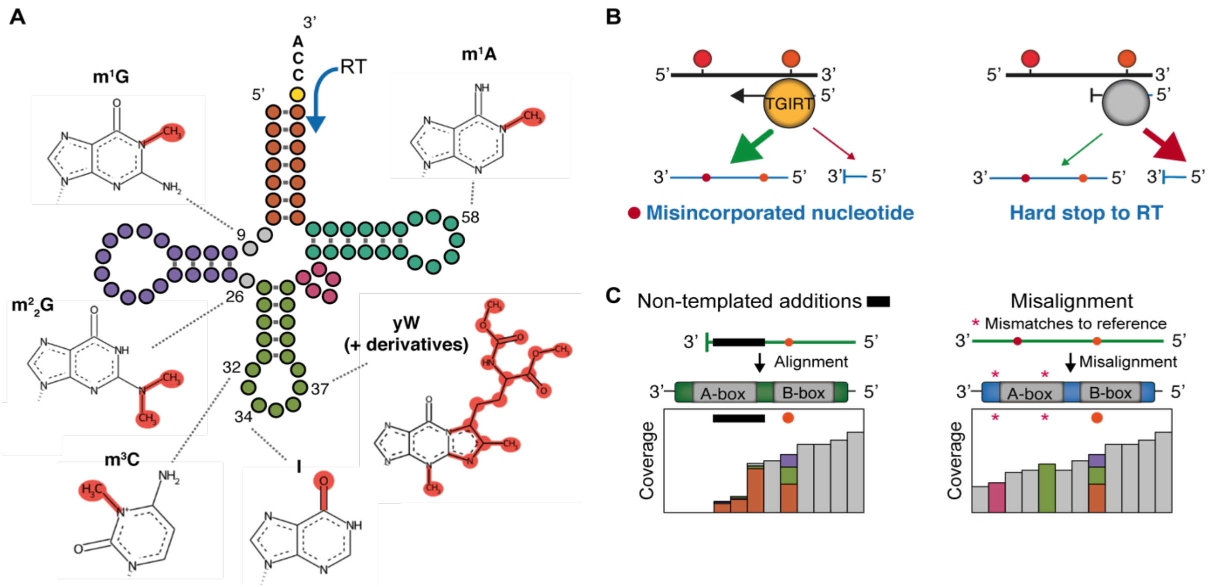


Figure 5 Challenges to tRNA reverse transcription and modification analysis.
(A) Schematics representing common barriers to reverse transcriptases in generating cDNA including tRNA secondary structure, and prevalent and conserved Watson-Crick face nucleotide modifications. Several well-known and conserved modifications that posing significant blocks to RT are illustrated. m^{1}G: *N1*-methylguanosine; m$^2_2$G : *N2,N2*-dimethylguanosine; m^{3}C: N3-methylcytosine; I: inosine; yW: wybutosine; m^{1}A: *N1*-methyladenosine.
(B) cDNA synthesis with TGIRT improves modification readthrough, resulting in reads with misincorporations at Watson-Crick face modified sites (left). More common retroviral RTs are less processive and are less able to readthrough modifications, resulting in hard stops to RT and 3' coverage bias.
(C) Difficulties with accurate tRNA modification calling and analysis resulting from non-templated nucleotide additions by RTs (black box; left) and misalignment to similar tRNA references (asterisks indicate sequence differences between references; right). In both cases, risk of false identification of modified sites may result from mismatch profiles that present as potential misincorporations. True misincorporation signatures from modified sites indicated with orange dot.

Modification analysis using misincorporation signatures

The processivity of TGIRT also improved on previous attempts in investigating, predicting, and measuring misincorporation-inducing modifications in tRNA molecules[110,117]. Typically, modifications at the Watson-Crick face of nucleotides physically impede commonly used RTs of retroviral origin and result in much lower levels of misincorporation. However, increased processivity of novel RTs, such as TGIRT, have shown promise in increased modification readthrough, often resulting in significant misincorporation at modified positions (**Figure 5B**; left).

The benefit is twofold; a larger proportion of cDNA representing full-length tRNA transcripts is generated, reducing the typical 3' sequence coverage bias seen with standard RTs such as SuperScript III[12,14] and resulting in a more representative sequencing library. Secondly, modifications can be identified and studied as sites showing sequence polymorphism, or mismatch to the reference sequence. Despite this, current methods using

TGIRT relieve only a proportion of stops to RT from modified sites, requiring the combined analysis of misincorporation and RT stop frequency for modification abundance estimation[110]. RT stops may arise from multiple sources however, such as RNA structure and degradation, and might also represent imprecise positions of modifications along an RNA transcript. Enabling efficient readthrough, and removing the need for analysis of stops would significantly improve the accuracy of modification identity and stoichiometry analysis.

Even with improved readthrough, leveraging misincorporations for modification analysis is often not as simple as searching for mismatches. Many experimental, sequencing, and read alignment artifacts and biases can result in false positive identification of modified sites, and reduced signal:noise ratios making accurate modification calling difficult (**Figure 5C**)[117,118]. For example, RTs are prone to non-templated nucleotide addition at cDNA 3' ends, which can be falsely identified as misincorporation signatures (**Figure 5C**; left)[119–121]. An initial misincorporation-based map of m^1A modifications in the human transcriptome found nearly 500 putative modified sites[122]. However, reanalysis of this data found that ~10% of these sites might represent true modifications, while up to 50% were falsely identified non-templated additions at cDNA 3' ends[123].

Another significant proportion of these false positives in this study likely arose from misalignment of reads containing misincorporations to an incorrect reference (**Figure 5C**; right). Short-read alignment algorithms that attempt to find the best positional match for a sequencing read in a reference sequence space can have a particularly arduous task with reads containing mismatches. This is complicated by the numerous sources and frequencies of mismatches in reads, which include; modified RNA residues that induce misincorporations (to varying degrees and proficiency depending on RT choice and reaction conditions), innate error rates of reverse transcriptases ($\sim 1\times10^{-4}$)[124,125], and NGS sequencing platforms such as Illumina (per-base error of 1×10^{-2} - 1×10^{-3})[126], to name a few. These extra sources of variation can be difficult to distinguish from true modification-induced mismatches. In the case of tRNAs, where gene redundancy and similarity, and the propensity for misalignment is high, such cases often present with highly homozygous mismatches at unexpected positions along tRNA genes (**Figure 5C**; right)

Despite this, progress has been made in identifying sources of bias and error, and in developing potential solutions using computational methods for RNA modification detection in sequencing datasets[118,127,128]. For example, it is now known that modification-induced misincorporations are rarely characterized by the misincorporation of only one other nucleotide, with the exception of inosine, which is always read as guanosine (G) during sequencing. Instead, complex signatures are frequently seen that may be modification, sequence-context, and reaction condition-specific. One method, high-throughput annotation of modified ribonucleotides (HAMR)[129], uses a binomial test for significance on mismatch frequencies at individual positions of annotated RNAs using aligned RNA-seq data. HAMR tests a conservative null hypothesis that only the true genotype can be biallelic. This means that sites that look like heterozygous and homozygous SNPs are excluded from being called as significant, and only misincorporations present as composite signatures of multiple misincoporated nucleotides are considered[129].

tRNA modification analysis, in comparison to RNA modification analysis in other RNA types, is especially complicated for two primary reasons: 1) low sequence diversity and high duplication rates amongst tRNA genes[7] result in significant difficulties in unambiguous alignment, increasing misalignment and spurious mismatch calling. 2) As the most modified RNA in any cell, expected misincorporations at Watson-Crick face modifications range between 0 and ~7 per tRNA molecule in human, for example[108]. This requires significantly elevated fidelity and processivity during RT, and additional care in computational analysis and interpretation, which, for the most part have been lacking so far[117]. These factors bolster the requirement for optimized sequencing library preparation protocols, and dedicated computational tools and algorithms tailored to the processing and analysis of complex tRNA sequencing datasets.

Computational challenges to tRNA sequencing data analysis

Advances in tRNA-seq methodology have primarily focused on the biochemistry associated with producing bias-free, representative cDNA libraries from purified tRNA pools for sequencing purposes. Very little emphasis has been placed on the concurrent development of appropriate computational tools and statistical frameworks needed to extract estimates of abundance, modification, charging, and general metrics of library quality control (QC) from the resulting datasets, while addressing the particularities of tRNA-seq data.

This oversight in algorithm and computational method development has resulted in the use of more routine analyses of tRNA sequencing datasets, such as those used for transcriptome analysis[10,11]. These are prone to biases when analyzing tRNA-seq data that are not as great a concern for other datasets, particularly with regards to alignment using popular short-read alignment algorithms (**Figure 6**). Two of the major unaddressed concerns for tRNA-seq data analysis include; 1) generating appropriate sequence references that minimize redundancy and ambiguity in read alignment while maintaining resolution for tRNA transcripts, and 2) handling reads with numerous misincorporations at modified tRNA sites (and terminally added non-templated nucleotides from RT activity) without resulting in significant data loss.

tRNA reference choice for alignment

The choice of reference sequences for sequencing data alignment can have pronounced effects on the accuracy and sensitivity of the aligner, the proportion of usable or “mappable” data, and the resolution of quantitation[130]. Aligning reads to the full set of predicted tRNA genes would ideally offer the best resolution and most informative results. However, in eukaryotes, redundancy of tRNA gene sequences at conserved structural elements, gene duplication, low sequence diversity within and between isodecoder families (**Figure 6A**), and significant processing of pre-tRNA transcripts (**Figure 2A**) all contribute to the difficulty associated with unambiguously aligning reads derived from mature tRNA to one locus of origin, most often resulting in multi-mapping reads[130,131].

Of course, the biological question and library generation protocol must be considered when deciding on an appropriate reference, and this should guide the choice. For example, limiting the reference to only mature, intronless tRNA sequences lacking 5' leader and 3' trailer sequences can be extremely beneficial for accuracy and speed, but will not be informative if one wishes to investigate pre-tRNAs. In this case, genomic tRNA sequences with flanking sequence, or even a full genome reference would be more appropriate. Additionally, eukaryotic tRNAs are appended with the conserved 3'-CCA tail post-transcriptionally while some archaeal and most eubacterial CCA ends are genomically encoded[132], a factor that needs to be considered when designing a reference to ensure efficient alignment at tRNA 3' ends.

Potentially the most advantageous approach to reduce alignment ambiguity and multi-mapping has been through tRNA clustering, wherein similar tRNA sequences are collapsed and used as alignment references[129–131]. This is particularly effective in reducing redundancy in complex eukaryotic genomes where whole-genome duplication, tRNA gene duplication, and tRNA pseudogenization are more pronounced. For example, in human, the reference genome hg38 includes 620 predicted tRNA genes[7]. Collapsing identical mature tRNA transcripts results in 424 unique sequences, of which 72 represent duplicate tRNAs with tRNA-Asp-GTC-2 being the largest with 11 loci throughout the genome producing identical mature tRNAs. In zebrafish, this redundancy is compounded, potentially by an extra whole-genome duplication event in fish relative to other vertebrates[133], resulting in 8,676 predicted high-confidence tRNA genes, excluding ~12,000 potential pseudogenes and low-scoring repetitive elements[7]. These 8,676 only produce 3,180 unique tRNA transcripts. An astounding 955 of these represent duplicated sequences (**Figure 6A**).

Although collapsing identical tRNAs has become a popular strategy[10,11,14,110,130], it is often still not enough to reduce multimapping or misalignment between highly similar tRNA sequences. In human, a tRNA-Ser-IGA isodecoder (containing an inosine at position 34; I34) and a tRNA-Ser-UGA isodecoder only differ by one nucleotide at position 34 in the anticodon (**Figure 6B**). Even though these tRNAs decode different codons, their sequence is similar enough to easily allow reads from one to be misaligned to the other, especially because mismatch tolerance offered by short read aligners is often position-agnostic. More lenient

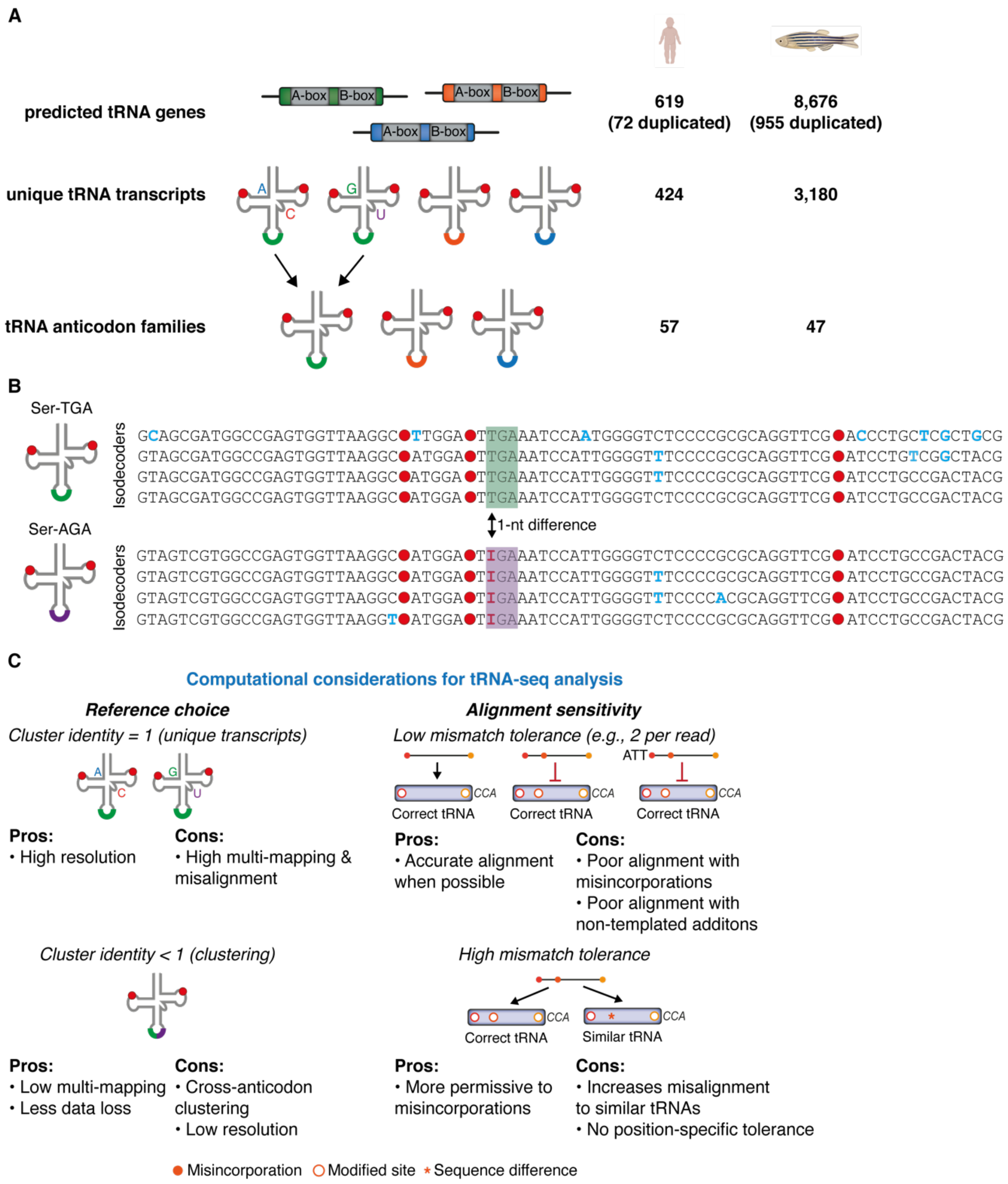

A
predicted tRNA genes
A-box
B-box
619
(72 duplicated)
8,676
(955 duplicated)
unique tRNA transcripts
A
C
G
U
424
3,180
tRNA anticodon families
57
47
B
Ser-TGA
Isodecoders
GCAGCGATGGCCGAGTGGTTAAGGC●TTGGA●TTGAAATCCAATGGGGTCTCCCCGCGCAGGTTCG●ACCCTGCTCGCTGCG
GTAGCGATGGCCGAGTGGTTAAGGC●ATGGA●TTGAAATCCATTGGGGTTTCCCCGCGCAGGTTCG●ATCCTGTCGGCTACG
GTAGCGATGGCCGAGTGGTTAAGGC●ATGGA●TTGAAATCCATTGGGGTTTCCCCGCGCAGGTTCG●ATCCTGCCGACTACG
GTAGCGATGGCCGAGTGGTTAAGGC●ATGGA●TTGAAATCCATTGGGGTCTCCCCGCGCAGGTTCG●ATCCTGCCGACTACG
1-nt difference
Ser-AGA
Isodecoders
GTAGTCGTGGCCGAGTGGTTAAGGC●ATGGA●TIGAAATCCATTGGGGTCTCCCCGCGCAGGTTCG●ATCCTGCCGACTACG
GTAGTCGTGGCCGAGTGGTTAAGGC●ATGGA●TIGAAATCCATTGGGGTTTCCCCGCGCAGGTTCG●ATCCTGCCGACTACG
GTAGTCGTGGCCGAGTGGTTAAGGC●ATGGA●TIGAAATCCATTGGGGTTTCCCCACGCAGGTTCG●ATCCTGCCGACTACG
GTAGTCGTGGCCGAGTGGTTAAGGT●ATGGA●TIGAAATCCATTGGGGTCTCCCCGCGCAGGTTCG●ATCCTGCCGACTACG
C
Computational considerations for tRNA-seq analysis
Reference choice
Cluster identity = 1 (unique transcripts)
A
C
G
U
Pros:
• High resolution
Cons:
• High multi-mapping & misalignment
Cluster identity < 1 (clustering)
Pros:
• Low multi-mapping
• Less data loss
Cons:
• Cross-anticodon clustering
• Low resolution
Alignment sensitivity
Low mismatch tolerance (e.g., 2 per read)
ATT
CCA
Correct tRNA
Correct tRNA
Correct tRNA
Pros:
• Accurate alignment when possible
Cons:
• Poor alignment with misincorporations
• Poor alignment with non-templated additons
High mismatch tolerance
CCA
Correct tRNA
Similar tRNA
Pros:
• More permissive to misincorporations
Cons:
• Increases misalignment to similar tRNAs
• No position-specific tolerance
● Misincorporation ○ Modified site * Sequence difference

Figure 6 Eukaryotic tRNA gene organization and similarity, and implications for computational analysis of tRNA-seq data.

(A) Predicted tRNA gene sets in *Homo sapiens* and *Danio rerio* and their organization into unique transcript-producing loci, and anticodon families. Numbers given in parentheses are genes which produce duplicate tRNA transcripts.

(B) Sequence similarity between the four isodecoders of Ser-TGA and Ser-AGA in *Homo sapiens.* Only one nucleotide difference exists between two of the isodecoders from different anticodon families which can cause misalignment and incorrect abundance estimation. Orange dots: misincorporations; Blue text: mismatches to other isodecoders.

(C) Two main considerations for computational analysis of tRNA-seq data include reference choice (left) pertaining to clustering parameters of tRNA genes, and alignment sensitivity (right) with regards to handling of mismatches in tRNA reads.

clustering, with lower sequence identity thresholds perform much better, but suffer from clustering tRNAs with different anticodons together (for instance, the Ser-IGA and Ser-UGA sequences discussed above), requiring complex analysis of cross-mapping for predictions of anticodon-level quantitation[129,130]. So far, no method provides efficient clustering of tRNA genes with the ability to restore transcript-level resolution from cluster-aligned sequencing data. Such a method would combine reduced alignment ambiguity with transcript-level resolution of tRNA abundance.

Misincorporation-sensitive alignment

Given the promise of RTs such as TGIRT to read through and misincorporate at Watson-Crick face modified residues, it is surprising that no method so far attempts to account for these position-specific mismatches that ultimately hinder alignment. Most popular short-read aligners, such as Bowtie[134], Bowtie 2[135], and even newer, more robust algorithms such as STAR[136], often only allow a specific number of mismatches per read independent of their position. Furthermore, this mismatch tolerance is designed to account for low-frequency errors introduced during cDNA synthesis, PCR, or sequencing, and natural heterogeneity in genomic sequence among individuals. This tolerance is therefore unsuited for the additional and prevalent level of misincorporations introduced at modified sites (**Figure 6C**).

Since misincorporations and true sequence diversity between tRNAs are both treated indiscriminately as mismatches during alignment, simply increasing mismatch tolerance to account for misincorporations is also not suitable. In this scenario, mismatches between different tRNAs that should not be tolerated might be allowed, increasing misalignment. This problem could be further compounded in the scenario of clustered tRNA genes where members of a cluster are represented by a parent or as a consensus sequence (**Figure 6C**)[129,130]. During alignment to this representative sequence, additional mismatches are expected where members of the cluster differ. These should also be specifically tolerated, in combination with misincorporations, but should again not lead to misalignment due to high mismatch tolerance outside of these sites.

Lastly, RTs are well-known to introduce non-templated nucleotide additions at read ends[119–121]. During alignment, these should also be considered and appropriately handled. In particular, these might be soft-clipped during alignment, such that they do not count towards the total mismatch tolerance or interfere with modification analysis.

Accounting for these sources of variation separately from misincorporations at modified sites is crucial in retaining qualitative and quantitative information about modifications for their analysis downstream, and to permit correct handling of alignments to clustered references. It would also allow separate control of mismatch tolerance to minimize multimapping and misalignment. However, no method to date attempts to regulate alignment at misincorporation sites by implementing novel alignment algorithms, or repurposing one better suited to dynamically control mismatch tolerance. Furthermore, this has certainly not been considered within the context of clustered references and the challenges associated with the resulting alignment information.

Outline of this thesis

Global analysis and quantitation of eukaryotic tRNA pools by sequencing remains a challenge, both biochemically during library construction, and with regards to computational methods and tools for the analysis of the resulting datasets. Progress in such methods, however, is crucial for understanding multiple facets of tRNA biology that have so far been intractable and hindered by inaccurate and biased methodology. Open questions, whose investigation can be aided by improved quantitation methods, include; evaluating the extent of tRNA transcript, anticodon pool, and modification regulation between different cell types, tissues, or disease contexts; understanding the functional and biological relevance of regulation at each of these levels; investigating how such changes globally impact translation and proteostasis; clarifying the mechanisms underlying tRNA-related pathologies, and how treatment and therapy can be targeted and tailored for the best outcome.

This thesis focuses specifically on addressing outstanding hurdles to tRNA-seq methodology through the development of novel methods targeted specifically at overcoming these issues. Furthermore, we aim to provide useful, open-source resources for such workflows in the form of a user-friendly computational tool with documentation, a detailed step-by-step protocol, troubleshooting guide and in-depth updates and enhancements to the computational package, and active community support. With respect to computational packages for tRNA-seq, such resource availability, transparency, and community engagement has not been implemented so far, but offers opportunities for enhancements, custom functionality, and problem solving through interaction with users and other experts in the field.

Chapter 2 focusses on the development of the modification-induced misincorporation tRNA sequencing (mim-tRNAseq) method. Introduced here are crucial optimizations to many steps in library generation, most importantly the reverse transcription reaction with TGIRT. An accompanying computational package is presented, which combines features of tRNA clustering, misincorporation-sensitive alignment, and a novel deconvolution algorithm able to restore transcript-level resolution to cluster-aligned reads. Comparisons to other available methods provide evidence supporting the improved accuracy and sensitivity of mim-tRNAseq, while metrics for alignment and coverage show superior modification readthrough and alignment sensitivity. Investigations of differential transcript abundance highlight important tRNA regulation in multicellular eukaryotes. Furthermore, functionality for detailed modification analysis is presented, wherein quantitative rigor is shown using yeast modification-deficient strains. Strikingly, these analyses highlight the interdependence of modifications at distinct sites, hinting at the complexity of modification pathways for tRNA transcripts.

Chapter 3 details a step-by-step protocol for the use of mim-tRNAseq, from RNA isolation, to data analysis. Furthermore, expected analysis outcomes are covered, which describe various data outputs and visualizations and how to utilize these for optimization and quality control. Since the initial release of mim-tRNAseq key updates to cluster deconvolution have been implemented and are described here. Lastly, typical problems and errors with their matching solutions are described to further improve accessibility of the method.

Chapter 4 summarizes the work in this thesis and provides a general discussion on the context, impact, and future prospects of the work.

References

1. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science (80-.).* **147**, 1462–1465 (1965).
2. Motorin, Y. & Helm, M. tRNA stabilization by modified nucleotides. *Biochemistry* **49**, 4934–4944 (2010).
3. Agris, P. F. *et al.* Celebrating wobble decoding: Half a century and still much is new. *RNA Biol.* **15**, 537–553 (2018).
4. Grosjean, H., Söll, D. G. & Crothers, D. M. Studies of the complex between transfer RNAs with complementary anticodons: I. Origins of enhanced affinity between complementary triplets. *J. Mol. Biol.* **103**, 499–519 (1976).
5. Rubio, M. A. T. *et al.* An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7821–7826 (2007).
6. Torres, A. G. *et al.* Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.* **43**, 5145–5157 (2015).
7. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
8. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-Specific Differences in Human Transfer RNA Expression. *PLoS Genet.* **2**, e221 (2006).
9. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
10. Zheng, G. *et al.* Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* **12**, 835–837 (2015).
11. Pinkard, O., McFarland, S., Sweet, T. & Coller, J. Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nat. Commun.* **11**, 1–15 (2020).
12. Gogakos, T. *et al.* Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell Rep.* **20**, 1463–1475 (2017).
13. Shigematsu, M. *et al.* YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Res.* gkx005 (2017) doi:10.1093/nar/gkx005.
14. Cozen, A. E. *et al.* ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods* **12**, 879–884 (2015).
15. Gingold, H. *et al.* A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell* **158**, 1281–1292 (2014).
16. Schmitt, B. M. *et al.* High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Res.* **24**, 1797–1807 (2014).
17. Oler, A. J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.* **17**, 620–628 (2010).
18. Orellana, E. A., Siegal, E. & Gregory, R. I. tRNA dysregulation and disease. *Nat. Rev. Genet. 2022* 1–14 (2022) doi:10.1038/s41576-022-00501-9.
19. Blaze, J. & Akbarian, S. The tRNA regulome in neurodevelopmental and neuropsychiatric disease. *Mol. Psychiatry 2022* 1–10 (2022) doi:10.1038/s41380-022-01585-9.
20. Suzuki, T., Nagao, A. & Suzuki, T. Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases. *http://dx.doi.org/10.1146/annurev-genet-110410-132531* **45**, 299–329 (2011).
21. Suzuki, T. The expanding world of tRNA modifications and their disease relevance. *Nat. Rev. Mol. Cell Biol.* 1–18 (2021) doi:10.1038/s41580-021-00342-0.
22. Dolgin, E. tRNA therapeutics burst onto startup scene. *Nat. Biotechnol.* **40**, 283–286 (2022).
23. Porter, J. J., Heil, C. S. & Lueck, J. D. Therapeutic promise of engineered nonsense suppressor tRNAs. *Wiley Interdiscip. Rev. RNA* **12**, e1641 (2021).
24. Sprinzl, M., Horn, C., Brown, M., Loudovltch, A. & Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26**, 148–153 (1998).
25. Crothers, D. M., Seno, T. & Söll, G. Is there a discriminator site in transfer RNA? *Proc. Natl. Acad. Sci. U. S. A.* **69**, 3063–3067 (1972).
26. Dieci, G., Conti, A., Pagano, A. & Carnevali, D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 296–305 (2013).
27. Pieler, T., Appel, B., Oei, S. L., Mentzel, H. & Erdmann, V. A. Point mutational analysis of the Xenopus laevis 5S gene promoter. *EMBO J.* **4**, 1847–1853 (1985).
28. Galli, G., Hofstetter, H. & Birnstiel, M. L. Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature* **294**, 626–631 (1981).

29. Hofstetter, H., Kressmann, A. & Birnstiel, M. L. A split promoter for a eucaryotic tRNA gene. *Cell* **24**, 573–585 (1981).
30. Sadowski, C. L., William Henry, R., Lobo, S. M. & Hernandez, N. Targeting TBP to a non-TATA box cis-regulatory element: a TBP-containing complex activates transcription from snRNA promoters through the PSE. *Genes Dev.* **7**, 1535–1548 (1993).
31. Waldschmidt, R., Wanandi, I. & Seifart, K. H. Identification of transcription factors required for the expression of mammalian U6 genes in vitro. *EMBO J.* **10**, 2595–2603 (1991).
32. Baker, R. E., Camier, S., Sentenac, A. & Hall, B. D. Gene size differentially affects the binding of yeast transcription factor tau to two intragenic regions. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 8768 (1987).
33. Dieci, G., Bosio, M. C., Fermi, B. & Ferrari, R. Transcription reinitiation by RNA polymerase III. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 331–341 (2013).
34. Arimbasseri, A. G. & Maraia, R. J. Mechanism of Transcription Termination by RNA Polymerase III Utilizes a Non-template Strand Sequence-Specific Signal Element. *Mol. Cell* **58**, 1124–1132 (2015).
35. Braglia, P., Percudani, R. & Dieci, G. Sequence context effects on oligo(dT) termination signal recognition by Saccharomyces cerevisiae RNA polymerase III. *J. Biol. Chem.* **280**, 19551–19562 (2005).
36. Graczyk, D., Cieśla, M. & Boguta, M. Regulation of tRNA synthesis by the general transcription factors of RNA polymerase III - TFIIIB and TFIIIC, and by the MAF1 protein. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1861**, 320–329 (2018).
37. White, R. J. RNA polymerases I and III, non-coding RNAs and cancer. *Trends Genet.* **24**, 622–629 (2008).
38. Czech, A., Wende, S., Mörl, M., Pan, T. & Ignatova, Z. Reversible and rapid transfer-RNA deactivation as a mechanism of translational repression in stress. *PLoS Genet.* **9**, e1003767 (2013).
39. Björk, G. R., Wikström, P. M. & Byström, A. S. Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science (80-.).* **244**, 986–989 (1989).
40. Wu, X. qi & Gross, H. J. The long extra arms of human tRNA((Ser)Sec) and tRNA(Ser) function as major identify elements for serylation in an orientation-dependent, but not sequence-specific manner. *Nucleic Acids Res.* **21**, 5589–5594 (1993).
41. Giegé, R., Sissler, M. & Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.* **26**, 5017–5035 (1998).
42. Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17224–17229 (2014).
43. Morselli, M. & Dieci, G. Epigenetic regulation of human non-coding RNA gene transcription. *Biochem. Soc. Trans.* **50**, 723–736 (2022).
44. Helbo, A. S., Lay, F. D., Jones, P. A., Liang, G. & Grønbæk, K. Nucleosome Positioning and NDR Structure at RNA Polymerase III Promoters. *Sci. Reports 2017 71* **7**, 1–11 (2017).
45. Van Bortle, K., Phanstiel, D. H. & Snyder, M. P. Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation. *Genome Biol. 2017 181* **18**, 1–18 (2017).
46. Oler, A. J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.* **17**, 620–628 (2010).
47. Barski, A. *et al.* Pol II and its associated epigenetic marks are present at Pol III–transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol. 2010 175* **17**, 629–634 (2010).
48. Canella, D. *et al.* A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. *Genome Res.* **22**, 666–680 (2012).
49. Zhang, G., Lukoszek, R., Mueller-Roeber, B. & Ignatova, Z. Different sequence signatures in the upstream regions of plant and animal tRNA genes shape distinct modes of regulation. *Nucleic Acids Res.* **39**, 3331–3339 (2011).
50. Trivedi, A., Young, L. S., Ouyang, C., Johnson, D. L. & Sprague, K. U. A TATA Element Is Required for tRNA Promoter Activity and Confers TATA-binding Protein Responsiveness in DrosophilaSchneider-2 Cells. *J. Biol. Chem.* **274**, 11369–11375 (1999).
51. Huang, Y. & Maraia, R. J. Comparison of the RNA polymerase III transcription machinery in Schizosaccharomyces pombe, Saccharomyces cerevisiae and human. *Nucleic Acids Res.* **29**, 2675 (2001).
52. Hamada, M., Huang, Y., Lowe, T. M. & Maraia, R. J. Widespread use of TATA elements in the core promoters for RNA polymerases III, II, and I in fission yeast. *Mol. Cell. Biol.* **21**, 6870–6881 (2001).

53. Turowski, T. W. & Tollervey, D. Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochem. Soc. Trans.* **44**, 1367 (2016).
54. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* 1–13 (2021) doi:10.1038/s41588-021-00782-6.
55. Pluta, K. *et al.* Maf1p, a Negative Effector of RNA Polymerase III in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **21**, 5031 (2001).
56. Reina, J. H., Azzouz, T. N. & Hernandez, N. Maf1, a New Player in the Regulation of Human RNA Polymerase III Transcription. *PLoS One* **1**, e134 (2006).
57. Michels, A. A. *et al.* mTORC1 directly phosphorylates and regulates human MAF1. *Mol. Cell. Biol.* **30**, 3749–3757 (2010).
58. Girbig, M. *et al.* Cryo-EM structures of human RNA polymerase III in its unbound and transcribing states. *Nat. Struct. Mol. Biol.* **28**, 210–219 (2021).
59. Vorländer, M. K. *et al.* Structural basis for RNA polymerase III transcription repression by Maf1. *Nat. Struct. Mol. Biol. 2020 273* **27**, 229–232 (2020).
60. Wang, Q. *et al.* Structural insights into transcriptional regulation of human RNA polymerase III. *Nat. Struct. Mol. Biol.* **28**, 220–227 (2021).
61. Turowski, T. W. *et al.* Global analysis of transcriptionally engaged yeast RNA polymerase III reveals extended tRNA transcripts. *Genome Res.* **26**, 933–944 (2016).
62. Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Publ. Gr.* **16**, 98–112 (2014).
63. Ishimura, R. *et al.* Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science (80-.).* **345**, 455–459 (2014).
64. Kapur, M. *et al.* Expression of the Neuronal tRNA n-Tr20 Regulates Synaptic Transmission and Seizure Susceptibility. *Neuron* **108**, 193 (2020).
65. Kirchner, S. *et al.* Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLoS Biol.* **15**, (2017).
66. Patil, A. *et al.* Translational infidelity-induced protein stress results from a deficiency in Trm9-catalyzed tRNA modifications. *RNA Biol.* **9**, 990 (2012).
67. Zhang, Z. *et al.* Global analysis of tRNA and translation factor expression reveals a dynamic landscape of translational regulation in human cancers. *Commun. Biol. 2018 11* **1**, 1–11 (2018).
68. Goodarzi, H. *et al.* Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell* **165**, 1416–1427 (2016).
69. Thornlow, B. P. *et al.* Transfer RNA genes experience exceptionally elevated mutation rates. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8996–9001 (2018).
70. Doran, J. L., Bingle, W. H. & Roy, K. L. Two human genes encoding tRNA(GCCGly). *Gene* **65**, 329–336 (1988).
71. Urbonavičius, J., Qian, Q., Durand, J. M. B., Hagervall, T. G. & Björk, G. R. Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J.* **20**, 4863–4873 (2001).
72. Asano, K. *et al.* Metabolic and chemical regulation of tRNA modification associated with taurine deficiency and human disease. *Nucleic Acids Res.* **46**, 1565 (2018).
73. Nedialkova, D. D. & Leidel, S. A. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell* **161**, 1606–1618 (2015).
74. Huang, B., Johansson, M. J. O. & Byström, A. S. An early step in wobble uridine tRNA modification requires the Elongator complex. *RNA* **11**, 424–436 (2005).
75. Johansson, M. J. O., Xu, F. & Byström, A. S. Elongator—a tRNA modifying complex that promotes efficient translational decoding. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1861**, 401–408 (2018).
76. Karlsborn, T., Tükenmez, H., Chen, C. & Byström, A. S. Familial dysautonomia (FD) patients have reduced levels of the modified wobble nucleoside mcm5s2U in tRNA. *Biochem. Biophys. Res. Commun.* **454**, 441–445 (2014).
77. Grunberger, D., Weinstein, I. B. & Mushinski, J. F. Deficiency of the Y base in a hepatoma phenylalanine tRNA. *Nature* **253**, 66–67 (1975).
78. Kuchino, Y., Borek2, E., Grunberger3, D., Mushinski4, J. F. & Nishimural, S. Changes of post-tanscriptional modification of wye base in tumor-specific tRNAPhe. *Nucleic Acids Res.* **10**, (1982).
79. Rosselló-Tortella, M. *et al.* Epigenetic loss of the transfer RNA-modifying enzyme TYW2 induces ribosome frameshifts in colon cancer. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20785–20793 (2020).
80. Waas, W. F., Druzina, Z., Hanan, M. & Schimmel, P. Role of a tRNA Base Modification and Its Precursors in Frameshifting in Eukaryotes *. *J. Biol. Chem.* **282**, 26026–26034 (2007).

81. Gerber, A. P. & Keller, W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science (80-.).* **286**, 1146–1149 (1999).
82. Richter, U., McFarland, R., Taylor, R. W. & Pickett, S. J. The molecular pathology of pathogenic mitochondrial tRNA variants. *FEBS Letters* vol. 595 1003–1024 (2021).
83. Tsutomu, S., Asuteka, N. & Takeo, S. Human mitochondrial diseases caused by lack of taurine modification in mitochondrial tRNAs. *Wiley Interdiscip. Rev. RNA* **2**, 376–386 (2011).
84. Yasukawa, T., Suzuki, T., Ishii, N., Ohta, S. & Watanabe, K. Wobble modification defect in tRNA disturbs codon-anticodon interaction in a mitochondrial disease. *EMBO J.* **20**, 4794–4802 (2001).
85. Morscher, R. J. *et al.* Mitochondrial translation requires folate-dependent tRNA methylation. *Nat. Publ. Gr.* (2018) doi:10.1038/nature25460.
86. Dunbar, D. R., Moonie, P. A., Zeviani, M. & Holt, I. J. Complex I deficiency is Associated with 3243G:C Mitochondrial DNA in Osteosarcoma Cell Cybrids. *Hum. Mol. Genet.* **5**, 123–129 (1996).
87. Goto, Y. I., Nonaka, I. & Horai, S. A mutation in the tRNALeu(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nat. 1990 3486302* **348**, 651–653 (1990).
88. Rogalski, M., Karcher, D. & Bock, R. Superwobbling facilitates translation with reduced tRNA sets. *Nat. Struct. Mol. Biol.* **15**, 192–198 (2008).
89. Antonio Enriquez, J., Chomyn, A. & Attardi, G. MtDNA mutation in MERRF syndrome causes defective aminoacylation of tRNALys and premature translation termination. (1995).
90. Richter, U. *et al.* RNA modification landscape of the human mitochondrial tRNALys regulates protein synthesis. *Nat. Commun.* **9**, (2018).
91. Wiener, D. & Schwartz, S. The epitranscriptome beyond m6A. *Nat. Rev. Genet. 2020 222* **22**, 119–131 (2020).
92. Schaefer, M. R. The Regulation of RNA Modification Systems: The Next Frontier in Epitranscriptomics? *Genes 2021, Vol. 12, Page 345* **12**, 345 (2021).
93. Motorin, Y. & Helm, M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. *Genes 2019, Vol. 10, Page 35* **10**, 35 (2019).
94. Helm, M. & Motorin, Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.* **18**, 275–291 (2017).
95. Torres, A. G., Batlle, E. & Ribas de Pouplana, L. Role of tRNA modifications in human diseases. *Trends Mol. Med.* **20**, 306–314 (2014).
96. Ko, W., Porter, J. J., Sipple, M. T., Edwards, K. M. & Lueck, J. D. Efficient suppression of endogenous CFTR nonsense mutations using anticodon-engineered transfer RNAs. *Mol. Ther. - Nucleic Acids* **28**, 685–701 (2022).
97. Lueck, J. D. *et al.* Engineered transfer RNAs for suppression of premature termination codons. *Nat. Commun. 2019 101* **10**, 1–11 (2019).
98. Temple, G. F., Dozy, A. M., Roy, K. L. & Wai Kan, Y. Construction of a functional human suppressor tRNA gene: an approach to gene therapy for β-thalassaemia. *Nat. 1982 2965857* **296**, 537–540 (1982).
99. Bordeira-Carriço, R. *et al.* Rescue of wild-type E-cadherin expression from nonsense-mutated cancer cells by a suppressor-tRNA. **22**, 1085–1092 (2014).
100. Eggertsson, G. & Söll, D. Transfer ribonucleic acid-mediated suppression of termination codons in Escherichia coli. *Microbiol. Rev.* **52**, 354 (1988).
101. Mort, M., Ivanov, D., Cooper, D. N. & Chuzhanova, N. A. A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat.* **29**, 1037–1047 (2008).
102. Shi, X. *et al.* Missense mutation of the sodium channel gene SCN2A causes Dravet syndrome. *Brain Dev.* **31**, 758–762 (2009).
103. Wang, J. *et al.* AAV-delivered suppressor tRNA overcomes a nonsense mutation in mice. *Nat. 2022 6047905* **604**, 343–348 (2022).
104. Zuko, A. *et al.* tRNA overexpression rescues peripheral neuropathy caused by mutations in tRNA synthetase. *Science* **373**, 1161–1166 (2021).
105. Dittmar, K. A., Mobley, E. M., Radek, A. J. & Pan, T. Exploring the Regulation of tRNA Distribution on the Genomic Scale. *J. Mol. Biol.* **337**, 31–47 (2004).
106. Trewick, S. C., Henshaw, T. F., Hausinger, R. P., Lindahl, T. & Sedgwick, B. Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. *Nat. 2002 4196903* **419**, 174–178 (2002).
107. Falnes, P., Johansen, R. F. & Seeberg, E. AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli. *Nat. 2002 4196903* **419**, 178–182 (2002).

108. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkx1030.

109. Dai, Q., Zheng, G., Schwartz, M. H., Clark, W. C. & Pan, T. Selective Enzymatic Demethylation of N2,N2-Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. *Angew. Chemie Int. Ed.* **56**, 5017–5020 (2017).

110. Clark, W. C., Evans, M. E., Dominissini, D., Zheng, G. & Pan, T. tRNA base methylation identification and quantification via high-throughput sequencing. *RNA* **22**, 1771–1784 (2016).

111. Karaca, E. *et al.* Human CLP1 mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. *Cell* **157**, 636–650 (2014).

112. Arimbasseri, A. G. *et al.* RNA Polymerase III Output Is Functionally Linked to tRNA Dimethyl-G26 Modification. *PLoS Genet.* **11**, e1005671 (2015).

113. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).

114. Katibah, G. E. *et al.* Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12025–12030 (2014).

115. Qin, Y. *et al.* High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22**, 111–128 (2016).

116. Lambowitz, A. M. & Zimmerly, S. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003616 (2011).

117. Schwartz, S. & Motorin, Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA Biol.* **14**, 1124–1137 (2017).

118. Sas-Chen, A. & Schwartz, S. Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods* (2018) doi:10.1016/j.ymeth.2018.10.011.

119. Golinelli, M. P. & Hughes, S. H. Nontemplated nucleotide addition by HIV-1 reverse transcriptase. *Biochemistry* **41**, 5894–5906 (2002).

120. Jaju, M., Beard, W. A. & Wilson, S. H. Human immunodeficiency virus type 1 reverse transcriptase. 3'- azidodeoxythymidine 5'-triphosphate inhibition indicates two-step binding for template-primer. *J. Biol. Chem.* **270**, 9740–9747 (1995).

121. Chen, D. & Patton, J. T. Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. *Biotechniques* **30**, 574–582 (2001).

122. Li, X. *et al.* Base-Resolution Mapping Reveals Distinct m 1 A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. *Mol. Cell* **68**, 993-1005.e9 (2017).

123. Schwartz, S. m1A within cytoplasmic mRNAs at single nucleotide resolution: a reconciled transcriptome-wide map. *RNA* **24**, 1427–1436 (2018).

124. Menéndez-Arias, L. Mutation Rates and Intrinsic Fidelity of Retroviral Reverse Transcriptases. *Viruses* **1**, 1137 (2009).

125. Svarovskaia, E. S., Cheslock, S. R., Zhang, W. H., Hu, W. S. & Pathak, V. K. Retroviral mutation rates and reverse transcriptase fidelity. *Front. Biosci.* **8**, 117–134 (2003).

126. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 1–15 (2016).

127. Safra, M. *et al.* The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* **551**, 251–255 (2017).

128. Legrand, C. *et al.* Statistically robust methylation calling for wholetranscriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. *Genome Res.* **27**, 1589–1596 (2017).

129. Ryvkin, P. *et al.* HAMR: high-throughput annotation of modified ribonucleotides. *RNA* **19**, 1684–1692 (2013).

130. Hoffmann, A. *et al.* Accurate Mapping of tRNA Reads. *Bioinformatics* **366**, 1 (2017).

131. Pichot, F., Marchand, V., Helm, M. & Motorin, Y. Non-Redundant tRNA Reference Sequences for Deep Sequencing Analysis of tRNA Abundance and Epitranscriptomic RNA Modifications. *Genes 2021, Vol. 12, Page 81* **12**, 81 (2021).

132. Marck, C. & Grosjean, H. tRNomics: Analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **8**, 1189–1232 (2002).

133. Meyer, A. & Schartl, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999).

134. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 1–10 (2009) doi:gb-2009-10-3-r25 [pii]\r10.1186/gb-2009-10-3-r25.

135. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
136. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
137. Berg, M. D. & Brandl, C. J. Transfer RNAs: diversity in form and function. *RNA Biology* vol. 18 316–339 (2021).

CHAPTER 2

High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq

This chapter was published in Molecular Cell following peer review, and is available through open-access. The article was reformatted and inserted here.

Behrens, A., Rodschinka, G. & Nedialkova, D. D. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol. Cell* **81**, 1–14 (2021). https://doi.org/10.1016/j.molcel.2021.01.028

Overview

tRNA abundance and modification status is critical for regulating efficient mRNA decoding and ensuring efficient protein synthesis. Defects and perturbations to these properties have been linked to cancer invasion, neurodevelopmental disorders, and problems with cellular differentiation. Surprisingly, defects in tRNA biogenesis and regulation show heterogeneity in tissue vulnerability, implicating the composition of tRNA pools in diverse cell types in the maintenance of proteome integrity. However, investigations into tRNA regulation and modification status, and their biological significance, have been hindered by technical limitations to global quantitation methods. Extensive modifications on the Watson-Crick face of many tRNA residues and their secondary structure act as physical barriers to reverse transcription (RT), which limits the accuracy of quantitation by high-throughput sequencing methods, while gene duplication and sequence similarity complicates computational analysis.

To overcome these hurdles, we developed modification-induced misincorporation tRNA sequencing (mim-tRNAseq), which combines an optimized workflow for library generation from cellular tRNA, and a computational package for the analysis of the resulting data. The library generation protocol facilitates extensive modification readthrough at common Watson-Crick face modifications that pose as blocks to RT. This results in a majority of reads representing full-length tRNA transcripts, eliminating much of the coverage bias present in other methods.

The computational analysis pipeline introduces multiple novel algorithms for the accurate alignment of tRNA reads, including adjustable tRNA clustering, misincorporation-sensitive alignment, and read deconvolution that restores transcript-level resolution to quantitation and modification analysis. Moreover, the package allows easy customization of many parameters, allowing users to tailor the analysis to their organism of interest and their specific dataset, and facilitates the “one-click” analysis of tRNA coverage, abundance, charging fractions, and differential expression and modification status.

We show the efficacy of mim-tRNAseq in yeast, fly, and human cells and show the improvements of the method by extensive comparison to DM-tRNAseq, hydro-tRNAseq, and QuantM-tRNAseq. We demonstrate the accuracy and sensitivity of mim-tRNAseq to detect differences in tRNA transcript abundance, modification identity and stoichiometry, and differences in charging. Distinct misincorporation signatures and near-perfect linear regression in calibration curves of expected versus observed modification stoichiometry further bolster the efficacy of the modification detection and analysis pipeline of mim-tRNAseq.

Using mim-tRNAseq to investigate dynamics in eukaryotic tRNA pools we find:

i) a dramatic heterogeneity of tRNA pools among three human cell lines (K562, HEK293T, and hiPSC);

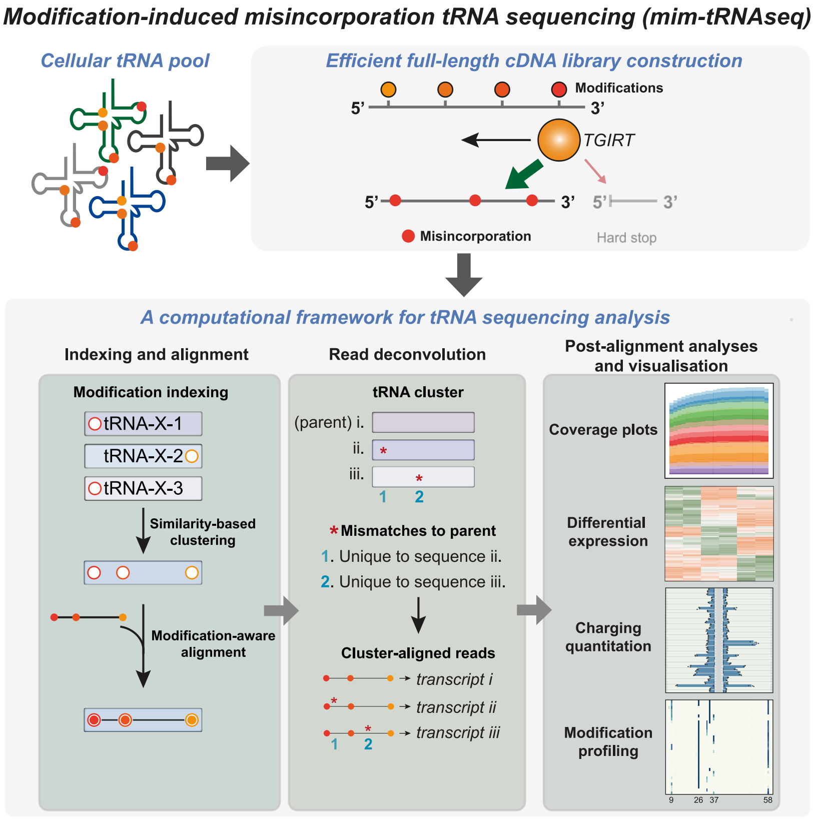
ii) large differences in modification stoichiometry among individual tRNAs from four eukaryotic species;

iii) a surprising interdependence of modifications at distinct sites within the same tRNA molecule.

Contribution: With advice and feedback from Danny Nedialkova, I conceptualised and implemented all computational analyses and workflows, including the design, programming, and maintenance of the mim-tRNAseq package. Data analysis and visualisation was performed jointly by myself and Danny Nedialkova, as was writing of the first draft of the manuscript. All subsequent reviews and edits to the manuscript were jointly shared between all three authors.

Abstract

Measurements of cellular tRNA abundance are hampered by pervasive blocks to cDNA synthesis at modified nucleosides and the extensive similarity among tRNA genes. We overcome these limitations with modification-induced misincorporation tRNA sequencing (mim-tRNAseq), which combines a workflow for full-length cDNA library construction from endogenously modified tRNA with a comprehensive and user-friendly computational analysis toolkit. Our method accurately captures tRNA abundance and modification status in yeast, fly, and human cells, and is applicable to any organism with a known genome. We applied mim-tRNAseq to discover a dramatic heterogeneity of tRNA isodecoder pools among diverse human cell lines and a surprising interdependence of modifications at distinct sites within the same tRNA transcript.



Introduction

Transfer RNAs (tRNAs) are short, abundant molecules required for translating genetic information into protein sequence. The composition of cellular tRNA pools is critical for efficient mRNA decoding and proteome integrity. tRNA expression is dynamically regulated in different tissues and during development[1–4], and defective tRNA biogenesis is linked to neurological disorders and cancer[5].

Nevertheless, the regulation of tRNA levels and its physiological significance remain under-appreciated due to a lack of accurate, high-resolution methods for tRNA quantitation. A major challenge is posed by the stable structure and pervasive Watson-Crick face modifications, which block reverse transcriptase (RT)[6]. Library generation workflows without a strategy for overcoming RT blocks yield mostly short reads due to premature RT stops at modified sites, as for instance in QuantM-tRNAseq[7]. Hybridization-based approaches can circumvent the need for cDNA synthesis, but they can only distinguish tRNAs differing by at least eight nucleotides[1]. This limitation is problematic given the extensive sequence similarity among tRNA transcripts, which can differ by a single nucleotide even if they read different codons[8]. Strategies to overcome structure- and modification-induced RT barriers have included tRNA fragmentation[9–11], the use of a thermostable template-switching RT in thermostable group II intron RTsequencing (TGIRT-seq and DM-tRNAseq)[12–14], and enzymatic removal of some base methylations in AlkB-facilitated RNA methylation sequencing (ARM-seq) and DM-tRNAseq[14,15].

While these methods have improved tRNA representation in sequencing libraries, several limitations remain. First, all of these methods relieve only a fraction of RT blocks, which can bias recovery towards tRNA subsets with few modified sites or those that are better substrates for demethylation *in vitro*. Second, removing modifications eliminates information about their presence and stoichiometry, which could be inferred from signatures of RT stops and misincorporations[6,12–14,16–21]. RNA modification profiling based solely on misincorporation signatures would be advantageous, as RT stops can also arise from RNA degradation or structure. Conditions that enable readthrough of Watson-Crick face modified sites while abrogating stops, however, have not been described for any RT so far[22]. A variant of the HIV-1 RT with improved readthrough of *N*^{1}-methyladenosine (m^{1}A) was recently derived by protein evolution[23], but whether it can also overcome any of the other types of RT-blocking tRNA modifications is unknown.

The computational analysis of tRNA sequencing data also presents significant challenges that are often overlooked. The number of predicted tRNA anticodon families in different genomes ranges from 33 in *M. hominis* to 57 in humans, with many tRNAs encoded by multiple gene copies. In eukaryotes, there is also considerable sequence variation among tRNAs with identical anticodons, which becomes more pronounced with increasing organismal complexity[24]. While the 41 tRNA anticodon families in budding yeast comprise 54 distinct tRNA transcripts, ~400 unique tRNA molecules can be potentially produced in human cells[8]. Some can have tissue-specific functions even in the presence of closely related isodecoders (tRNAs that share an anticodon but differ in sequence elsewhere)[2].

The exceptional degree of tRNA sequence similarity can undermine alignment accuracy, particularly for short reads resulting from premature RT stops[7] or tRNA fragmentation[9,10]. The problem is compounded by multiple mismatches between tRNA-derived reads and the genomic reference that arise from RT misincorporation during modification readthrough. Current alignment approaches allow mismatches at any position of a read[7,9,10,12–14,25], which can decrease mapping accuracy for nearly identical tRNAs. The total number of mismatches is also limited in some approaches, which can eliminate reads from highly modified tRNAs. Computational tool choice can thus substantially impact measurements of tRNA abundance and modification.

Here, we present a novel workflow that overcomes the experimental and computational hurdles to quantitative tRNA profiling through modification-induced misincorporation tRNA sequencing (mim-tRNAseq). We combine a sensitive method for cDNA library construction from endogenously modified tRNAs with a new computational framework for read alignment, data analysis and visualization. By identifying conditions that enable efficient RT readthrough of modified sites, we achieve uniform sequence coverage of tRNA pools from yeast, fly, and human cells while retaining modification signatures. In parallel, we developed a comprehensive and user-friendly computational toolkit, which yields measurements of tRNA abundance, charging fractions, and modification profiles with unprecedented accuracy and resolution. mim-tRNAseq identified a wide variation in tRNA isodecoder abundance among different human cell lines and an interdependence among tRNA modifications at distinct sites. As our workflow is sensitive, robust, and applicable to any organism with a known genome, we anticipate it will help shed new light on previously intractable aspects of tRNA biology.

Design

Efficient sequencing library generation from native eukaryotic tRNA pools

To develop a method for high-resolution tRNA quantitation, we focused on improving the efficiency of full-length cDNA synthesis from endogenously modified tRNAs by TGIRT. This enzyme can attach adapter sequences to RNA by template switching[26], which circumvents potential hindrances to 3’ adapter ligation and RT posed by tRNA structure[12–14]. TGIRT can also read through a subset of Watson-Crick face modifications more efficiently than other commercial RTs[18], albeit with reduced fidelity[12–14]. Despite these advantages, RT stops at modified sites in tRNA are still pervasive in TGIRT-mediated reactions[14,16] and cDNA yield is extremely low[14,27].

As TGIRT is active in a wide range of conditions[26], we asked whether its efficiency on tRNA templates can be further improved. To test this, we first purified tRNA pools from *S. cerevisiae* and human K562 cells by gel size selection of 60-100 nt RNAs from total RNA. We then used these, along with a synthetic unmodified *E. coli* tRNA-Lys-UUU, in template-switching TGIRT reactions. The cDNA yield from all templates was minimal under conditions previously used for tRNA sequencing (450 mM salt, 60°C)[12–14], but dramatically improved at lower temperatures and salt concentration (**Figure 1A**). While a considerable fraction of cDNAs we obtained were full-length, some RT stops still occurred, and larger products potentially derived from two tRNA molecules linked by template switching were also present (**Figure 1A**). To circumvent these issues and the known sequence bias of TGIRT during template switching[28], we introduced DNA adapters at the 3' end of tRNA with T4 RNA ligase

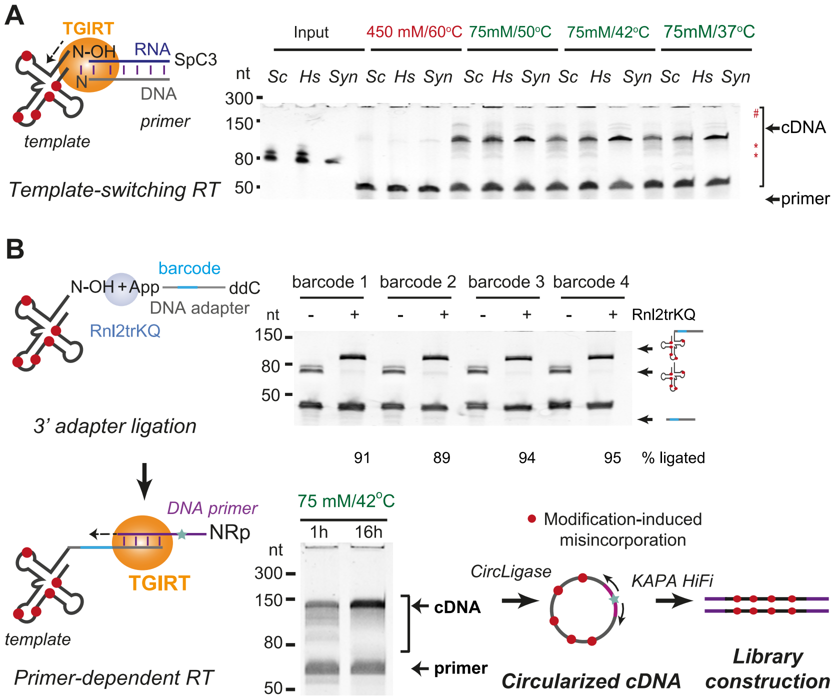


Figure 1 An optimized workflow for full-length cDNA library construction from eukaryotic tRNA pools.
(A) Schematic of template-switching TGIRT reactions primed by an RNA/DNA duplex with a single-nucleotide 3' overhang and a gel image of cDNA products from endogenously modified tRNA pools from S. cerevisiae (Sc), K562 cells (Hs) or a synthetic unmodified tRNA (Syn) at different reaction temperatures and salt concentration. Red: reaction conditions previously used for tRNA library construction; asterisks: premature stops to cDNA synthesis; hash: potential products from end-to-end linkage of tRNAs.
(B) Schematic of the mim-tRNAseq library generation workflow. Top gel image: 3' adapter ligation reactions with four barcoded adapters. Ligation efficiency was measured by normalizing input tRNA band intensity to that in reactions where Rnl2trKQ was omitted. Bottom gel image: comparison of cDNA yield in short (1 h) or extended (16 h) primer-dependent TGIRT RT on a mix of adapter-ligated tRNA pools from S. cerevisiae and human K562 and HEK293T cells. See also **Figure S1** and Methods.

2. We reasoned that the stable structure of mature tRNAs would not pose a challenge, as their 3’ ends contain the stretch of at least two unpaired nucleotides that is required for efficient 3’ adapter ligation[29]. To further minimize potential bias and enable sample pooling prior to RT, we designed four barcoded adapters with limited potential to co-fold with tRNA, and confirmed they can be ligated to size-selected yeast tRNA pools with 89% - 95% efficiency (**Figure 1B**). Pooled adapter-containing tRNA samples were then subjected to primer-dependent RT with TGIRT in a low-salt buffer at 42°C. Strikingly, we found that extending the reaction time eliminated nearly all premature RT stops on endogenously modified yeast and human tRNAs (**Figure 1B**) without compromising template integrity (**Figure S1A**). The primer for cDNA synthesis contained a 5’ RN dinucleotide to ensure efficient cDNA circularization[30,31] prior to PCR amplification with KAPA HiFi Polymerase, which exhibits minimal bias for fragment length or GC content[32]. This optimization enabled us to construct Illumina sequencing libraries starting from as little as 50 ng of endogenously modified tRNA with only five to six PCR cycles, minimizing sample input requirements and amplification bias.

A comprehensive computational framework for tRNA sequencing data analysis

We reasoned that the increase in full-length cDNA reads would reduce alignment ambiguity. However, given TGIRT’s low fidelity at modified sites, we expected many tRNA-derived reads to contain multiple mismatches to the reference genome. Another source of mismatches are non-templated nucleotides added to 3’ cDNA ends by TGIRT and other RTs[26,33]. Such read extensions are penalized by most algorithms, but can be recognized and dynamically processed (“soft-clipped”) by some. We therefore asked how two short-read aligners commonly used for tRNA analysis – Bowtie[34] and Bowtie 2[35] - would perform on a tRNA sequencing dataset from human HEK293T cells obtained with our improved library construction protocol (**Figure 1B**).

We first generated a non-redundant reference of 420 mature tRNA transcripts from 596 curated nuclear- and mitochondrial-encoded tRNA genes retrieved from GtRNAdb and mitotRNAdb[8,36] (**Figure 2A** and *Methods*). Alignment was performed with Bowtie or Bowtie 2 with parameters previously used for tRNA sequencing analysis[12–16]. Bowtie end-to-end alignment allows a maximum of three mismatches to the reference at any position. Its inability to distinguish modification-induced misincorporations from other mismatches can lead to data loss for highly modified tRNAs, or misalignment for highly similar tRNAs. Indeed, only 25% of reads from our HEK293T tRNA library aligned with Bowtie, with a third of those mapping to multiple tRNA references (**Figure 2A**). Trimming a fixed number of nucleotides from 5’ read ends prior to alignment, which can remove non-templated nucleotides, expectedly improved mapping rates (**Figure S1B**). The variable length of non-templated additions, however, makes such a trimming approach imprecise, and many trimmed reads still failed to align or were multi-mapped (**Figure S1B**).

In contrast, Bowtie 2's lack of mismatch restrictions and ability to soft-clip read ends make it seem more suited for tRNA read mapping. High mismatch tolerance, however, compounds the problem of misalignment: while Bowtie 2 increased alignment rates of our HEK293T-derived dataset to 82%, most mapped reads (85%) could not be assigned to a single reference (**Figure 2A**). Multi-mapping rates were similarly high when human QuantM-tRNAseq data were aligned using Bowtie 2 with the published settings[7] (85%, **Figure S1C**). These high rates of data loss indicate that standard read alignment approaches are poorly suited to the complexity of tRNA sequencing data, with consequences for the accuracy of all downstream analyses.

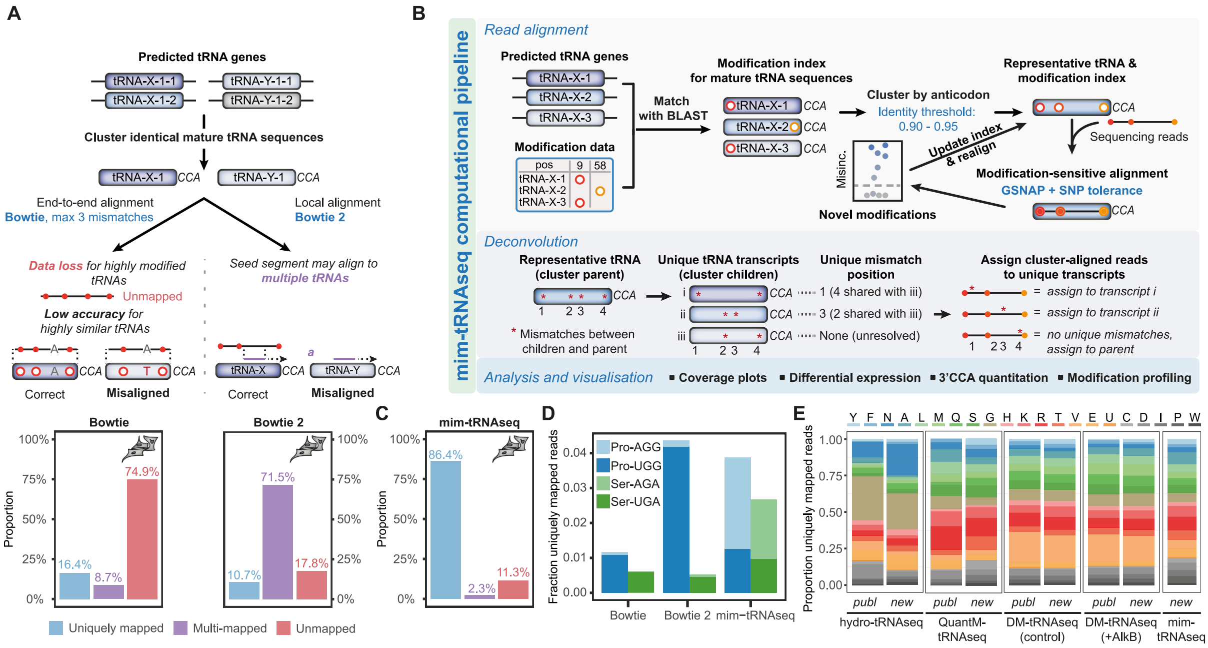


Figure 2 The mim-tRNAseq computational pipeline: a comprehensive framework for tRNA sequencing data analysis.

(A) Bowtie and Bowtie 2 alignment strategies and mapping statistics for a tRNA library from HEK293T cells constructed with the mim-tRNAseq workflow (n = 1).

(B) Outline of the mim-tRNAseq computational pipeline.

(C) Alignment statistics of HEK293T data (as in (A); n =1) using the mim-tRNAseq pipeline.

(D) Uniquely aligned read proportions for inosine 34 (I34) and uridine 34 (U34)-containing Ser and Pro tRNA isoacceptors using the three alignment strategies on a HEK293T dataset.

(E) Distribution of uniquely aligned reads among tRNA isotypes in published datasets and mim-tRNAseq from HEK293-derived cell lines (hydro-tRNAseq and QuantM-tRNAseq: HEK293 T-Rex Flp-IN; DM-tRNAseq control or AlkB-treated (+AlkB) and mim-tRNAseq library construction: HEK293T). Proportions were obtained from published counts per tRNA ("publ") or after re-analysis of the datasets with the mim-tRNAseq pipeline ("new"). tRNA families that carry the same amino acid (isotypes) are sorted by the number of RT barriers annotated in MODOMICS (decreasing from top to bottom; greyscale: isotypes without MODOMICS annotation). See also **Figure S1** and Methods.

Given these limitations, we reasoned that an accurate tRNA read analysis workflow requires solutions to two main challenges: alignment bias against reads with modification-induced misincorporations, and multi-mapping of reads from nearly identical tRNAs. To tackle the first issue, we took advantage of the comprehensive annotation of tRNA modifications in MODOMICS[37], and utilized this data to enable position-specific mismatch tolerance during alignment (**Figure 2B top panel**). To achieve this, we chose GSNAP, an aligner designed for detecting complex variants in sequencing reads[38]. Unlike most other algorithms, GSNAP considers alignments to a reference and an alternate allele equally in SNP-tolerant alignment mode while also effectively soft-clipping read ends. To address multi-mapping, we devised a strategy to cluster reference sequences by a sequence identity (ID) threshold. Given that many reads still map to multiple references with the commonly used strategy of clustering only completely identical tRNA genes[14,16,25] (ID=1, **Figure 2A**), we reasoned that alignment ambiguity could be decreased by lowering the sequence ID threshold. To maintain isoacceptor resolution, we chose to only cluster tRNA transcripts that share an anticodon regardless of sequence ID.

Based on these premises, we developed a new computational workflow to suit the intricacies of tRNA sequencing data (**Figure 2B** and *Methods*). To generate an alignment reference, mature tRNA transcript sequences are matched to MODOMICS to index all known modified sites, and clustered by anticodon according to sequence ID. Reads are aligned to the resulting indexed reference using GSNAP in SNP-tolerant mode. Unannotated potentially modified sites are detected by a mismatch rate of >10% and included in an updated index, followed by re-alignment of all reads with a more stringent tolerance to mismatches outside of modified sites to further boost alignment accuracy. To restore single-transcript resolution for subsequent analyses, we developed a deconvolution algorithm that assigns cluster-aligned reads to unique tRNA species (**Figure 2B middle panel** and *Methods*). For this, each cluster is assessed for single-nucleotide differences that distinguish unique tRNA sequences, based on which each read is separated from the cluster "parent" and assigned to an individual transcript. Analysis of coverage, 3' CCA, differential tRNA abundance, and modification profiling are then performed after read deconvolution (**Figure 2B bottom panel**). The entire computational framework for tRNA read alignment, analysis, and visualization is packaged in an open-source tool with a command-line interface and a broad set of customizable parameters.

This computational workflow dramatically improved both the efficiency and accuracy of tRNA read alignment. Both clustering and SNP tolerance at modified sites prevented data loss for defined tRNA subsets. A cluster ID of 0.95 maximized unique transcript resolution and minimized multi-mapping for human tRNAs (**Figure S1D**), yielding 86% uniquely mapped and only 2.5% ambiguously aligned reads (**Figure 2C**). Multi-mapping rates were five-fold higher when only completely identical tRNA transcripts were clustered, resulting in data loss for selected tRNAs (e.g. tRNA-Asn-GTT-2 and tRNA-Pro-AGG-1, **Figure S1D,E**). Aligning without SNP tolerance had similar effects, particularly for transcripts with inosine at position

34 (I34), which is encoded as an A but yields a G in cDNA libraries. The number of reads mapping to tRNA-Val-AAC, for example, increased by 300-fold in SNP-tolerant mode, and virtually all of these contained a G34 (**Figure S1F,G**). This high mismatch rate at I34 also presented obvious challenges for Bowtie and Bowtie 2. Almost no reads mapped to the I34-containing tRNA-Ser-AGA and tRNA-Pro-AGG with these algorithms, while many were assigned to tRNA-Ser-UGA and tRNA-Pro-UGG instead (**Figure 2D**). The same dramatic under-representation of tRNA-Ser-AGA and tRNA-Pro-AGG was evident in published counts for QuantM-tRNAseq libraries, which were generated by Bowtie 2 local alignment (**Figure S1H**). By contrast, our computational workflow yielded a more balanced representation of these four tRNA species for both mim-tRNAseq (**Figure 2D**) and QuantM-tRNAseq libraries (**Figure S1H**). The choice of read alignment parameters can thus yield very different tRNA abundance estimates

Results

The mim-tRNAseq workflow alleviates tRNA sequencing bias

To benchmark our workflow, we used mim-tRNAseq to analyze HEK293T tRNAs and compared our results to those published for the same cell type with DM-tRNAseq[14], and from the closely related HEK293 T-Rex Flp-IN line[39] obtained with hydro-tRNAseq[10] or QuantM-tRNAseq[7]. To distinguish experimental from computational differences, we also re-analyzed the published datasets using our computational pipeline (**Figure 2B)**. Reads from tRNA isotypes with a single known barrier to RT[37] were substantially overrepresented in DM-tRNAseq (tRNA-Val, 19%-21%) and hydro-tRNAseq (tRNA-Gly, 30%) compared to our dataset (~6%). In QuantM-tRNA-seq, tRNA-Arg comprised 16% of published tRNA counts versus 3.5% in hydro-tRNAseq, 7–9% in DM-tRNAseq, and 9% in our dataset. This isotype over-representation persisted regardless of analysis method (**Figure 2E**; “publ” vs “new”), suggesting it originated during library construction. By contrast, tRNA-Tyr, which has five known RT-blocking modifications, comprised ~4% of mapped reads in our dataset versus only 1% for published hydro-tRNAseq and DM-tRNAseq counts, and 0.3% for QuantM-tRNAseq. This under-representation was largely relieved when DM-tRNAseq and QuantM-tRNAseq datasets were re-analyzed with our computational pipeline (**Figure 2E)**. Thus, mim-tRNAseq recovers highly modified tRNAs more efficiently than current methods through a combination of advances in library construction and data analysis.

mim-tRNAseq improves tRNA coverage and abundance estimates

We extended our analysis to single-cell and multicellular eukaryotes by preparing mim-tRNAseq libraries from exponentially growing *S. cerevisiae* and *S. pombe,* as well as *D.*

melanogaster BG3-c2 cells and human induced pluripotent stem cells (hiPSC) with a normal karyotype. We determined the optimal cluster ID threshold as 0.90 for budding yeast and 0.95 for fission yeast, *Drosophila*, and human tRNA pools (**Figure S1D** and **S2A**). These settings yielded between 85% and 91% of uniquely mapped reads (**Figure 3A**), with a median of 65% - 83% full-length ones (**Figure 3B, C**). By contrast, unique alignment rates were lower for datasets from DM-tRNAseq, QuantM-tRNAseq, and for libraries we generated with the standard TGIRT protocol (**Figure S2B**). tRNA coverage in those datasets also had substantial 3’ end bias, consistent with RT stops at modified sites (**Figure S2, C-E)**. Accordingly, unique tRNA transcripts were represented by a median of < 11% and 6% full-length reads in DM-tRNAseq and QuantM-tRNAseq, respectively **(Figure S2, F-H)**.

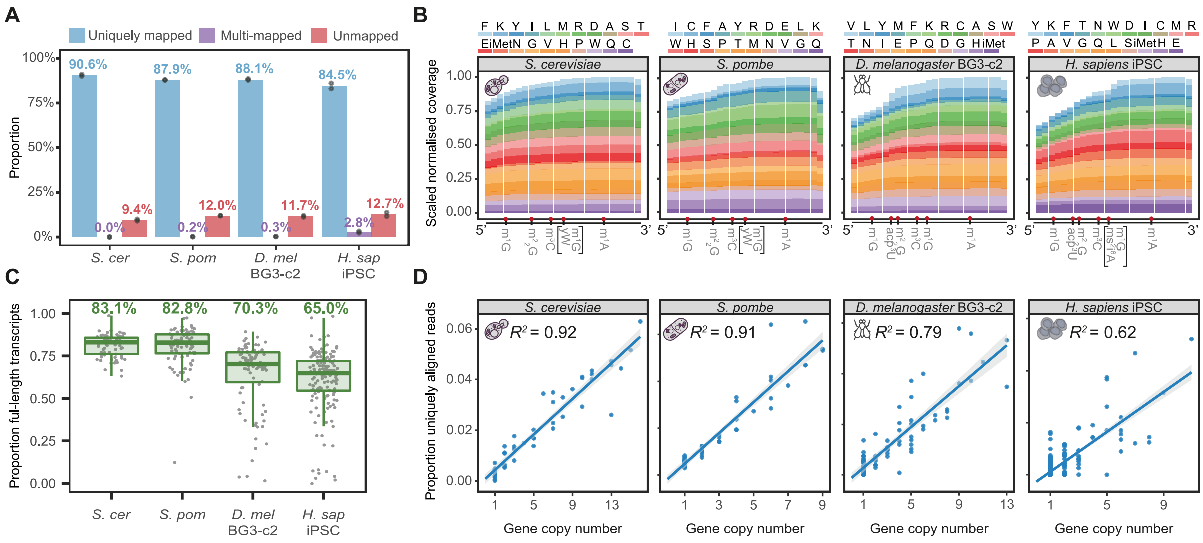


Figure 3 mim-tRNAseq improves quantitative analysis of tRNA pools in cells from diverse eukaryotes.
(A) Alignment statistics for mim-tRNAseq datasets from the indicated cell types. Bars and labels indicate average values, dots show individual sample values (n = 2).
(B) Metagene analysis of scaled sequence coverage across nuclear-encoded tRNA isotypes ordered per sample by differences between 3' and 5’ coverage (decreasing order from top to bottom; n = 1). Y-axis values normalized to the second-to-last bin from the 3’ end. Each x-axis bin represents 4% of tRNA length. Indicated are major known barriers to RT.
(C) Box plots of full-length fraction per tRNA transcript in datasets from (B) (center line and label, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range).
(D) Correlation plots of unique tRNA gene copy number and corresponding proportion of uniquely aligned tRNA reads in single replicates (same samples as (B)) from *S. cerevisiae*, *S. pombe*, *D. melanogaster* BG3-c2 cells, and hiPSC. Solid blue lines: linear regression model; shaded gray: 95% CI. See also **Figure S2** and **Figure S3**.

Most reads in mim-tRNAseq datasets mapped to cytosolic tRNA, with mitochondrial tRNA fractions ranging from 0.5% in budding yeast to 3% in hiPSC (**Figure S2I**). Importantly, nearly all mapped reads (>96%) spanned the post-transcriptionally added 3’ CCA stretch (**Figure S3A-D**), indicating they originate from mature tRNA. This was not due to bias towards A-ending RNA species, as our workflow accurately captured the 3:1 ratio of two synthetic *E. coli* tRNA-Lys-UUU tRNAs with either 3’-CCA or 3’-CC spiked in prior to library construction. cDNA circularization also did not introduce appreciable length bias, since tRNA coverage after alignment mirrored initial cDNA size (**Figure 3B** and **1A**). Moreover, circularization sequence context is very similar for all cDNAs, as most have a stretch of 1 to 3 non-templated Ts at their 5' ends, corresponding to non-templated A added to cDNA 3' ends by TGIRT (**Figure S3E,F**), which were effectively soft-clipped during GSNAP alignment. Indeed, nucleotide frequencies downstream of non-templated nucleotides were highly similar to those obtained by aligning the 5' ends of catalogued tRNA transcripts (**Figure S3E,F)**.

We asked whether these experimental and computational advances would enable more accurate tRNA quantitation. We first sought to compare our measurements of absolute tRNA abundance to data obtained with an orthogonal, hybridization-based approach. Absolute RNA quantification by e.g. Northern blotting or arrays requires highly specific probes and careful comparisons of signal in serial sample dilutions to calibration curves with known target amounts. The design of specific probes for tRNAs, however, is extremely challenging: even with full-length probes, a difference of at least 8 nucleotides is required to avoid cross-hybridization[1,40]. Probe design is particularly problematic for human tRNA pools, which can contain >400 tRNA species from 57 anticodon families. Since the major tRNA transcript for each anticodon family can differ between cell types[2], probe selection can unduly influence measurement accuracy. By contrast, the 41 anticodon families of *S. cerevisiae* consist of only 54 tRNA species, and most major anticodon variants differ sufficiently in sequence to be distinguished by hybridization. We therefore compared fluorescence intensity measurements for 39 out of the 41 budding yeast anticodon families obtained by direct hybridization to a tRNA microarray[41] to the fraction of reads mapping to those anticodon families in mim-tRNAseq datasets. This comparison yielded a Pearson's r=0.75 (p=3.8 x 10^{-8}), corroborating the quantitative nature of mim-tRNAseq (**Figure S3G**).

The main regulatory elements for tRNA transcription are intrinsic and overlap with conserved structural regions of mature tRNAs, and it remains unclear how selective tRNA gene expression is achieved in metazoans[2–4]. In rapidly growing yeast cells, however, nearly all tRNA loci are transcribed[42,43]. tRNA gene copy number thus positively correlates with the abundance of tRNA anticodon families during exponential growth measured by hybridization (R^2 = 0.47 in microscale thermophoresis[44] and R^2=0.60 in tRNA microarray[41]). We leveraged mim-tRNAseq's superior resolution to probe this relationship at the level of individual tRNA transcripts (**Figure 3D)**. We obtained the highest correlation between gene copy number and tRNA abundance reported so far (adjusted R^2 = 0.92 for *S. cerevisiae* and 0.91 for *S. pombe*, p< 3.71 x 10^{-30}), further underscoring the quantitative nature of mim-tRNAseq. This correlation decreased substantially for *S. cerevisiae* libraries from budding yeast generated by template-switching in otherwise identical RT conditions (R^2 = 0.61, **Figure S3H, I**), consistent with 3' sequence preferences of TGIRT in this set-up[28]. An even more drastic reduction was seen in

S. cerevisiae libraries generated with Superscript III (R^2 = 0.31), which displayed substantial 3' end coverage bias despite high rates of unique read alignment (**Figure S3H, J, K**).

The correlation between gene copy number and tRNA abundance was also lower in *Drosophila* BG3-c2 cells (adjusted R^2 = 0.79) and hiPSC (adjusted R^2 = 0.62). The values were similar regardless of whether we used copy numbers for all predicted human tRNA genes or only the high-confidence tRNA gene set (**Figure S3L**). These findings are consistent with differential tRNA gene usage in distinct cell types[1–4] and highlight that mechanisms beyond gene copy number shape metazoan tRNA pools.

mim-tRNAseq captures differences in tRNA abundance and aminoacylation

To establish whether mim-tRNAseq can accurately detect differences in tRNA abundance, we first compared the tRNA pools of karyotypically normal hiPSC to those in two aneuploid human cell lines (K562 and HEK293T). Of the 368 cytosolic tRNA species resolved quantitatively by mim-tRNAseq, 205 were undetectable in one or more cell lines (≤0.005% of tRNA-mapped reads). Remarkably, more than half of the detectable tRNAs were differentially expressed, some by up to three orders of magnitude (adjusted p<0.05, **Figure 4A** and **Table S1**). By contrast, the relative levels of tRNAs with a given anticodon differed by only up to 1.7-fold among the three cell lines (**Figure 4B**). Of the 47 tRNA anticodon families passing our detection threshold, 11 differed in abundance between HEK293T cells and hiPSC and 21 differed in abundance between K562 cells and hiPSC (**Figure 4B** and **Table S1**). Each cell line exhibited a distinct pattern of tRNA expression, with differences being more pronounced for low-abundance transcripts (**Figure 4C**; base mean expression given by line plot in rightmost panel). These data suggest that different cell types can converge on similar anticodon pools via distinct tRNA transcript subsets, possibly through the relatively stable expression of major tRNA isodecoders[3].

We validated the changes in relative abundance by Northern blotting for two tRNA species: tRNA-Arg-UCU-4 and tRNA-Gly-CCC-2, which differ sufficiently from their isodecoders to avoid probe cross-hybridization, and represent tRNAs with a low and high abundance. tRNA-Arg-UCU-4 and its mouse ortholog are highly expressed in the central nervous system and are also present at low levels in HEK293T cells[2,45]. mim-tRNA seq detected 6 to 8-fold lower levels of tRNA-Arg-UCU-4 in K562 and hiPSCs versus HEK293T (**Table S1**) and a similar 5 to 10-fold decrease was observed by Northern blotting (**Figure 4D,E**). Differential abundance estimates by mim-tRNAseq and Northern blotting were also highly concordant for the abundant tRNA-Gly-CCC-2 (~1% of tRNA-mapped reads; **Figure 4D,E**).

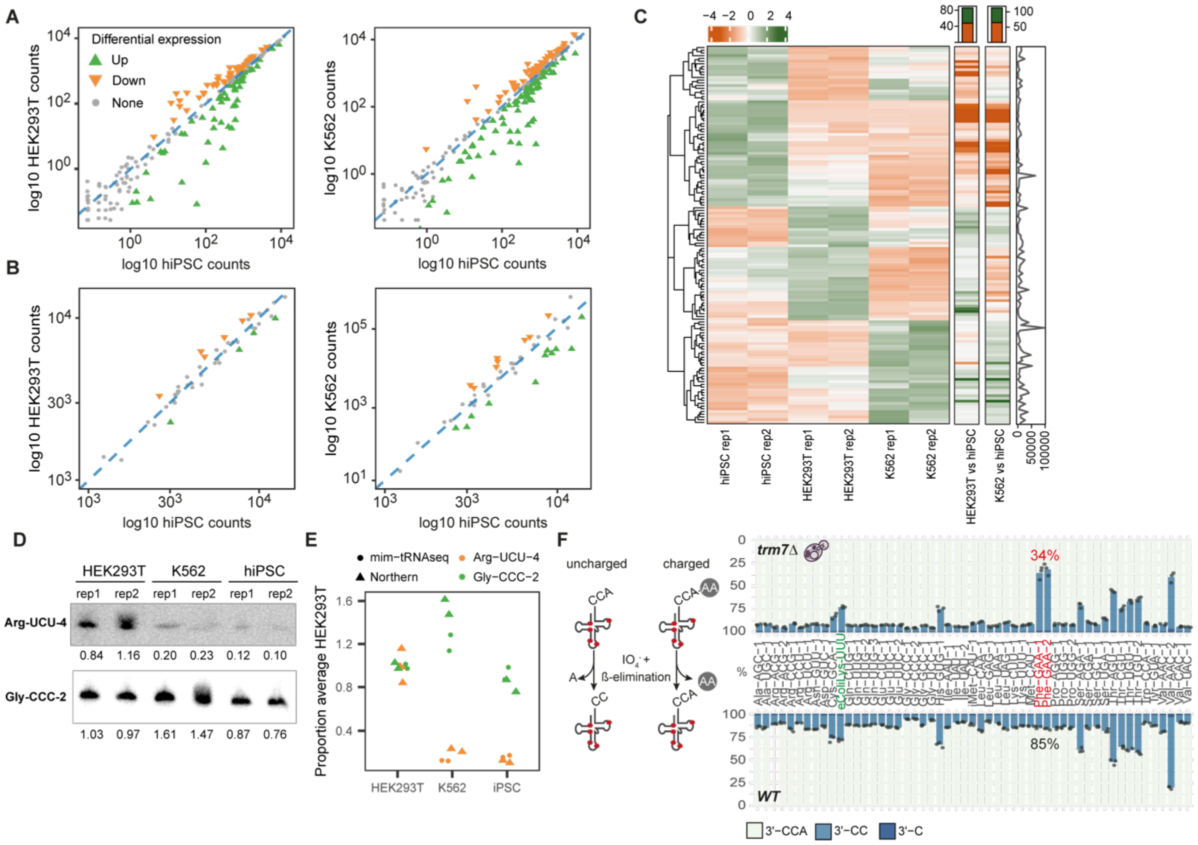


Figure 4 mim-tRNAseq accurately captures differential tRNA expression and aminoacylation with single-transcript resolution.

(A) Differential expression analysis of unique tRNA transcripts in HEK293T and K562 relative to hiPSC. Axes represent log-transformed normalized read counts from DESeq2, with significant down- and up-regulation in hiPSCs indicated with closed orange and green triangles, respectively (FDR adjusted one-sided Wald test p-value ≤ 0.01, n = 2).

(B) Differential expression analysis as in (A) for counts per tRNA anticodon family.

(C) Left panel: hierarchically clustered expression heatmap showing scaled z-score of normalized unique transcript counts in HEK293T, K562 and hiPSC (n = 2). Middle panels: differential expression for HEK293T and K562 relative to iPSC (values: $\log_2$ fold-changes; bar plots: numbers of up- and down-regulated genes in green and orange, respectively). Right panel: base mean normalized per tRNA transcript across all samples.

(D) Northern blot analysis of tRNA-Arg-UCU-4 and tRNA-Gly-CCC-2 in HEK293T, K562, and hiPSC (n = 2, matched samples to those used for mim-tRNAseq). Band intensities were quantified by densitometry and normalized to the mean value for HEK293T.

(E) Relative abundance of tRNA-Arg-UCU-4 and tRNA-Gly-CCC-2 in HEK293T, K562 and hiPSCs measured by mim-tRNAseq (C) or Northern blotting (D), normalized to the mean value for HEK293T (n = 2, matched samples).

(F) tRNA charging analysis in wild type and *trm7Δ S. cerevisiae*. Charged tRNA are represented by proportion of reads with 3'-CCA ends (light green, in %). Light green bars and tRNA-Phe-GAA labels: average charged tRNA fractions (% CCA; n = 3). See also Table S1 and **Figure S3**.

We then confirmed the ability of mim-tRNAseq to accurately measure tRNA aminoacylation. Charged tRNAs have periodate-resistant 3' ends and can be quantified as a fraction of tRNAs with 3'-CCA versus 3'-CC following oxidation and β-elimination[46]. We compared mim-tRNAseq data from oxidized tRNA of wild-type yeast and a *trm7Δ* strain, which has a tRNA-Phe-GAA charging defect[47]. This defect was evident by a 2.5-fold decrease in 3'-CCA proportions for both tRNA-Phe-GAA isodecoders in tRNA pools from *trm7Δ* cells in the absence of other changes in aminoacylation status (**Figure 4F**). Thus, mim-tRNAseq enables the sensitive and accurate quantitation of differences in tRNA abundance or charging.

Improved readthrough facilitates the discovery and annotation of Watson-Crick face tRNA modifications

Mismatches to reference and/or premature RT stop signatures are frequently used to detect Watson-Crick face RNA modifications[6,12–14,16–21], but their analysis is prone to both experimental and computational artefacts[48]. Since tRNA-derived reads are particularly misalignment-prone with standard algorithms (**Figure 2A,D**), this could impact the accuracy of modification calling.

By contrast, mim-tRNAseq abrogated nearly all RT stops and yielded reproducibly high levels of mismatches coinciding with frequently modified tRNA positions (**Figure 5A**). We quantified the extent of readthrough at annotated Watson-Crick face tRNA modifications by calculating the proportion of aligned reads extending past a given position. We then took the minimum value in a 3-nucleotide window centered around it to avoid readthrough overestimation. The median readthrough values we obtained with this approach were ~100% at the most common RT barriers in tRNA such as m^1A, N^1-methylguanosine (m^1G), N^2,N^2-dimethylguanosine ($m^2{}_2G$), and N^3-methylcytosine (m^3C), as well as bulkier modifications like wybutosine (yW) and other wyosine derivatives (**Figure 5B**). All 162 annotated Watson-Crick face modifications in tRNA from budding yeast (100%) and 232 out of the 250 annotated ones in human tRNA (93%) had a readthrough efficiency of >80% (**Table S3**). This is due to both experimental and computational advances, as readthrough was much lower in libraries generated with standard TGIRT conditions or in DM-tRNAseq (**Figure S4A,B)**. By contrast, there was a large variation in bypass of the same modification type in different tRNAs in libraries made with Superscript IV (**Figure S4C**).

The only RT blocks remaining in mim-tRNAseq were at rare hypermodified positions. These include 2-methylthio-derivatives of A37 (ms^2t^6A/ms^2i^6A in human cytosolic tRNA-Lys-UUU and 3-4 mitochondrial tRNAs in *Drosophila* and human cells) and rare stretches of two modified sites ($m^2{}_2G$26/27 and 20/20a N^3-(3-amino-3-carboxypropyl)-uridines (acp^3U); **Figure 5B**; **Figure S2I** and **Figure S4D-E**). These few remaining RT stops do not impact tRNA quantitation, as the cDNA fragments derived from them are sufficiently long (39-56 nt) long for unambiguous read alignment with our pipeline.

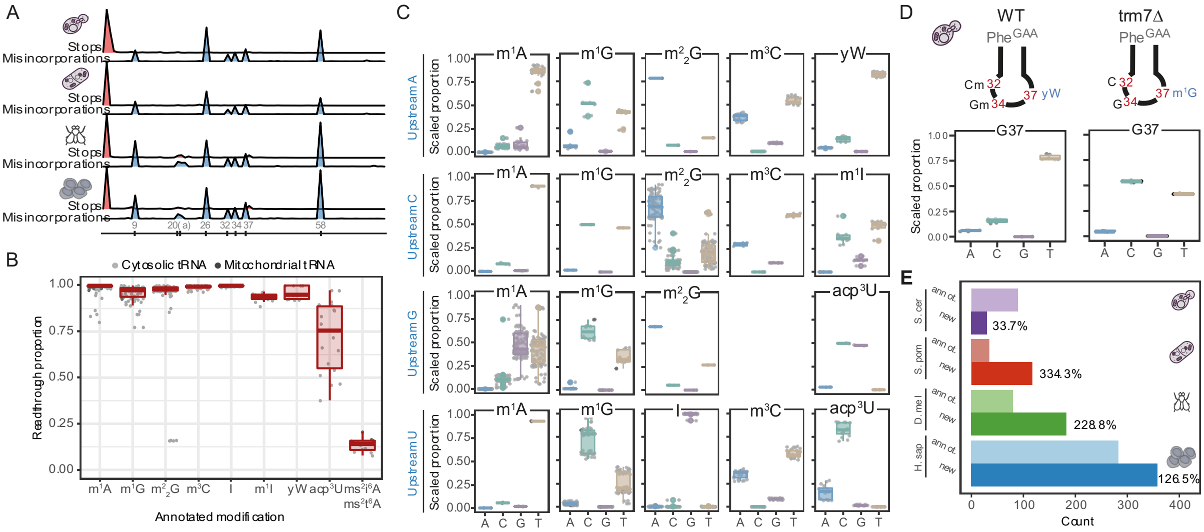


Figure 5 Near-complete modification readthrough in mim-tRNAseq datasets enables modification discovery and annotation.

(A) Average proportion of stops (red) and misincorporation rates (blue) per nucleotide for all tRNA unique transcripts (n = 2) in *S. cerevisiae*, *S. pombe*, *D. melanogaster* BG3-c2 cells, and hiPSC. X-axis: canonical tRNA position at major sites with known RT barriers.

(B) RT readthrough per annotated modification aggregated for cytosolic and mitochondrial tRNA from the four species.

(C) Box plots of misincorporation signatures for annotated modified sites as in (B) (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range). Signatures stratified by upstream context (rows) and modification type (columns); proportion per nucleotide scaled to total misincorporation at this site.

(D) Box plot of misincorporation signature at G37 of tRNA-Phe-GAA from WT and *trm7Δ S. cerevisiae* (n = 3).

(E) Modified site discovery by mim-tRNAseq ("new") compared to misincorporation-inducing modified sites previously annotated in MODOMICS ("annot."). Labels indicate percentage of newly detected sites relative to annotated ones. See also **Figure S4** and Table S2.

We then examined whether different modifications are marked by specific signatures of nucleotide misincorporation. This can depend on the processivity and fidelity of an RT, the reaction conditions, and the sequence context of the modified site[12,13,17,18,20]. Signature analysis is especially challenging when RT stops are pervasive, since mismatches at read ends stemming from non-templated nucleotide addition during RT may manifest as misincorporation and lead to spurious modification calls[48]. As mim-tRNAseq enables near-complete modification readthrough (**Figure 5B**), we examined misincorporation patterns at annotated sites as a function of modification type and sequence context. We found distinct and highly reproducible misincorporation signatures at specific modifications (**Figure 5C**). The ones at m^1G, m^2_2G, and m^3C were largely independent of sequence context, whereas those at m^1A and acp^3U were influenced by the upstream template nucleotide (**Figure 5C**). We also observed distinct signatures for wyosine derivatives, inosine and N^1-methylinosine (m^1I), where the tRNA sequence space is not sufficiently large to explore the impact of sequence context. In contrast, misincorporation signatures of Superscript IV were much less specific for distinct modifications, with a high prevalence of T mismatches regardless of modification type

(**Figure S4F**). A recent comparison of thirteen RTs found a similar lack of distinguishable signatures for m^1G and $m^2{}_2G$[22].

To validate the specificity of these signatures, we compared misincorporation patterns at G37 in tRNA-Phe-GAA from WT and *trm7Δ* yeast (**Figure 5D**). The conversion of m^1G37 to yW in this tRNA requires 2'-*O*-methylation of C32 and G34 by Trm7[49]. Accordingly, the misincorporation signature at G37 in tRNA-Phe-GAA from *trm7Δ* cells was distinct from that in WT (**Figure 5D**) and nearly identical to that of m^1G in our aggregate analysis (**Figure 5C**).

This remarkable consistency enables the use of misincorporation signatures not only for mapping RNA modifications, but also for predicting their identity. We therefore probed our datasets from *S. cerevisiae*, *S. pombe*, *Drosophila* BG3-c2 cells, and human cells for misincorporation-inducing modifications not annotated in MODOMICS. Such sites were identified by a mismatch frequency of >10% and the presence of a distinct misincorporation signature to limit spurious modification calls due to genomic misannotation or SNPs. Modification type was then predicted by combining information on the canonical tRNA position, nucleotide identity, and misincorporation signature in comparison to known sites (**Figure 5C**). Performing this analysis with single-transcript resolution revealed many uncatalogued modifications (**Figure 5E**; **Table S2**), including 30 sites in *S. cerevisiae* and 358 sites in human tRNAs, despite comprehensive existing annotation. Discovery rates were higher in poorly annotated species such as *S. pombe* and *D. melanogaster*. Our predictions generally agreed with prior annotation of modified sites based on RT stops and/or misincorporations (**Table S2**), with some important differences. First, we identified one m^1G9 site, two $m^2{}_2G26$ sites, and seven m^1A58 sites in tRNAs from *S. pombe*, which had not been detected by hydro-tRNAseq[9]. Second, we found no detectable misincorporation at G37 in human tRNA-Pro-AGG or C47d in human tRNA-Ser-AGA and tRNA-Ser-CGA, although these positions have been annotated as m^1G37 and m^3C47d, respectively[9,16]. These differences likely result from our workflow's improved resolution of nearly identical tRNAs, since human tRNA-Pro-UGG and tRNA-Ser-UGA contain m^1G37 and m^3C47d, respectively (**Figure 2D** and **Figure S5C**). These data demonstrate that mim-tRNAseq can map potentially modified tRNA sites and predict modification identity with high sensitivity and specificity.

Accurate quantitation of RNA modification stoichiometry based on misincorporation rates

Proportions of RT stops and/or misincorporations are widely used to estimate tRNA modification levels[9,10,16,19], but whether such measurements are quantitative is unknown. Misincorporation rates at individual modified positions in mim-tRNAseq datasets varied remarkably across tRNA species (**Figure 6A)** despite efficient readthrough (**Figure 5B** and

Figure S5). To test whether this variation reflects modification stoichiometry, we sequenced endogenously modified tRNA from wild-type and mutant yeast lacking m^1G9 (*trm10Δ*)[50] or $m^2{}_2G26$ (*trm1Δ*)[51] pooled in defined ratios prior to library construction. Misincorporations were predictably absent at G9 or G26 sites in samples from the knockout strains. Strikingly, their rates had a near-perfect linear correlation to initial pooling ratios in mixed samples (R^2 = 0.97, **Figure 6B, C**). Mismatch proportions in mim-tRNAseq datasets thus accurately reflect the stoichiometry of m^1G and $m^2{}_2G$, and possibly all other misincorporation-inducing modified tRNA bases. Calibration curves with endogenously modified tRNAs are not feasible for all misincorporation-inducing modifications (**Figure 5B**), however, as some of them are essential for cell viability[52,53].

These findings enabled us to profile modified tRNA fractions with single-transcript resolution in cells from four eukaryotic species. Misincorporation rates were ~100% at all instances of I34 and of wyosine derivatives at position 37, suggesting these modifications are present in stoichiometric levels (**Figure 6D** and **Figure S6A)**. We observed a similar trend for $m^2{}_2G26$, with a clear separation between a majority of fully modified tRNAs and a very small number of transcripts with 10-30% misincorporation. By contrast, the modified fractions of m^1G, m^3C, and m^1A varied substantially among individual tRNAs independently of sequence context (**Figure 6D, Figure S6A,B**). Instances of very high misincorporation were detectable for all three modifications (m^1A: 100%; m^3C: 94%; m^1G: 88%), indicating that mim-tRNAseq can capture high stoichiometry at these sites if it is present (**Figure 6D, Table S3**). However, some tRNAs seem to contain these modifications at sub-stoichiometric levels. Sub-stoichiometric m^3C32 and m^1G37 are consistent with the regulatory rather than structural roles of modifications within the tRNA anticodon loop. The stoichiometry of m^1G37 measured by mim-tRNAseq ranged from 14% to 80% in tRNAs from the four eukaryotic species (**Table S3**). In bacteria, m^1G37 in tRNA-Pro-UGG and tRNA-Pro-GGG aids in reading frame maintenance[54,55]. Eukaryotic cells, however, lack tRNA-Pro-GGG due to toxicity from its high miscoding capacity[56]. A recent study estimated m^1G37 stoichiometry in bacterial tRNA-Pro-UGG by primer extension at 68% in *E. coli* and 73% in *Salmonella enterica*[57]. Our workflow estimated m^1G37 stoichiometry at 53% in yeast tRNA-Pro-UGG and 72% for tRNA-Leu-UAA (**Table S3**). Gel-based primer extension assays with AMV RT, which is blocked by m^1G, were consistent with these measurements (**Figure S5D,E**), providing an orthogonal validation of mim-tRNAseq modification stoichiometry estimates.

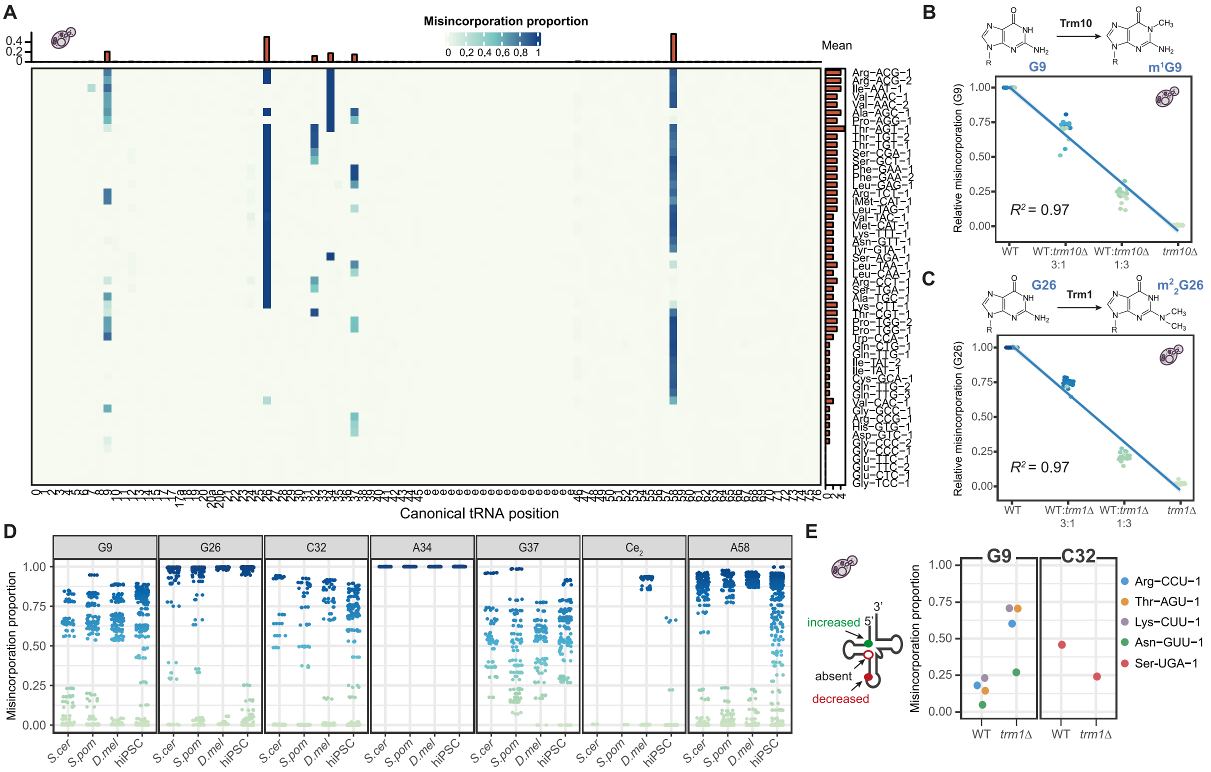


Figure 6 Misincorporation rates in mim-tRNAseq reflect modification stoichiometry.
(A) Global heatmap of average misincorporation proportions in *S. cerevisiae* per unique tRNA transcript with coverage above 2000 reads (n = 2; top bar graph: mean misincorporation per position; right bar graph: number of sites per transcript with detectable misincorporation signatures in ≥10% of reads spanning that position).
(B) Relative misincorporation proportions at G9 in samples from wild-type (WT) *S. cerevisiae* and trm10Δ (lacking m^1G9) or mixes thereof (filtered for clusters with ≥10% misincorporation in WT and scaled to WT proportion; solid blue line: linear regression model; shaded gray: 95% CI).
(C) Analysis as in (B) but for misincorporation at G26 in samples from WT *S. cerevisiae* or *trm1Δ* (lacking m^2_2G26).
(D) Misincorporation proportions per canonical nucleotide position and identity (aggregated per species; e2: second nucleotide of variable loop).
(E) Significant changes in misincorporation rates in *trm1Δ* relative to WT *S. cerevisiae* (FDR-adjusted Chi-square p-value ≤ 0.01, log2 fold-change≥0.5; n=1). See also **Figure S5** and **Figure S6**, and Table S3.

In contrast to the regulatory roles of anticodon loop modifications, m^1A58 is important for the maturation and stability of initiator tRNA-Met in yeast[53] and may play a similar role in other eukaryotic tRNA species. A sequence comparison of budding yeast tRNAs with high or low m^1A58 levels revealed no notable differences, however, indicating that sequence alone is unlikely to be a major determinant of modification stoichiometry at this position (**Figure S6C)**.

To examine whether the stoichiometry of misincorporation-inducing tRNA modifications differs in distinct cell types or states, we calculated log odds ratios of misincorporation proportions across all tRNA positions (see *Methods*). There were very few statistically significant changes when comparing mim-tRNAseq datasets from hiPSCs and HEK293T or K562 cells (**Figure S6D, E**), suggesting most tRNAs are modified to a similar extent in these cell lines. A comparison of datasets from WT and *trm10Δ* or *trm1Δ* yeast, however, revealed the striking precision of our approach in detecting transcripts with large reductions in m^1G9 or m^2_2G26 (**Figure S6F, G**). Unexpectedly, in *trm1Δ* yeast cells that lack m^2_2G26, there were also differences in modification levels at other tRNA sites. These included a 3- to 6.5-fold increase in m^1G9 levels in four tRNAs (tRNA-Lys-CUU-1, tRNA-Thr-AGU-1, tRNA-Arg-CCU-1, tRNA-Asn-GUU-1) and a 2.4-fold decrease in m^3C32 of tRNA-Ser-UGA-1 (**Figure 6E** and **Figure S6G**). m^1G9 levels in tRNA-Lys-CUU-1 and tRNA-Thr-AGU-1 also increase upon Trm10 overexpression in yeast[58]. Sequence comparisons between tRNAs with increased versus unchanged m^1G9 levels in *trm1Δ* cells indicate that a U7:A66 pair rather than G7:C66 pair may be linked to m^1G9 hypermethylation in the absence of $m^2{}_2G26$ (**Figure S6H**). These findings reveal an interdependence between Watson-Crick face modifications at distinct tRNA sites, and suggest that their stoichiometry is determined by structural features.

Discussion

The abundance, charging, and modification status of individual tRNA species can differ in distinct cellular environments. Measuring these properties on a global scale, however, has not been feasible due to technical limitations. No library construction method so far allows the efficient reverse transcription of these highly modified RNAs, while the lack of computational tools suited to the complexity of tRNA sequencing data has been another major methodological gap.

We describe conditions that permit near-complete tRNA modification readthrough by TGIRT, dramatically improving cDNA yield and the fraction of full-length products from tRNA templates. All but one rare tRNA modification roadblock are resolved by mim-tRNAseq, which alleviates the bias of existing tRNA quantification methods towards low-modified tRNAs species. Our library construction protocol circumvents the need to purify enzymes for modification removal[14] or RT[23], which can introduce unwanted variation. We also describe multiple conceptual advances in the analysis of tRNA sequencing data, including the use of modification annotation, which permits position-specific mismatch tolerance during read alignment. Collectively, these advances enable the efficient and accurate mapping and analysis of tRNA-derived reads with single-transcript resolution.

One poignant example of the substantial improvements in our computational workflow concerns tRNAs with I34, which is essential for wobble pairing during decoding. Inosines are interpreted as cytosines during RT, resulting in the stoichiometric presence of G in sequencing reads. When using Bowtie or Bowtie 2 to align tRNA datasets from human cells, we found that reads with G34 were frequently mapped to nearly-identical tRNA isoacceptors with U34. Such misalignment can have wide-ranging implications, since it would not only skew abundance estimates, but can also lead to spurious conclusions about tRNA modification status and stoichiometry. These findings highlight the importance of both sensitivity and accuracy of read alignment in the context of analyzing tRNA transcriptomes.

The robust misincorporation signatures deposited by TGIRT reveal the location, type, and stoichiometry of Watson-Crick face base modifications in tRNA. Calibration measurements of observed versus expected modified fractions in existing approaches for sequencing-based modification analysis are either lacking[14,19] or display a non-linear relationship[23], likely because of persistent RT stops. By contrast, mim-tRNAseq enables efficient readthrough of almost all tRNA modifications, while modification identity is also discernible by highly specific misincorporation patterns. Improved readthrough permits accurate measurements of modification stoichiometry from misincorporation rates alone, evident from calibration curves with near-perfect linear regression for m^1G and m^2_2G (R^2 = 0.97). Performing this calibration with mixtures of endogenously modified tRNA pools shows that our entire workflow is free of bias towards low-modified tRNAs.

Remarkably, we find that while some tRNA positions are almost always fully modified (e.g. m^2_2G26 and I34), others are sub-stoichiometric in some tRNA species. This is in line with a model in which some modifications are deposited because of overlapping substrate specificities in RNA modification enzymes[59]. Indeed, methylation at G9 in some yeast tRNAs is enhanced when they lack m^2_2G26, while methylation of C32 is decreased, suggesting that a conformational change upon m^2_2G26 loss[60] might change the affinity of other modification enzymes for individual tRNAs.

In summary, mim-tRNAseq is a sensitive and accurate start-to-finish technique for quantitation of tRNA abundance and charging, which also reports on the presence and stoichiometry of misincorporation-inducing RNA modifications. The robust library construction workflow and the easy-to-use and freely available computational toolkit make mim-tRNAseq broadly applicable for studying key aspects of tRNA biology in a range of organisms and cell types. Our experimental workflow can also be implemented for the discovery and quantitation of modified sites in other RNA species.

Limitations

mim-tRNAseq currently reports on the presence and stoichiometry of those Watson-Crick face tRNA modifications that elicit robust misincorporation during RT with TGIRT. Various protocols for chemical treatment of the “RT-silent” modifications (e.g. pseudouridine, 5-methylcytosine, 7-methylguanosine) have been developed to enable their detection via misincorporation[61]. Combining them with mim-tRNAseq can expand the modification range detectable in a single sequencing reaction. Our stoichiometry measurements for m^1G and m^2_2G were validated with mixtures of endogenously modified tRNA pools from wild-type and modification-deficient strains, but such validation is not feasible for modifications essential for cell viability. Finally, mim-tRNAseq requires low starting material, but is not compatible with single-cell tRNA profiling.

STAR Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Commercial Reagents		
mTeSR1	STEMCELL Technologies	Cat# 85850
Micro Bio-Spin P30 columns, RNase-free	BioRad	Cat# 7326251
Glycogen	Ambion	Cat# AM9510
T4 Polynucleotide Kinase	New England Biolabs	Cat# M0201L
T4 RNA ligase 2 (truncated KQ)	New England Biolabs	Cat# M0373L
SUPERase In	Ambion	Cat# AM2694
TGIRT	InGex	Cat# TGIRT50
Superscript III	Invitrogen	Cat# 18080044
AMV RT	Promega	Cat# M9004
CircLigase ssDNA ligase	Lucigen	Cat# CL4115K
KAPA HiFi DNA Polymerase	Roche	Cat# KK2102
DNA Clean&Concentrator-5 PCR purification kit	Zymo Research	Cat# D4013
Immobilon NY+	Millipore	Cat# INYC00010
Deposited Data		
Raw and analyzed sequencing data	This paper	GEO: GSE152621
DM-tRNAseq raw data for *H. sapiens* HEK293T	Zheng at al.[14]	GEO: GSE66550
Hydro-tRNAseq raw data for *H. sapiens* HEK293 T-Rex Flp-IN	Gogakos et al.[10]	GEO: GSE95683
QuantM-tRNAseq raw data for *H. sapiens* HEK293 T- Rex Flp-IN	Pinkard et al.[62]	GEO: GSE141436
Experimental Models: Cell Lines		
D. melanogaster BG3-c2 cells	P. Becker, LMU	N/A
HEK293T cells	O. Griesbeck, MPIN	N/A
HPSI0214i-kucg_2 cells	Kilpinen et al.[63]; ECACC	Cat# 77650065
Experimental Models: Organisms/Strains		
S. cerevisiae: strain BY4741	Euroscarf	N/A
S. cerevisiae: strain BY4741 trm1Δ::kanMX	Euroscarf	N/A
S. cerevisiae: strain BY4741 trm7Δ::kanMX	Euroscarf	N/A
S. cerevisiae: strain BY4741 trm10Δ::kanMX	Euroscarf	N/A
S. pombe: strain ED668 h+	S. Braun, LMU	N/A
Oligonucleotides		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNA sequences, primers for library construction, and probes for primer extension and Northern blotting, see Table S4	This paper	N/A
Software and Algorithms		
mim-tRNAseq v0.2.5.6	This paper	https://github.com/nedialkova-lab/mim-tRNAseq
Bowtie v1.2.2	Langmead et al.[34]	http://bowtie-bio.sourceforge.net/index.shtml
Bowtie2 v2.3.3.1	Langmead and Salzberg[35]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
GSNAP v2019-02-26	Wu and Nacu[38]	http://research-pub.gene.com/gmap/
Samtools v1.11	Li et al.[64]	http://samtools.sourceforge.net/
Bedtools v2.29.2	Quinlan and Hall[65]	https://bedtools.readthedocs.io/en/late
BLAST+ v2.9.0	Camacho et al.[66]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
Infernal v1.1.2	Nawrocki and Eddy[67]	http://eddylab.org/infernal/
usearch v10.0.240_i86linux32	Edgar[68]	https://www.drive5.com/usearch
R/DESeq2 v1.26.0	Love et al.[69]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
R/ComplexHeatmap v2.2.0	Gu et al.[70]	https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html
Python/Biopython v1.70	Cock et al.[71]	https://biopython.org/
Other		
Detailed protocol for mim-tRNAseq library construction	This paper	Methods S1

Resource Availability

Lead Contact

Please direct any requests for further information or reagents to the Lead Contact, Danny Nedialkova (nedialkova@biochem.mpg.de).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Raw sequencing data have been deposited in the Gene Expression Omnibus (accession number: GSE152621). The mim-tRNAseq computational pipeline is available under a GNU public License v3 at https://github.com/nedialkova-lab/mim-tRNAseq. A package description and installation guide are available at https://mim-trnaseq.readthedocs.io/.

Experimental Model and Subject Details

Cell Lines and Strains

S. cerevisiae cells (BY4741 wild-type, *trm7*Δ*, trm1*Δ and *trm10*Δ) were grown in yeast extract-peptone-dextrose (YPD) medium. *S. pombe* cells (ED668 h+, ade6-M216 ura4-D18 leu1-32) were cultured in yeast Extract with supplements (YES). Overnight cultures were diluted to an optical density 600 (OD_{600}) of 0.05, grown at 30°C at 250 revolutions per minute, and harvested at OD_{600}=0.5 by rapid filtration and snap-freezing in liquid nitrogen. *D. melanogaster* BG3-c2 cells were cultured at 26°C in Schneider's *Drosophila* Medium (Gibco) supplemented with 10% fetal calf serum, 1% penicillin/streptomycin, and 10 µg/ml human insulin. HEK293T cells were grown at 37°C and 5% CO_2 in DMEM supplemented with 10% fetal bovine serum (Sigma Aldrich). The HPSI0214i-kucg_2 human induced pluripotent stem cell line (obtained from HipSci[63]) was cultured at 37°C and 5% CO_2 in mTeSR1 (STEMCELL Technologies). K562 cells were grown at 37°C and 5% CO_2 in RPMI 1640 supplemented with 10% fetal calf serum and 2mM L-Glutamine.

Method Details

RNA isolation

RNA from *Drosophila* BG3-c2, HEK293T, and human iPS cells was isolated with Trizol (Sigma Aldrich) according to the manufacturer's instructions. For total RNA isolation from yeast, frozen cells were resuspended in 100 mM sodium acetate pH=4.5, 10 mM EDTA pH=8.0, 1% SDS (1 ml per 50 OD_{600} units). An equal volume of hot acid phenol (pH=4.3) was added, and the cell suspension was vortexed vigorously followed by incubation at 65°C for 5 min (*S. cerevisiae*) or 45 min (*S. pombe*) with intermittent mixing. After addition of 1/10 volume 1-Bromo-3-chloropropane (BCP, Sigma Aldrich), samples were centrifuged at 10,000 x *g* for 5 min and the aqueous phase was transferred to a new tube. Following an additional round of hot acid phenol/BCP and a round of BCP only extraction, RNA was precipitated from the aqueous phase by the addition of 3 volumes of ethanol. Pellets were washed in 80% ethanol, briefly air-dried, and resuspended in RNase-free water. For RNA isolation from yeast under conditions that preserve tRNA charging, frozen cells were resuspended in ice-cold 100 mM sodium acetate pH=4.5, 10 mM EDTA pH=8.0. One volume of cold acid phenol (pH=4.3) was added and cells were lysed with 500 μm-diameter glass beads by three rounds of vortexing for 45 sec with a 1-min incubation on ice in between. One-tenth volume of BCP was then added and the samples were centrifuged at 10,000 x *g*/4°C for 5 min, followed by a second round of cold phenol-BCP and one round of BCP-only extraction. RNA was ethanol-precipitated from the aqueous phase and pellets were washed in 80% ethanol containing 50 mM sodium acetate, pH=4.5, briefly air-dried, and resuspended in 50 mM sodium acetate pH=4.5, 1 mM EDTA pH=8.0. RNA concentration was determined with NanoDrop and samples were frozen at -80°C in single-use aliquots.

RNA oxidation and β-elimination

To measure tRNA charging levels, RNA oxidation and β-elimination were performed as described[46] with minor modifications. 25 μg of total RNA were resuspended in 10 mM sodium acetate pH 4.5 and oxidized by the addition of freshly prepared $NaIO_4$ to a final concentration of 50 mM in a 58-μL volume for 30 min at 22°C. The reaction was quenched by addition of 6 μL 1 M glucose for 5 min at 22°C. RNA was purified with Micro Bio Spin P30 columns (BioRad) followed by two rounds of ethanol precipitation in the presence of 0.3M sodium acetate pH=4.5. Pellets were resuspended in 20 μL RNAse-free water and β-elimination was performed by addition of 30 μl 100 mM sodium borate pH=9.5 (freshly prepared) for 90 min at 45°C. RNA was recovered with Micro Bio Spin P30 columns followed by ethanol precipitation, resuspended in RNAse-free water, quantified on a NanoDrop, and stored at -80°C in single-use aliquots.

tRNA purification by gel size selection

Two synthetic RNA standards corresponding to *E. coli* tRNA-Lys-UUU with intact 3'-CCA (5'-GGGUCGUUAGCUCAGUUGGUAGAGCAGUUGACUUUUAAUCAAUUGGUCGCAGGUUC GAAUCCUGCACGACCCACCA-3') or a 3'-CC (5'-GGGUCGUUAGCUCAGUUGGUAGAG CAGUUGACUUUUAAUCAAUUGGUCGCAGGUUCGAAUCCUGCACGACCCACC-3') were added to 5 - 10 μg of total RNA in a 3:1 molar ratio at 0.06 pmol/μg, followed by incubation at 37°C in 50 mM Tris-HCl pH=9.0 to deacylate tRNAs. Deacylation was omitted for samples subjected to oxidation and β-elimination. Total RNA was subsequently dephosphorylated with 10U T4 PNK (NEB) at 37°C for 30 min and purified by ethanol precipitation in 0.3M sodium acetate pH=4.5 with 25 μg glycogen (Ambion) as a carrier. RNA was resolved on a denaturing 10% polyacrylamide/7M urea/1xTBE gel alongside Low Range ssRNA marker (NEB) and visualized with SYBR Gold. Species migrating at the size range of mature tRNAs (60 – 90 nt) were excised and gel slices were crushed with disposable pestles. Low-retention tubes and tips (Biotix, Axygen) were used for all subsequent steps of sequencing library construction to maximize nucleic acid recovery. Following addition of 400 µl gel elution buffer (0.3M sodium acetate pH=4.5, 0.25% SDS, 1mM EDTA pH=8.0), the gel slurry was incubated at 65°C for 10 min, snap-frozen on dry ice, and thawed at 65°C for 5 min. RNA was eluted overnight at room temperature with continuous mixing. Gel pieces were removed with Costar Spin-X centrifuge tube filters and RNA was recovered from the flow-through by ethanol precipitation in the presence of 25 µg of glycogen. This protocol typically recovers 5-10% of total RNA in the 60 – 90 nt fraction, consistent with estimates of tRNA proportions in cells[72].

3' adapter ligation

50 to 200 ng of gel-purified tRNA was ligated to one of four adapters with distinct barcodes (I1:5'-pGAT*ATCGT*CAAGATCGGAAGAGCACACGTCTGAA/ddC/-3'; I2:5'-pGAT*AGCTA*CAAGATCGGAAGAGCACACGTCTGAA/ddC/-3'; I3:5'-pGAT*GCATA*CAAGATCGGAAGAGCACACGTCTGAA/ddC/-3'; I4:5'-pGAT*TCTAG*CAAGATCGGAAGAGCACACGTCTGAA/ddC/-3'; barcodes italicised; underlined sequence complementary to RT primer). The adapters are blocked by the 3' chain terminator dideoxycytidine to prevent concatemer formation, and 5'- phosphorylated to enable pre-adenylation by Mth RNA ligase prior to ligation[31]. Ligation was performed for 3 hours at 25°C in a 20-µl reaction volume containing pre-adenylated adapter and RNA substrate in a 4:1 molar ratio, 1x T4 RNA Ligase Reaction Buffer, 200 U of T4 RNA ligase 2 (truncated KQ; NEB), 25% PEG 8000, and 10 U SUPERase In (Ambion). Ligation products were separated from excess adapter on denaturing 10% polyacrylamide/7M urea/1xTBE gels. Bands migrating at 95-125 nt were excised and ligation products were recovered from crushed gel slices.

Reverse transcription

All reactions contained 125 nM primer, 125 nM template and 500 nM TGIRT (InGex) or 200 U Superscript III (Invitrogen). To prime reverse transcription in template-switching reactions, a synthetic RNA/DNA duplex with a single-nucleotide 3' overhang was generated by annealing an RNA oligonucleotide (5'-GAGCACACGUCUGAACUCCACUCUUUCCCUACACGACGCU CUUCCGAUCU-3') to a DNA oligonucleotide (5'-pRAGATCGGAAGAGCGTCGTGTAGGGA AAGAGTGGAGTTCAGACGTGTGCTCN-3'). The DNA oligonucleotide contained a phosphorylated A/G followed by a random nucleotide at its 5' end, which is a preferred substrate for CircLigase used in subsequent cDNA circularization[30,31]. For primer-dependent reverse transcription reactions, adapter-ligated tRNA and RT primer (5'-pRNAGATCGGAAGA GCGTCGTGTAGGGAAAGAG/iSp18/GTGACTGGAGTTCAGACGTGTGCTC-3'; underlined sequence complementary to 3' adapter, 5'-RN to ameliorate potential biases during circularization) were mixed in MAXYMum Recovery™ PCR Tubes (Axygen), denatured at 82°C for 2 min and annealed at 25°C for 5 min in a Thermocycler. TGIRT reactions were assembled in a 20-µl final volume by combining template and primer with 10 U SUPERase In, 5 mM DTT (from a freshly made 100 mM stock) and manufacturer-recommended TGIRT buffer (20 mM Tris-HCl pH=7.6, 450 mM NaCl, 5 mM $MgCl_2$) or low salt buffer (50 mM Tris-HCl pH=8.3, 75 mM KCl, 3 mM $MgCl_2$). After TGIRT addition, samples were pre-incubated at reaction temperature for 10 min (primer-dependent reactions) or 22°C for 30 min (template-switching reactions), initiated by addition of dNTPs to a final concentration of 1.25 mM, and incubated in a Thermocycler for 1 hour or 16 hours. For Superscript III RT, template and primer were denatured at 75°C for 5 min and chilled on ice, and reverse transcription was performed in the presence of 1X First-Strand Buffer, 5 mM DTT, 0.5 mM dNTPs, 10 U SUPERase In, and 200 U Superscript III (Invitrogen) at 57°C for 60 min.

Template RNA was subsequently hydrolyzed by the addition of 1 µl 5M NaOH and incubation at 95°C for 3 min and reaction products were separated from unextended primer on denaturing 10% polyacrylamide/7M urea/1xTBE gels. Gels were stained with SYBR Gold, the region between 60 and 150 nt was excised and cDNA was eluted from crushed gel slices in 400 µl 10 mM Tris-HCl pH=8.0, 1 mM EDTA at 70°C/2000 rpm for 1 hour in a Thermoblock, followed by ethanol precipitation in 0.3M sodium acetate pH=5.5 in the presence of 25 µg glycogen.

cDNA circularization and library construction PCR

Purified cDNA was circularized with CircLigase ssDNA ligase (Lucigen) in 1x reaction buffer supplemented with 1 mM ATP, 50 mM $MgCl_2$, and 1M betaine for 3 hours at 60°C, followed by enzyme inactivation for 10 min at 80°C. One-fifth of circularized cDNA was directly used for library construction PCR with a common forward (5'-AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCT*C-3') and unique indexed reverse primers (5'-CAAGCAGAAGACGGCATACGAGAT*NNNNNN*GTGACTGGAGTTCAGACGTGT*G-3', asterisks denote a phosphorothioate bond and *NNNNNN* corresponds to the reverse complement of an Illumina index sequence). Amplification was performed with KAPA HiFi DNA Polymerase (Roche) in 1x GC buffer with initial denaturation at 95°C for 3 min, followed by five to six cycles of 98°C for 20 sec, 62°C for 30 sec, 72°C for 30 sec at a ramp rate of 3°C/sec. PCR products were purified with DNA Clean&Concentrator 5 (Zymo Research) and resolved on 8% polyacrylamide/1xTBE gels alongside pBR322 DNA-MspI Digest (NEB). The 130-220 bp region of each lane was excised and DNA was eluted from crushed gel slices in 400 µl water with continuous mixing at room temperature overnight. After ethanol precipitation in 0.3M sodium acetate pH=5.5 and 25 µg glycogen, libraries were dissolved in 10 µl 10 mM

Tris-HCl pH=8.0, quantified with the Qubit dsDNA HS kit, and sequenced for 150 cycles on an Illumina NextSeq platform.

Northern blotting

Two micrograms of total RNA were resolved on denaturing 10% polyacrylamide/7M urea/1xTBE gels. RNA was transferred to Immobilon NY+ membranes (Millipore) in 1xTBE for 40 min at 4mA/cm^2 on a TransBlot Turbo semi-dry blotting apparatus (Bio-Rad) and crosslinked at 0.04 J in a Stratalinker UV crosslinker. Membranes were incubated at 80°C for one hour and pre-hybridized in 20 mM Na_2HPO_4 pH=7.2, 5xSSC, 7% SDS, 2x Denhardt, 40 µg/ml sheared salmon sperm DNA at 55°C for 4 hours. The buffer was exchanged and 10 pmol 5'-end ^{32}P-labelled probe (Arg-UCU-4: 5'-CGGAACCTCTGGATTAGAAGTCCAGCGCG CTCGTCC-3'; Gly-CCC-2: 5'-CGGGTCGCAAGAATGGGAATCTTGCATGATAC-3') was added, followed by hybridization at 55°C overnight. Membranes were washed three times in 25 mM Na_2HPO_4 pH=7.5, 3xSSC, 5% SDS, 10x Denhardt, once in 1xSSC, 10% SDS, and exposed to PhosphorImager screens, which were subsequently scanned on a Typhoon FLA 9000 (GE Healthcare). Band intensity was quantified with ImageQuant (GE Healthcare).

Primer extension analysis of m^{1}G37

The extent of RT arrest at m^{1}G37 in tRNA-Leu-UAA and tRNA-Pro-UGG from *S. cerevisiae* was quantified via primer extension with AMV RT, an enzyme with low processivity at this modification[22]. The primers were designed to enable a 4-nucleotide extension to m^{1}G37 (tRNA-Leu-UAA: 5'-CGCGGACAACCGTCCAAC-3'; tRNA-Pro-UGG: 5'-TGAACCCAGGGCC TCT-3') and end-labeled with **γ**-^{32}P-ATP. 3 µg of total RNA from exponentially growing cells was mixed with 1 pmol end-labeled primer and incubated at 95°C for 3 min followed by slow cooling to 37°C. RT reactions were assembled by adding 15 U AMV RT (Promega), 0.5 mM dNTPs, 20U SUPERase In (Ambion) and 1X AMV RT buffer in a 5-µl volume. Following incubation at 37°C for 45 min, reactions were stopped by addition of 5 µl 2X RNA loading dye (47.5% Formamide, 0.01% SDS, 0.01% bromophenol blue, 0.005% Xylene Cyanol, 0.5 mM EDTA), boiled at 95°C for 5 min, and resolved on a denaturing 15% PAA/7M urea/1X TBE gel. The gel was exposed at -80°C to a PhosphorImager screen, which was scanned on a Typhoon FLA 9000 (GE Healthcare). Band intensity was quantified with ImageQuant (GE Healthcare).

Quantification and Statistical Analysis

Read preprocessing

Sequencing libraries were demultiplexed using cutadapt v2.5[73] and a fasta file (barcodes.fa) of the first 10 nt for the four different 3' adapters (see *3' adapter ligation* above). Indels in the alignment to the adapter sequence were disabled with *--no-indels.* Following demultiplexing, reads were further trimmed to remove the two 5'-RN nucleotides introduced by circularization from the RT primer with *-u 2*. In both processing steps, reads shorter than 10 nt were discarded using *-m 10.* Example commands for demultiplexing and 5' nucleotide trimming:

```
    cutadapt --no-indels -a file:barcodes.fa -m 10 -o mix1_{name}_trim.fastq
.gz mix.fastq.gz
    cutadapt -j 40 -m 10 -u 2 -o mix1_barcode1_trimFinal.fastq.gz mix1_
barcode1_trim.fastq.gz
```

Modification indexing and clustering

mim-tRNAseq uses modification data from MODOMICS[37] to guide accurate alignment of short reads from tRNAs. A prepackaged set of data is available for *S. cerevisiae*, *S. pombe*, *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens* and *E. coli*, and can be specified with the *--species* parameter. For other organisms, mim-tRNAseq requires a fasta file of predicted genomic tRNA sequences (*-t*) and a tRNAscan-SE "out" file containing information about tRNA introns (*-o*), both of which should be obtained from GtRNAdb[8] or from running tRNAscan-SE[74] on the genome of interest. Lastly, a user-generated sample input file is required which contains two tab-separated columns specifying the path to trimmed tRNA-seq reads in fastq format, and the experimental condition of each fastq file. Additionally, a mitochondrial tRNA fasta reference is supplied with the prepackaged data inputs listed above, or may be supplied (*-m*) for custom genomes as a fasta file obtained from mitotRNAdb[36]. mim-tRNAseq automatically removes nuclear-encoded mitochondrial tRNAs (nmt-tRNAs) and tRNA species with undetermined anticodons (where applicable), generates mature, processed tRNA sequences (with appended 3'-CCA if necessary, 5'-G for tRNA-His, and spliced introns), and fetches species-matched MODOMICS entries accordingly. Transcript sequences are then matched to MODOMICS entries using BLAST in order to index all known instances of residues modified at the Watson-Crick face within each tRNA. An additional modifications file for modifications reported in the literature but not yet added to MODOMICS may be supplied and is automatically processed by the pipeline (e.g. I34 annotation[9,75]). tRNA clustering is enabled with the *--cluster* parameter, which utilizes the usearch *--cluster_fast* algorithm[68] to cluster tRNA sequences by a user-defined sequence identity threshold (customizable with *--cluster-id*). Regardless of the chosen threshold, only tRNAs sharing an anticodon are clustered to maintain isoacceptor resolution in cases where tRNA transcripts differ by a single nucleotide in the anticodon. The clusters are re-centered based on the number of identical sequences, and this is used to re-cluster and improve the selection of a representative centroid/parent sequence for each cluster (https://www.drive5.com/usearch/manual7/recenter.html). Polymorphisms between cluster members are recorded, and mismatches at these sites during alignment are tolerated, but they are not included in misincorporation analysis for modified sites. Since inosine is interpreted as a G during reverse transcription, annotated inosines are changed to G in tRNA reference sequences.

Read alignment and modification discovery

After clustering, reads are aligned using GSNAP to the representative centroid cluster sequences of mature tRNA transcripts. By enabling SNP-tolerant alignment with *--snp-tolerance,* the indexed modified sites are treated as pseudo-SNPs to allow modification-induced mismatches at these sites in a sequence- and position-specific manner. Soft-clipping during alignment in combination with the GSNAP parameter *--ignore-trim-in-filtering=1* ensures that non-templated nucleotide extensions are not counted as mismatches during alignment. Mismatch tolerance outside of indexed SNPs is controlled using the *--max-mismatches* parameter, where an integer of allowed mismatches per read can be provided, or a relative mismatch fraction of read length between 0.0 and 0.1 can be supplied (default 0.1). If *--remap* is specified, then misincorporation analysis is performed and new, unannotated modifications are called where *--misinc-thresh* (total misincorporation proportion at a residue; default is 0.1 or 10%) and *--min-cov* (minimum total coverage for a cluster) regulate the calling of new modifications, which exclude mismatch sites between cluster members appearing as misincorporations in this analysis. The existing SNP index is then updated with these new sites, and realignment of all reads is performed with a mismatch tolerance set using *--remap-mismatches*. New potential inosine sites are classified for position 34 where a reference A nucleotide is misincorporated with a G in 95% or more total misincorporation events. Both *--remap* and *--max-mismatches* are extremely useful for detecting unknown modifications in poorly annotated tRNAs, subsequently allowing more accurate and efficient read alignment, which improves the results of all downstream analyses. Users should consider a low mismatch tolerance during remap to avoid inaccuracy resulting from lenient alignment parameters. We recommend a relative mismatch fraction of 0.075 during remapping (*--remap-mismatches* 0.075). Only uniquely mapped reads are retained for post-alignment analyses.

Read deconvolution

This process aims to recapitulate the single-transcript resolution of *--cluster-id 1* (see above), but with the alignment accuracy and decreased multi-mapping achieved at lower *--cluster-id* values. The deconvolution algorithm first searches each cluster of tRNA reference sequences for single-nucleotide differences that distinguish among those. For this, each nucleotide in a reference sequence is assessed for uniqueness at that position when compared to all other reference sequences in the cluster. If a nucleotide is unique in position and identity for a specific tRNA reference in the cluster, it is catalogued. Then, after alignment, each read is assessed for mismatches to the cluster parent to which it was aligned. These are then scanned individually to find potential matches to the previously catalogued set for the cluster which can distinguish unique tRNA references. Based on the presence and identity of a unique distinguishing mismatch, a read is then be assigned to a specific tRNA reference within a cluster. Depending on the organism and/or cluster ID threshold, unique distinguishing mismatches may not always be present for all tRNA references in a cluster. Reads without distinguishing mismatches remain assigned to the cluster parent, which is then marked as not fully deconvoluted. Using this algorithm, uniquely aligned reads are assigned to individual tRNA sequences in the reference (where possible) before any of the downstream analyses detailed below. For differential expression analyses of reads summed per tRNA anticodon, read deconvolution is not necessary and therefore not performed.

Modification, RT stop, readthrough and 3' CCA end analyses

Following read deconvolution, all other mismatched positions for the read are extracted from alignment records in bam files, and converted into positions relative to the unique transcript to which the read was assigned (or the cluster parent if definitive assignment is not possible). The identity of the misincoporated nucleotide is recorded to enable signature analysis, and the counts of mismatches for each of the four nucleotides for all reads with the misincorporation are normalized relative to total read coverage at that position. Stops during reverse transcription are extracted from the alignment start position of each read relative to the reference (5' read ends correspond to cDNA 3' ends during RT) and normalized to total read counts for the unique tRNA. Similarly, readthrough for each position is calculated as the fraction of reads that stop at a position relative to read coverage at each position (as opposed to stop proportions which are normalized to total tRNA read coverage). This value is then subtracted from one to estimate the proportion of reads per position that extend beyond that site, and the minimum value in a 3-nucleotide window centered around the modification is recorded. Using a 3-nucleotide window ensures that potential variance in the position at which the RT stalls due to the modification is accounted for. Taking the minimum value of readthrough for these 3 nucleotides reduces the likelihood of readthrough overestimation. Misincorporation, stop data, and readthrough per unique tRNA sequence, per position are output as tab-separated files, and global heatmaps showing misincorporation and stop proportions across all unique tRNA sequences are plotted per experimental condition. Misincorporation signatures are also plotted for well-known conserved modified tRNA sites (9, 20, 26, 32, 34, 37 and 58) separated by upstream and downstream sequence context to assess potential factors influencing misincorporation signatures. Lastly, the dinucleotide at the

3' ends of reads is quantified, so long as the read aligns to the conserved 3'-CCA tail of the reference. Proportions of transcripts with absent 3' tails, 3'-C, 3'-CC and 3'-CCA are calculated per unique tRNA sequence and plotted pairwise between conditions for quantitation and comparison of functional tRNA pools, or tRNA charging fractions in periodate oxidation experiments.

Post-alignment analyses

The cluster deconvolution algorithm allows coverage analysis, novel modification discovery and read counting for tRNA quantitation to be done at the level of unique tRNA sequences. Coverage is calculated as the depth of reads at all positions across a tRNA sequence and plotted using custom R scripts. Cytosolic tRNAs with low read coverage can be filtered at the coverage analysis step by supplying a minimum coverage threshold to *--min-cov*. Unique tRNA sequences filtered out here are excluded from all downstream analyses, except differential expression analysis by DESeq2[69] where all unique tRNA sequences are included. Normalized coverage (read fraction relative to library size) is plotted per sample in 25 bins across gene length in a metagene analysis. Normalized coverage is also scaled relative to the second last bin to account for potential differences in 3' CCA intactness. Read counts per unique tRNA sequence are summed to calculate read counts per isoacceptor family (all tRNAs sharing an anticodon). These counts are subsequently used by a DESeq2 pipeline for count transformations, sample distance analysis using distance matrix heatmaps, PCA plots, and differential expression analysis at the level of isoacceptor families and unique tRNA transcripts (only for completely resolved clusters). In the case that only one experimental condition is supplied, or if there are no replicates for one or more conditions, differential expression analysis is not performed on these samples, but a normalized counts table is still produced for investigations into tRNA abundance.

Data analysis with the mim-tRNAseq package

The following parameters were used for the analysis of mim-tRNAseq generated sequencing datasets (see *mimseq --help* or https://mim-trnaseq.readthedocs.io/en/latest/intro.html for full explanations of parameters):

S. cerevisiae: `--cluster --cluster-id 0.90 --snp-tolerance --min-cov 2000 --max-mismatches 0.1 --control-condition Exp --cca-analysis --remap --remap-mismatches 0.075`

S. pombe: `--cluster --cluster-id 0.95 --snp-tolerance --min-cov 2000 --max-mismatches 0.1 --control-condition Exp --cca-analysis --remap --remap-mismatches 0.075`

D. melanogaster: `--cluster --cluster-id 0.95 --snp-tolerance --min-cov 2000 --max-mismatches 0.1 --control-condition bg3 --cca-analysis --remap --remap-mismatches 0.075`

H. sapiens: `--snp-tolerance --cluster --cluster-id 0.95 --min-cov 2000 --max-mismatches 0.1 --control-condition kiPS --cca-analysis --remap --remap-mismatches 0.075`

tRNA read alignment with Bowtie and Bowtie 2

To test previously used alignment strategies as in DM-tRNAseq[14] or ARM-seq[15], a non-redundant set of reference human tRNA transcripts was created by fetching the full set of 610 predicted tRNA genes for human genome hg19 from GtRNAdb[8] and the 22 mitochondrially encoded human tRNA genes from mitotRNAdb[36]. Following intron removal and addition of 3' CCA (for nuclear-encoded transcripts) and 5'-G (for tRNA-His), a curated set of 596 genes (excluding anticodon/isotype mismatch and nuclear-encoded mitochondrial tRNAs) were collapsed into 420 unique sequences. Corresponding Bowtie and Bowtie 2 indices were built from this set of references. Bowtie alignment was performed with a maximum of 3 allowed mismatches per read (*-v 3*), filtering for uniquely aligning reads (*-m 1*) and ensuring the best alignment from the best stratum (i.e. reads with the least number of mismatches) were reported (*--best --strata*). Bowtie 2 alignments were performed in very sensitive local mode (*--very-sensitive --local*) and up to 100 alignments per read were allowed (*-k 100*). Read quality scores were ignored for alignment score and mismatch penalty calculation (*--ignore-quals*) with increased penalties for ambiguous characters ("N") in reference or read (*--np 5*). Output alignments in SAM format were reordered to match read order in input fastq file (*--reorder*). The alignment commands for both algorithms are given below:

```
bowtie -v 3 -m 1 --best --strata --threads 40 -S
```

```
bowtie2 --local -x -k 100 --very-sensitive --ignore-quals --np 5 --reorder
-p 40 -U
```

QuantM-tRNAseq data for HEK293 T-Rex Flp-IN cells downloaded from the NCBI Gene Expression Omnibus repository was adapter-trimmed and analyzed with Bowtie 2 as described[7]:

```
bowtie2 --local --score-min G,1,8 -D 20 -R 3 -N 1 -L 10 -i S,1,0.5
```

Sequence logo analysis

Alignment files for uniquely aligned reads from human HEK293T and *S. cerevisiae* cells were utilized to generate frequency plots of untemplated nucleotide additions by TGIRT, and 5' sequence logos in each sample. Briefly, CIGAR strings for each unique alignment were assessed for GSNAP soft-clipped nucleotides representing untemplated additions. The number of additions per read were recorded and plotted as frequency histograms. Since a total of 3 additions or less were present in >90% of reads analyzed, we generated sequence logos using the Python package Logomaker[76] for these reads using soft-clipped residues and the first 10 nucleotides after them. For the logo representing all catalogued tRNA genes, we used mature tRNA transcript sequences from each genome present in GtRNAdb, and generated a multiple sequence alignment of these using Infernal[67]. A sequence logo was then generated from the first 11 nucleotides of each aligned tRNA transcript (in order to include G-1 for tRNA-His, plus 10 additional nucleotides as in the uniquely aligned read logo above).

Differential modification analysis

To test for global differential modification between two conditions, first, misincorporation proportion and coverage data generated by mim-tRNAseq were used to calculate absolute counts of modified and unmodified bases per position for each resolved tRNA transcript. Then, log odds ratios (logOR) were calculated for each position, *x,* as follows:

$$logOR_x = log\left(\frac{M_a/M_b}{U_a/U_b}\right)$$

where M_a and M_b are the counts of modified nucleotides at position *x* in condition *a* and *b*, and U_a and U_b are the counts of unmodified nucleotides at position *x* in condition *a* and *b*, respectively. Significance for each logOR was determined with chi-square tests using the respective modified and unmodified nucleotide counts for each condition in a two-dimensional contingency table for the Pearson's chi-square test. Correction for multiple testing was performed with the FDR method. Following significance tests, logOR values were filtered for FDR-adjusted p-values ≤ 0.01, absolute $\log_2$ fold-changes ≥ 1, and total misincorporation at the given position of 10% or more in at least one of the conditions to ensure only sites with high-confidence misincorporation levels are kept. The resulting logOR were used in generating heatmaps for individual contrasts between cell types or experimental conditions.

Acknowledgements

We are grateful to K. Strasser, F. Kliem and M. Moschref for assistance with library construction, the Max Planck Institute of Biochemistry Core Facility for high-throughput sequencing, S. Braun, P. Becker and O. Griesbeck for reagents, T. Wu for help with GSNAP, J. Gagneur and M. Colomé-Tatché for data analysis advice, A. Andersen (Life Science Editors) for manuscript feedback, F. Bonneau and E. Conti for sharing resources, and M. Helm for insightful suggestions. D.D.N. acknowledges Wellcome Trust Sanger Institute as the source of HPSI0214i-kucg_2 hiPSC line, which was generated under the Human Induced Pluripotent Stem Cell Initiative supported by the Wellcome Trust (WT098051) and the NIHR/Wellcome Trust Clinical Research Facility, and acknowledges Life Science Technologies Corporation as the provider of Cytotune. A.B. and G.R. were supported by the International Max Planck Research School for Molecular Life Sciences. This work was funded by the Max Planck Society and the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC-StG-2018 No. 803825 to D.D.N.).

Supplemental information

Figure S1

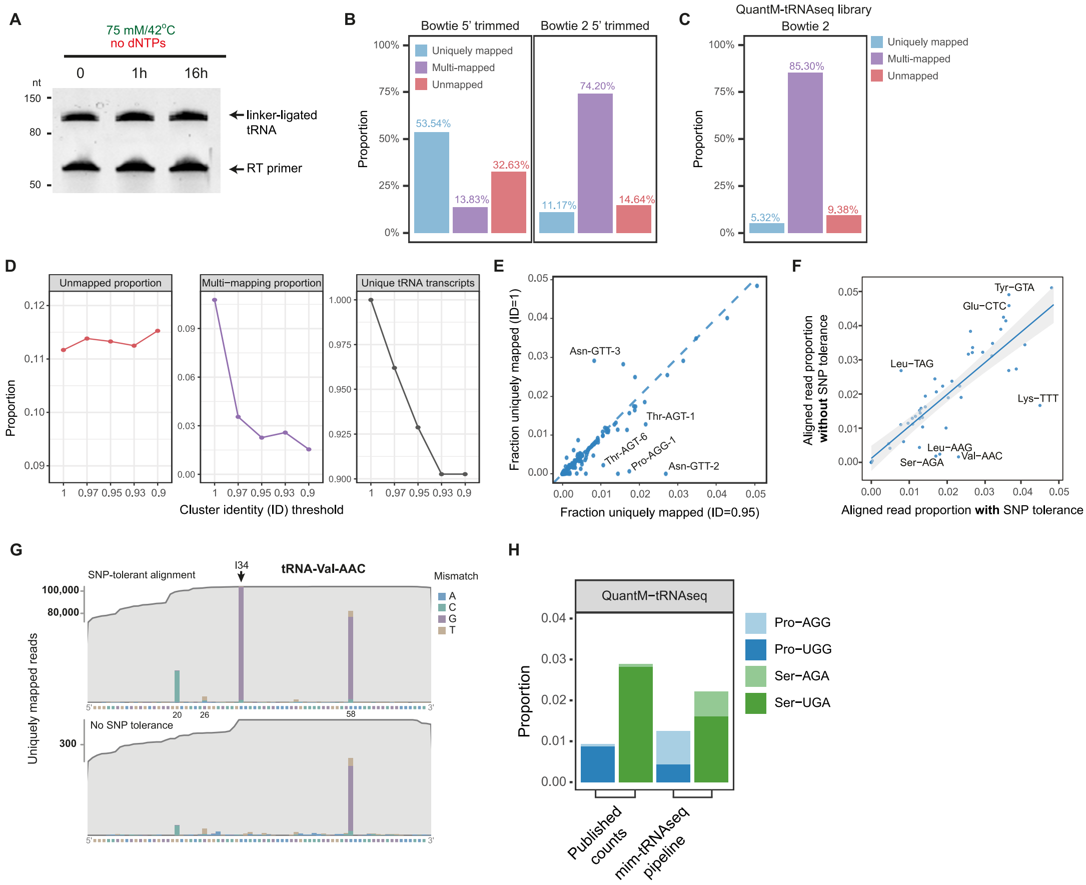


Figure S1 Development and optimization of the mim-tRNAseq workflow. Related to Figure 1 and Figure 2.
(A) Stability of linker-ligated tRNA from *S. cerevisiae* cells in TGIRT reactions. The reactions were assembled in the absence of dNTPs and stopped by addition of gel loading buffer immediately after assembly (0 h) or following incubation at 42°C for 1 h or 16 h and visualized by SYBR Gold staining after separation on a 10% denaturing polyacrylamide gel.
(B) Alignment statistics of HEK293T tRNA with Bowtie and Bowtie 2 after trimming 3 nucleotides from 5’ read ends (n = 1).
(C) Alignment statistics of published HEK293 T-Rex Flp-IN tRNA data from QuantM-tRNAseq, aligned using Bowtie 2 and collapsed unique tRNA sequences with the published settings (Pinkard et al., 2020).
(D) Unmapped read proportion, multi-mapping read proportion, and resolution of unique tRNA transcripts by the deconvolution algorithm with different cluster identity (ID) thresholds for tRNA from HEK293T (n = 1).
(E) Fraction of reads uniquely mapped to deconvoluted tRNA transcripts with optimized cluster ID for humans (0.95) versus a collapsed non-redundant reference (cluster ID = 1).
(F) Fraction of reads uniquely mapped to tRNA anticodon families in HEK293T data with and without GSNAP SNP-tolerance.
(G) Read coverage for HEK293T tRNA-Val-AAC analyzed with and without GSNAP SNP-tolerance. Y-axis shows coverage in read counts. X-axis shows tRNA-Val-AAC reference sequence. Mismatches in aligned reads depicted by stacked bar-plots for each mismatch position.
(H) Comparison of read proportions for inosine 34 (I34) and uridine 34 (U34)-containing serine (Ser) and proline (Pro) isoacceptors in HEK293 T-Rex Flp-IN tRNA data from QuantM-tRNAseq using published counts, or counts obtained after re-analyzing the same data with the mim-tRNAseq computational pipeline (n = 1).

Figure S2

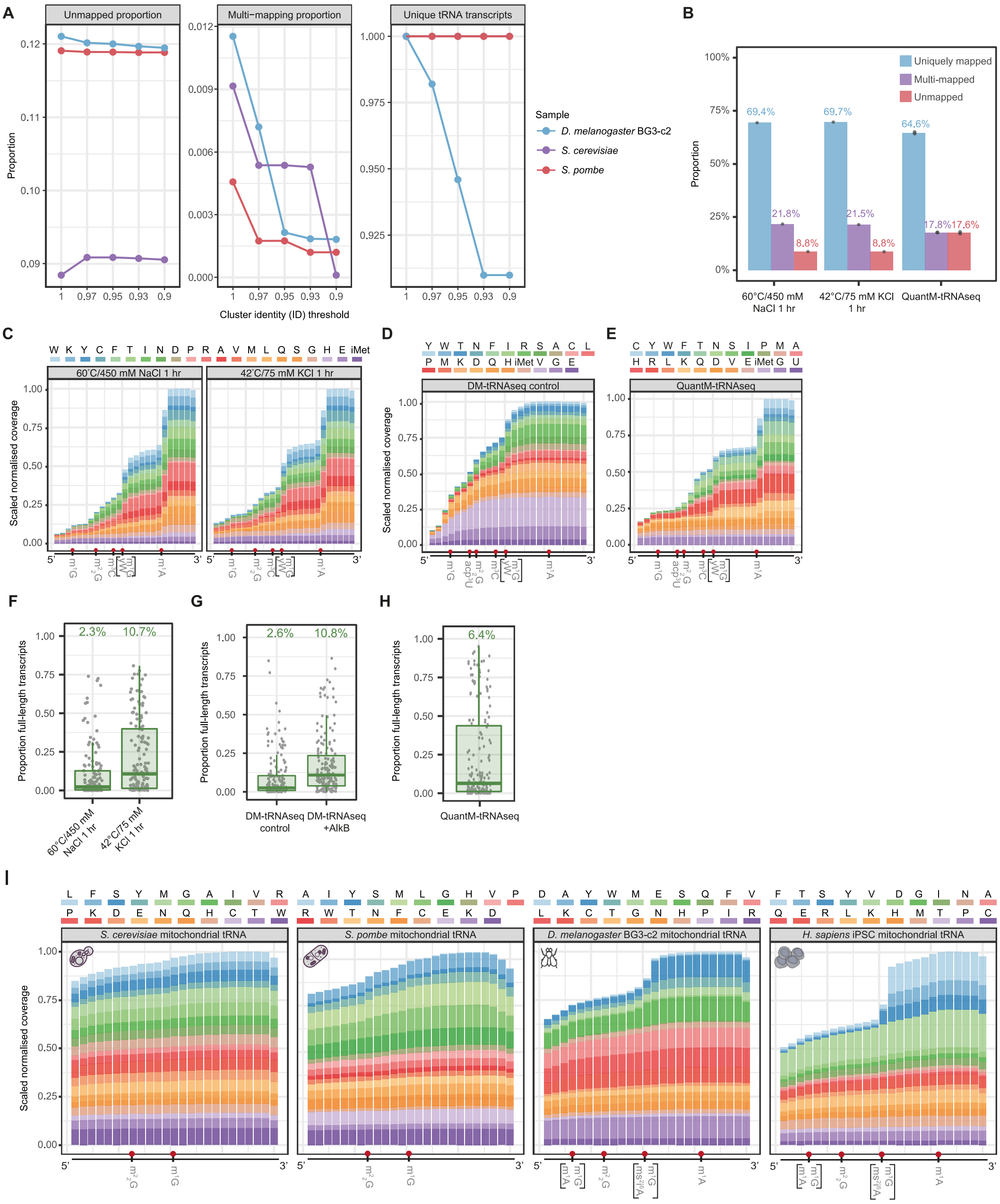


Figure S2 mim-tRNAseq improves resolution, coverage, and full-length transcript recovery. Related to Figure 3.
(A) Unmapped read proportion, multi-mapping read proportion, and resolution of tRNA transcripts with different cluster ID thresholds for tRNA sequencing data from *S. cerevisiae*, *S. pombe*, and *D. melanogaster* BG3-c2 cells (n = 1).
(B) Alignment statistics for tRNA sequencing data from HEK293-derived cell lines generated under the indicated RT reaction conditions (n = 1) or using published QuantM-tRNAseq datasets (n = 2). All data analyzed with the mim-tRNAseq pipeline.
(C - E) Metagene analysis of scaled sequence coverage across nuclear-encoded human tRNA isotypes in libraries generated under the RT reaction conditions given in panel labels (C) or published DM-tRNAseq (D) and QuantM-tRNAseq (E) datasets. All analyses were performed with the mim-tRNAseq computational pipeline.
(F - H) Box plot of full-length fraction per tRNA transcript in datasets from (C - E) (center line and label, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range).
(I) Metagene analysis of scaled sequence coverage across mitochondrially encoded tRNA isotypes (one-letter amino acid code) in mim-tRNAseq data from the indicated cells. Isotypes in all plots are ordered per sample by differences between 3' and 5' coverage (decreasing order from top to bottom; n = 1). Y-axis values normalized to the second-to-last bin from the 3' end. Each x-axis bin represents 4% of tRNA length. Indicated are annotated nuclear and mitochondrial tRNA modifications known to pose barriers to RT.

Figure S3

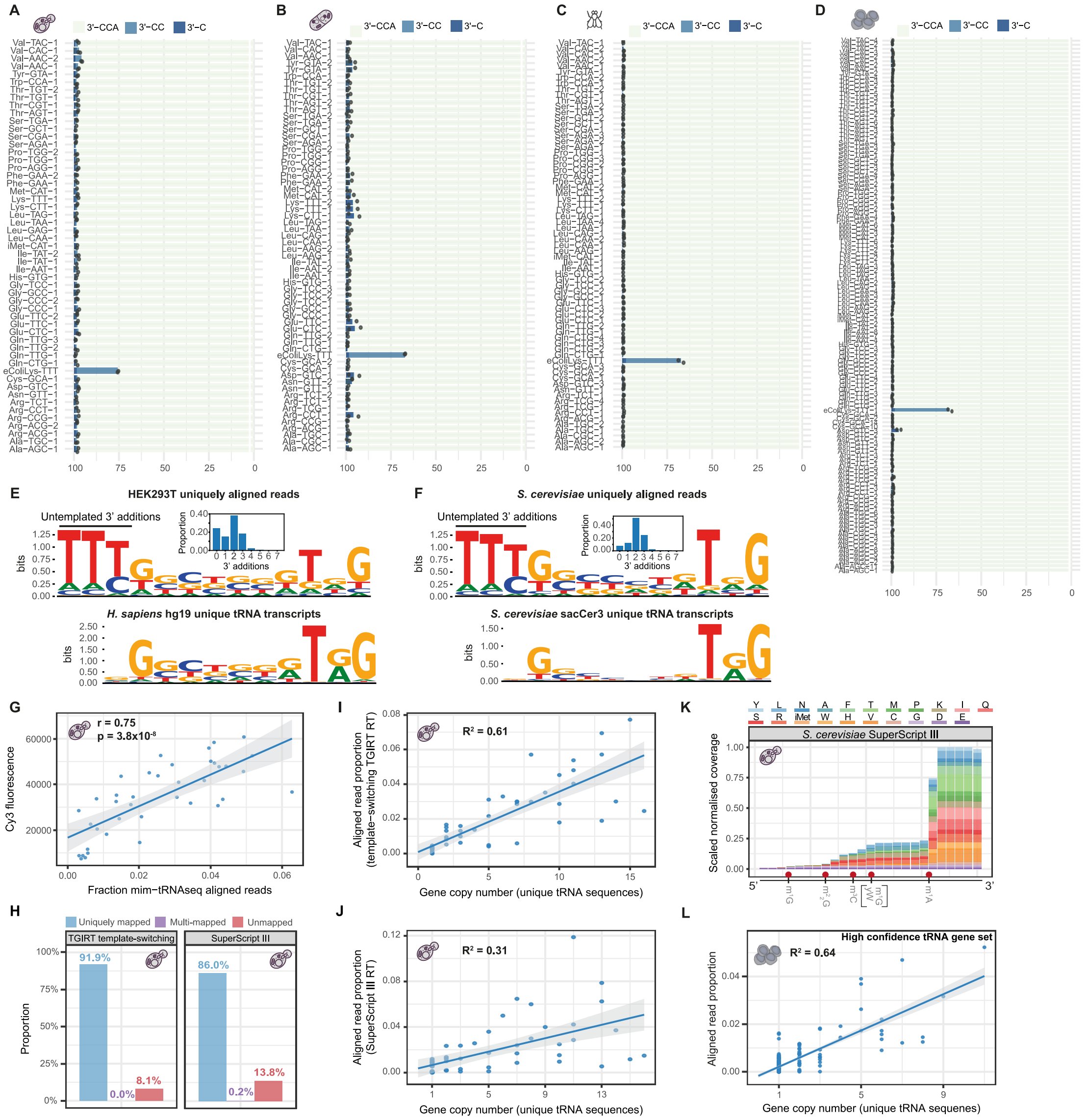


Figure S3 Improved tRNA quantification accuracy with mim-tRNAseq. Related to Figure 3.

(A - D) tRNA 3'-CCA completeness measured by mim-tRNAseq in datasets from *S. cerevisiae* (A), *S. pombe* (B), *D. melanogaster* BG3-c2 cells (C), and hiPSC (D). Average proportion of reads aligning to 3' ends of unique tRNA transcripts are shown by different bar colors. Individual proportions per sample indicated by dots (n = 2).

(E) Sequence logo analysis of uniquely aligned HEK293T reads aligned to soft-clipped untemplated 3' TGIRT additions (top panel; frequency of additions in bar plot insert), and multiple sequence alignment of unique tRNA transcripts in the hg19 genome (bottom panel).

(F) Sequence logo analysis as in (E) but using *S. cerevisiae* uniquely aligned tRNA reads (top) and unique tRNA transcripts from the sacCer3 genome (bottom).

(G) Correlation plots of *S. cerevisiae* tRNA expression measured using the mim-tRNAseq method and Cy3 fluorescence from published microarray data (n=1, Pearson's correlation coefficient and p-value; Tuller et al., 2010).

(H) Alignment statistics for *S. cerevisiae* libraries prepared with TGIRT template-switching RT reaction (left), and SuperScript III RT reaction (right).

(I) Correlation plot of S. cerevisiae unique tRNA transcript abundance to tRNA gene copy number using TGIRT template-switching reaction for cDNA synthesis.

(J) Same as (I) but for libraries prepared with SuperScript III RT.

(K) Metagene analysis of scaled sequence coverage across nuclear-encoded tRNA isotypes for libraries from *S. cerevisiae* prepared with SuperScript III RT.

(L) Correlation plots of unique tRNA gene copy number and corresponding proportion of aligned reads from hiPSC tRNA using the GtRNAdb high-confidence tRNA gene set for the H. sapiens hg19 genome assembly as an alignment reference.

Figure S4

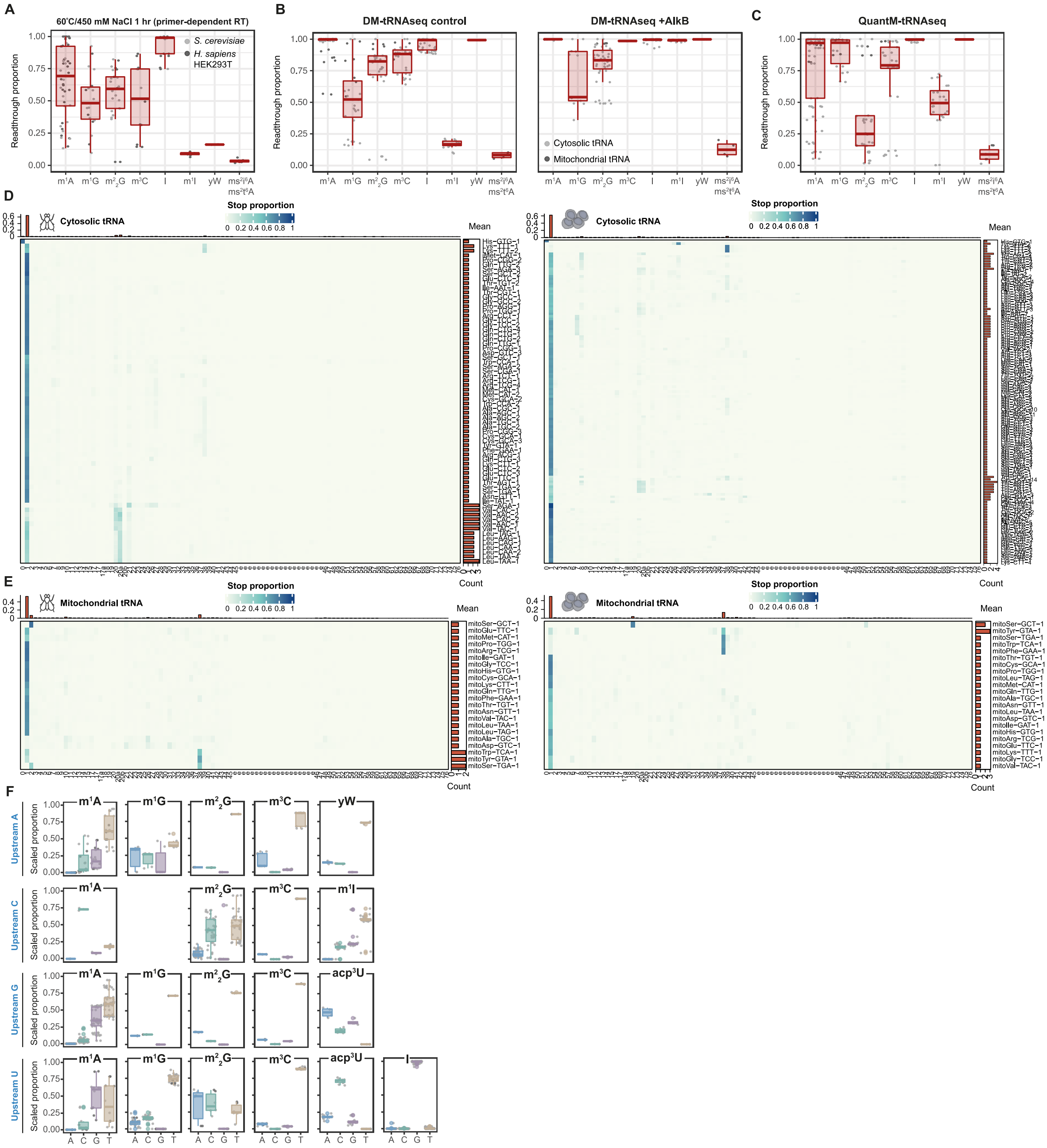


Figure S4 RT readthrough, stop, and misincorporation signature analysis at modified sites. Related to Figure 5

(A - C) RT readthrough per annotated Watson-Crick face modification using tRNA sequencing data from HEK293T cells generated with primer-depended RT by TGIRT in manufacturer-recommended conditions (A) or from publicly available DM-tRNAseq (B) without (left) and with (right) AlkB demethylation, and QuantM-tRNAseq data (C). All data analyzed with mim-tRNAseq computational pipeline. Modified sites filtered for misincorporation ≥ 10% and nucleotide coverage ≥ 2000 reads (note that acp3U was excluded due to insufficient coverage).

(D - E) Global heatmaps of average proportions of stops to RT for each unique tRNA transcript with coverage above 2000 reads in tRNA sequencing data from D. melanogaster BG3-c2 cells (left panels; n = 2) and hiPSC cells (right panels; n = 2) for cytosolic (D) and mitochondrial (E) tRNAs. Top bar graph indicates mean read proportion that stop at each position, right bar graph indicates number of sites of RT stops per transcript, where a stop is counted if more than 10% of the reads do not extend past that position. Column labels show canonical tRNA positions.

(F) Boxplots of misincorporation signatures for annotated modified sites in published human QuantM-tRNAseq data analyzed with the mim-tRNAseq computational pipeline (n = 2, center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range). Signatures stratified by upstream context (rows) and modification type (columns); proportion per nucleotide scaled to total misincorporation.

Figure S5

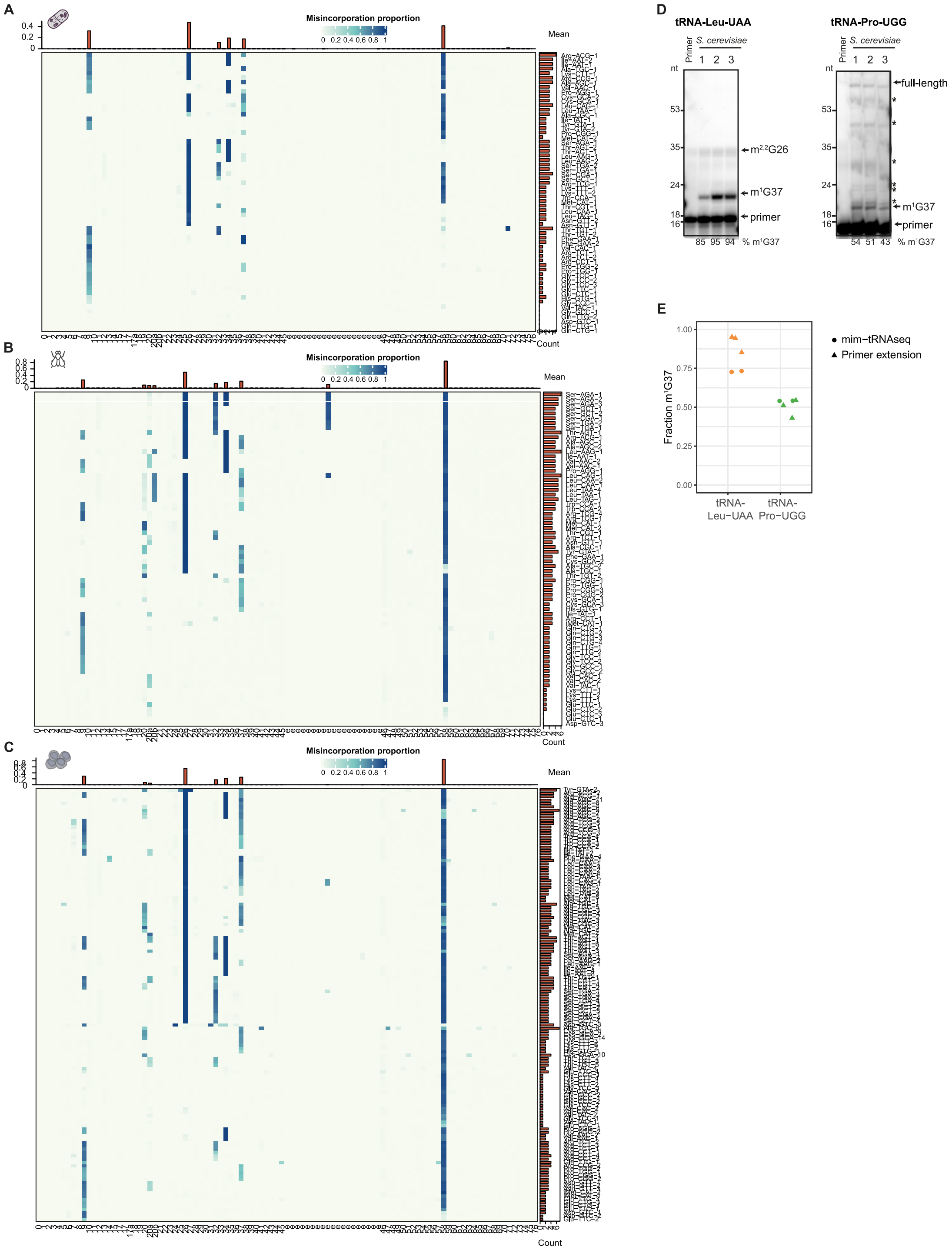


Figure S5 Modification stoichiometry estimates in tRNA from diverse eukaryotes. Related to Figure 6

(A - C) Global heatmaps of average misincorporation proportions in *S. pombe* (A), *D. melanogaster* BG3-c2 cells (B), and hiPSC (C) for each unique tRNA transcripts with coverage above 2000 reads (n = 2). Top bar graph indicates mean misincorporation at each position, right bar graph indicates number of sites of misincorporation per transcript, where a misincorporation is counted if at least 10% of the reads have a detectable mutation signature at that position. Column labels show canonical tRNA positions.

(D) Primer extension analysis of m1G37 in tRNA-Leu-UAA and tRNA-Pro-UGG with AMV RT using total RNA from *S. cerevisiae* (n = 3). % m1G37 was calculated by dividing the band intensity of the primer stop at position 37 by the sum of all stops (indicated with asterisks) and read-through to position 1.

(E) Comparison of misincorporation rates measured by mim-tRNAseq (n = 2) and RT stop fractions measured by primer extension (n = 3) at m1G37 in tRNA-Leu-UAA and tRNA-Pro-UGG from *S. cerevisiae*.

Figure S6

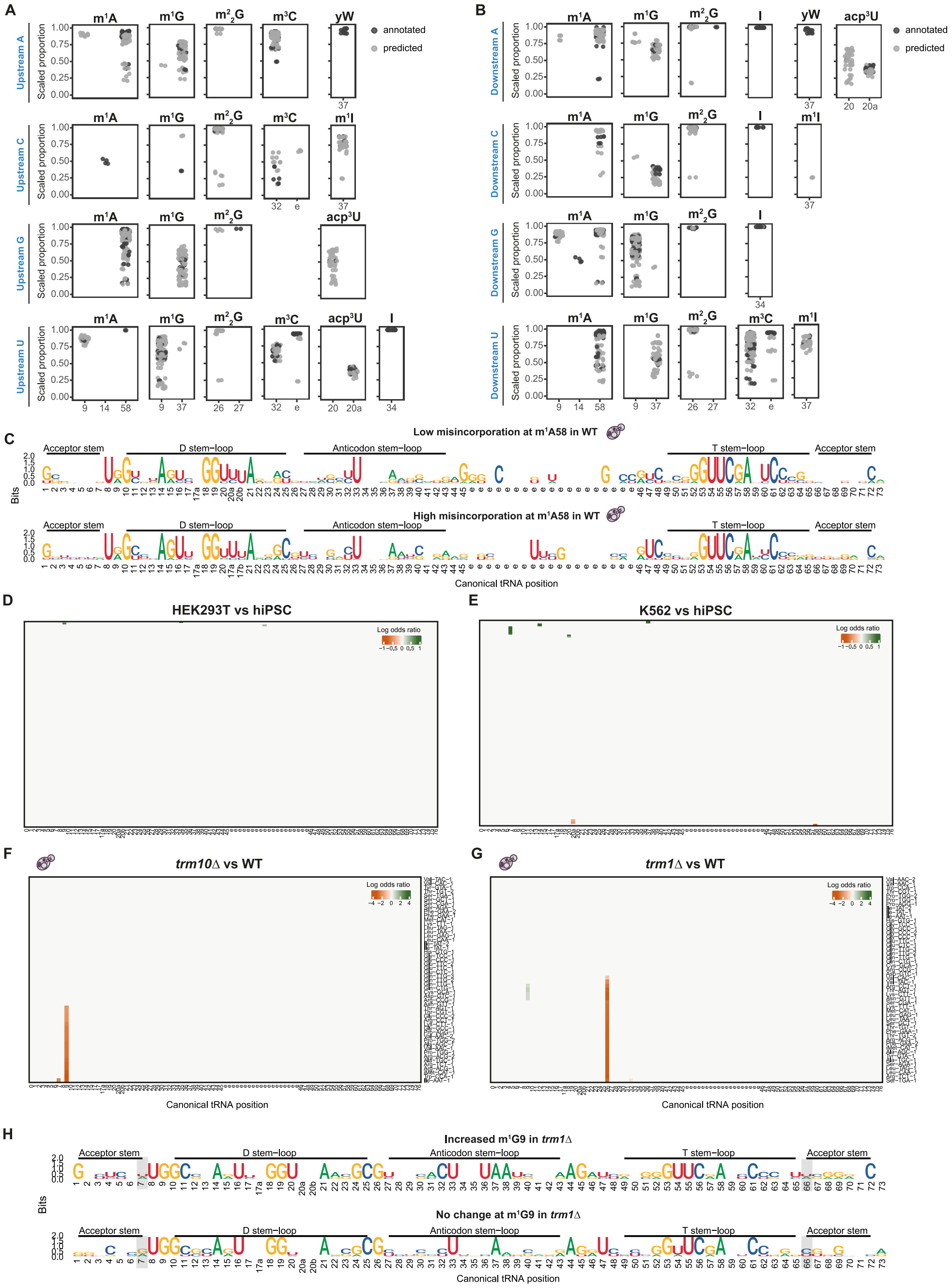


Figure S6 Determinants of tRNA modification status and stoichiometry in yeast and human cells. Related to Figure 6.
(A - B) Misincorporation proportions at annotated (dark grey) and mim-tRNAseq-predicted (light gray) sites with Watson-Crick face modifications in tRNA transcripts from the four species used in this study. Panels are separated by modification type and upstream nucleotide (A) or downstream nucleotide (B) relative to RT direction. X-axis labels indicate the canonical tRNA position of the modification.
(C) Sequence logos of aligned tRNA transcripts from budding yeast with low (< 50%; n = 18) or high (> 50%; n = 32) misincorporation at m^{1}A58 sites. The main tRNA structural domains are labeled. X-axis indicates canonical tRNA positions.
(D - E) Global heatmaps of log odd ratios (logOR) of average misincorporation in unique tRNA transcripts in HEK293T vs hiPSC (D) and K562 vs hiPSC (E) (n = 2).
(F -G) logOR heatmaps as in (C) and (D) for budding yeast *trm10Δ* vs WT (F) and *trm1Δ* vs WT (G) (n = 2). All logOR values in heatmaps (D - G) filtered for significance (Chi-square FDR-adjusted p-value ≤ 0.01) and effect size (average misincorporation log2 fold-change ≥ 0.5) for sites detected as modified by mim-tRNAseq. Column names show canonical tRNA position.
(H) Sequence logos of aligned tRNA transcripts with m$^2_2$G26, in which misincorporation at m^{1}G9-modified sites was significantly increased (upper panel; n = 4) or did not change (lower panel; n = 6) in *trm1Δ* datasets.

References

1. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-Specific Differences in Human Transfer RNA Expression. *PLoS Genet.* **2**, e221 (2006).
2. Ishimura, R. *et al.* Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science (80-.).* **345**, 455–459 (2014).
3. Kutter, C. *et al.* Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* **43**, 948–955 (2011).
4. Schmitt, B. M. *et al.* High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Res.* **24**, 1797–1807 (2014).
5. Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Publ. Gr.* **16**, 98–112 (2014).
6. Motorin, Y., Muller, S., Behm-Ansmant, I. & Branlant, C. Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods. *Methods Enzymol.* **425**, 21–53 (2007).
7. Pinkard, O., McFarland, S., Sweet, T. & Coller, J. Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nat. Commun.* **11**, 1–15 (2020).
8. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
9. Arimbasseri, A. G. *et al.* RNA Polymerase III Output Is Functionally Linked to tRNA Dimethyl-G26 Modification. *PLoS Genet.* **11**, e1005671 (2015).
10. Gogakos, T. *et al.* Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell Rep.* **20**, 1463–1475 (2017).
11. Karaca, E. *et al.* Human CLP1 mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. *Cell* **157**, 636–650 (2014).
12. Katibah, G. E. *et al.* Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12025–12030 (2014).
13. Qin, Y. *et al.* High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22**, 111–128 (2016).
14. Zheng, G. *et al.* Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* **12**, 835–837 (2015).
15. Cozen, A. E. *et al.* ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods* **12**, 879–884 (2015).
16. Clark, W. C., Evans, M. E., Dominissini, D., Zheng, G. & Pan, T. tRNA base methylation identification and quantification via high-throughput sequencing. *RNA* **22**, 1771–1784 (2016).
17. Hauenschild, R. *et al.* The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* **43**, 9950–9964 (2015).
18. Li, X. *et al.* Base-Resolution Mapping Reveals Distinct m 1 A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. *Mol. Cell* **68**, 993-1005.e9 (2017).
19. Ryvkin, P. *et al.* HAMR: high-throughput annotation of modified ribonucleotides. *RNA* **19**, 1684–1692 (2013).
20. Safra, M. *et al.* The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nat. Publ. Gr.* **551**, 251–255 (2017).
21. Ebhardt, H. A. *et al.* Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* **37**, 2461 (2009).
22. Werner, S. *et al.* Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. *Nucleic Acids Res.* **48**, 3734–3746 (2020).
23. Zhou, H. *et al.* Evolution of a reverse transcriptase to map N1-methyladenosine in human messenger RNA. *Nat. Methods 2019 1612* **16**, 1281–1288 (2019).
24. Goodenbour, J. M. & Pan, T. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res.* **34**, 6137 (2006).
25. Hoffmann, A. *et al.* Accurate Mapping of tRNA Reads. *Bioinformatics* **366**, 1 (2017).
26. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).
27. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195 (2018).
28. Xu, H., Yao, J., Wu, D. C. & Lambowitz, A. M. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci. Rep.* **9**,

1–17 (2019).
29. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res.* **40**, (2012).
30. Heyer, E. E., Ozadam, H., Ricci, E. P., Cenik, C. & Moore, M. J. An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. *Nucleic Acids Res.* **43**, e2–e2 (2015).
31. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
32. Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. (2012) doi:10.1038/nmeth.1814.
33. Chen, D. & Patton, J. T. Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. *Biotechniques* **30**, 574–582 (2001).
34. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 1–10 (2009).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
36. Jühling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **37**, D159–D162 (2009).
37. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkx1030.
38. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
39. Lin, Y. C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun. 2014 51* **5**, 1–12 (2014).
40. Dittmar, K. A., Mobley, E. M., Radek, A. J. & Pan, T. Exploring the Regulation of tRNA Distribution on the Genomic Scale. *J. Mol. Biol.* **337**, 31–47 (2004).
41. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
42. Harismendy, O. *et al.* Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J.* **22**, 4738–4747 (2003).
43. Roberts, D. N., Stewart, A. J., Huff, J. T. & Cairns, B. R. The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 14695–14700 (2003).
44. Jacob, D. *et al.* Absolute Quantification of Noncoding RNA by Microscale Thermophoresis. *Angew. Chem. Int. Ed. Engl.* **58**, 9565–9569 (2019).
45. Torres, A. G., Reina, O., Attolini, C. S. O. & De Pouplana, L. R. Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8451–8456 (2019).
46. Evans, M. E., Clark, W. C., Zheng, G. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic Acids Res.* **45**, e133–e133 (2017).
47. Han, L., Guy, M. P., Kon, Y. & Phizicky, E. M. Lack of 2&apos;-O-methylation in the tRNA anticodon loop of two phylogenetically distant yeast species activates the general amino acid control pathway. *PLoS Genet.* **14**, e1007288 (2018).
48. Sas-Chen, A. & Schwartz, S. Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods* (2018) doi:10.1016/j.ymeth.2018.10.011.
49. Guy, M. P. *et al.* Yeast Trm7 interacts with distinct proteins for critical modifications of the tRNAPhe anticodon loop. *RNA* **18**, 1921–1933 (2012).
50. Jackman, J. E., Montange, R. K., Malik, H. S. & Phizicky, E. M. Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *RNA* **9**, 574–585 (2003).
51. Ellis, S. R., Morales, M. J., Li, J. M., Hopper, A. K. & Martin, N. C. Isolation and characterization of the TRM1 locus, a gene essential for the N2,N2-dimethylguanosine modification of both mitochondrial and cytoplasmic tRNA in Saccharomyces cerevisiae. *J. Biol. Chem.* **261**, 9703–9709 (1986).
52. Gerber, A. P. & Keller, W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science (80-.).* **286**, 1146–1149 (1999).
53. Anderson, J. *et al.* The essential Gcd10p–Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA. *Genes Dev.* **12**, 3650 (1998).
54. Gamper, H. B., Masuda, I., Frenkel-Morgenstern, M. & Hou, Y. M. Maintenance of protein

synthesis reading frame by EF-P and m(1)G37-tRNA. *Nat. Commun.* **6**, (2015).

55. Maehigashi, T., Dunkle, J. A., Miles, S. J. & Dunham, C. M. Structural insights into +1 frameshifting promoted by expanded or modification-deficient anticodon stem loops. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12740–12745 (2014).
56. Pernod, K. *et al.* The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic Acids Res.* **48**, 6170–6183 (2020).
57. Masuda, I. *et al.* tRNA Methylation Is a Global Determinant of Bacterial Multi-drug Resistance. *Cell Syst.* **8**, 302-314.e8 (2019).
58. Swinehart, W. E., Henderson, J. C. & Jackman, J. E. Unexpected expansion of tRNA substrate recognition by the yeast m1G9 methyltransferase Trm10. *RNA* **19**, 1137–1146 (2013).
59. Phizicky, E. M. & Alfonzo, J. D. Do all modifications benefit all tRNAs? (2009) doi:10.1016/j.febslet.2009.11.049.
60. Steinberg, S. & Cedergren, R. A correlation between N2-dimethylguanosine presence and alternate tRNA conformers. *RNA* **1**, 886 (1995).
61. Motorin, Y. & Helm, M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. *Genes 2019, Vol. 10, Page 35* **10**, 35 (2019).
62. Pinkard, O., McFarland, S., Sweet, T. & Coller, J. Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nat. Commun.* **11**, (2020).
63. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
64. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
66. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
67. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
68. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
70. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
71. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
72. Warner, G. J. *et al.* Inhibition of selenoprotein synthesis by selenocysteine tRNA[Ser]Sec lacking isopentenyladenosine. *J. Biol. Chem.* **275**, 28110–28119 (2000).
73. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
74. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
75. Torres, A. G. *et al.* Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.* **43**, 5145–5157 (2015).
76. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

CHAPTER 3

Experimental and computational workflow for the analysis of tRNA pools from eukaryotic cells by mim-tRNAseq

This chapter was published in STAR Protocols following peer review, and is available through open-access. The article was reformatted and inserted here.

Behrens, A. & Nedialkova, D. D. Experimental and computational workflow for the analysis of tRNA pools from eukaryotic cells by mim-tRNAseq. *STAR Protoc.* **3**, 101579 (2022). https://doi.org/10.1016/j.xpro.2022.101579

Overview

Since its initial release, mim-tRNAseq underwent significant and continual development, upgrades, and functionality additions. Some of this development was prompted from the strong community-based input we received, particularly via the GitHub repository page for the package. In addition, we received many questions and requests about mim-tRNAseq installation, usage, and troubleshooting.

This encouraged the development of a detailed, peer-reviewed protocol for both the mim-tRNAseq library generation steps, and the usage of the `mimseq` computational package. In this manuscript, step-by-step details can be found for the complete process for sequencing and analysis of eukaryotic tRNA pools, with example data analysis pertaining to human cell-derived data from the original publication. Furthermore, additional attention is paid to detailing critical steps and possible alternatives throughout the process. The most common errors and inefficiencies, with potential solutions are discussed in the troubleshooting guide.

We utilize this opportunity to describe the major updates to mim-tRNAseq, specifically the computational package in its current form (version 1 and above), and explain the significant algorithmic changes to cluster deconvolution and its impact on the improved accuracy and sensitivity of mim-tRNAseq. Expected outcomes are discussed, with focus on the most crucial outputs and plots from `mimseq` to enable quality control at both the library generation and computational analysis steps of the protocol.

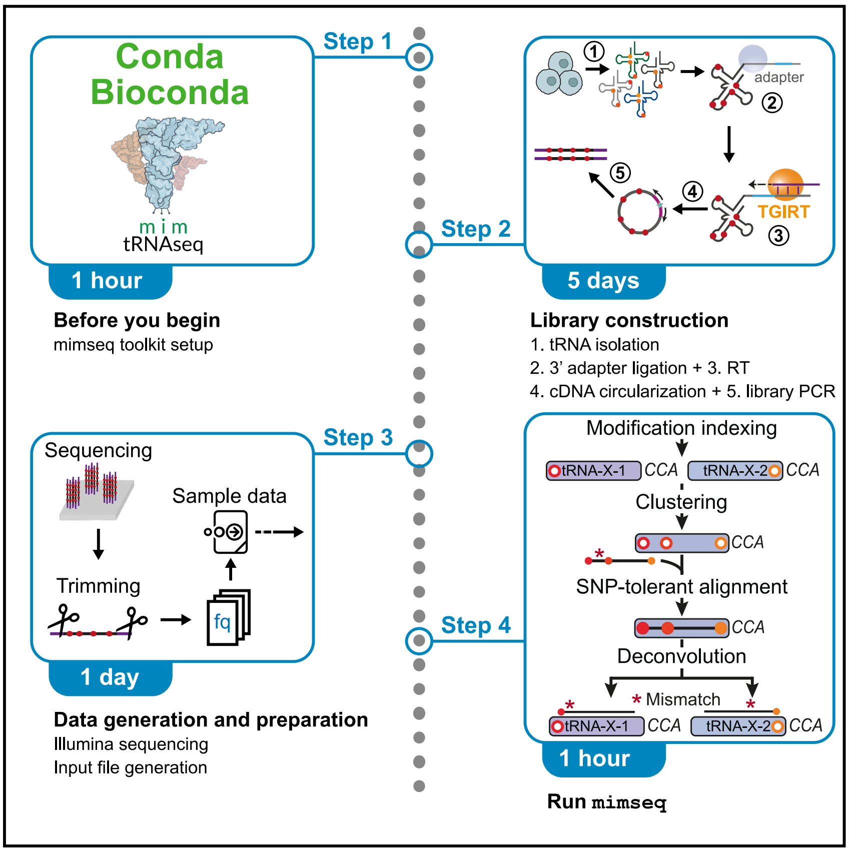
In the spirit of open-source, community-driven development of robust methods and computational packages, we envision that the mim-tRNAseq protocol will further aid its use and improvement over time. Moreover, protocols such as this are crucial for combatting the ongoing reproducibility crisis in the biological sciences.

Contribution: All updates to the mim-tRNAseq computational package, including algorithms, data processing, and visualisation were jointly conceptualised by Danny Nedialkova and myself and implemented by me. Both Danny Nedialkova and I shared writing, review and editing of the manuscript. More specifically, I wrote the original draft of all computational steps of the protocol, expected outcomes, details regarding updates to deconvolution, and problems and solutions pertaining to computational analyses. All data visualisations and figures were prepared by Danny Nedialkova and me.

Summary

Quantifying tRNAs is crucial for understanding how they regulate mRNA translation, but is hampered by their extensive sequence similarity and premature termination of reverse transcription at multiple modified nucleotides. Here, we describe the use of modification-induced misincorporation tRNA sequencing (mim-tRNAseq), which overcomes these limitations with optimized library construction and a comprehensive toolkit for data analysis and visualization. We outline algorithm improvements that enhance the efficiency and accuracy of read alignment and provide details on data analysis outputs using example datasets.

For complete details on the use and execution of this protocol, please refer to Behrens et al. (2021)[1].



Before you begin

mim-tRNAseq consists of an optimized protocol for cDNA library construction from eukaryotic tRNA pools and a suite of computational tools for the analysis, quantitation, and visualization of the resulting high-throughput sequencing data. It is based on the efficient synthesis of full-length cDNAs from tRNA transcripts with a thermostable group II intron reverse transcriptase (TGIRT) in reaction conditions that enable the readthrough of nearly all nucleotide modifications. To account for extensive sequence similarity among eukaryotic tRNA genes and the nucleotide misincorporations in cDNA introduced at modified tRNA sites by TGIRT, we developed the mim-tRNAseq computational toolkit, which includes multiple novel algorithms that are specifically tailored to the analysis of tRNA-derived sequencing data. Details of the development and optimization of mim-tRNAseq can be found in the original publication[1] and in the package documentation (https://mim-trnaseq.readthedocs.io/en/latest).

Here, we describe in detail the steps required to implement the library generation and data analysis workflow, starting with samples from cultured human cell lines. We also present an updated v1.1 of the mim-tRNAseq toolkit, which contains improvements to various aspects of the computational workflow that increase the accuracy and efficiency of data analysis.

Note: For all computational steps throughout the protocol, commands to be entered on the command line are given in text boxes (starting with ">") along with comments describing the function of the command (starting with "#").

Preparation of the mim-tRNAseq toolkit environment

Timing: 1 h

Before use, the computational toolkit needs to be installed. The package (named `mimseq`) is available on bioconda (https://bioconda.github.io/recipes/mimseq/README.html), GitHub (https://github.com/nedialkova-lab/mim-tRNAseq), and archived on Zenodo (https://doi.org/10.5281/zenodo.6694873).

Note: We strongly recommend using the conda package and environment manager to install the package with all of its dependencies from the bioconda channel.

Note: The GitHub repository is a useful source of community discussion, issues and solutions, and requested upgrades and functionality (see Issues tab).

Note: The toolkit has been tested extensively on multiple Linux-based servers and computing clusters. We recommend a Linux server with at least 32GB of memory and 8 CPU cores. These requirements will change depending on the number of samples and size of the datasets and tRNA reference, as will the processing time required by `mimseq`. Although running `mimseq` on a personal computer is theoretically possible, and multi-processing is customizable before runtime, we have not tested this.

1. Retrieve and install miniconda.

```
# Download miniconda
> wget https://repo.anaconda.com/miniconda/Miniconda3-py39_4.10.3-Linux-x86_64.sh

# Run the installation script
> bash Miniconda3-py39_4.10.3-Linux-x86_64.sh
```

After running the second command, follow the on-screen prompts and accept the defaults (unless otherwise required).

Following installation, restart the shell session by reconnecting to the remote server in order to activate the conda installation.

2. Initialize and configure the `mimseq` environment.

```
# Create the mimseq environment with the correct Python version
> conda create -n mimseq python=3.7

# Activate the environment
> conda activate mimseq

# Configure environment channels
> conda config --add channels conda-forge
```

3. Install mamba.

```
> conda install -c conda-forge mamba
```

4. Use mamba to install `mimseq` and all dependencies.

```
> mamba install -c bioconda mimseq
```

5. usearch is not available on conda and needs to be installed manually.

```
# Download and unzip usearch
> wget https://drive5.com/downloads/usearch10.0.240_i86linux32.gz
> gunzip usearch10.0.240_i86linux32.gz

# Make binary executable and rename
> chmod +x usearch10.0.240_i86linux32
> mv usearch10.0.240_i86linux32 usearch

# If root access is available then copy the binary into an accessible PATH
location, for example:
> cp usearch /usr/local/bin

# However, if this is not possible, add the path to the usearch binary to
your PATH variable
> export PATH=$PATH:full/path/to/usearch

# where “full/path/to/usearch” should be replaced by the path to the
usearch binary that was just unzipped and modified.
```

Note: Exporting the usearch binary to a user-specific path as in the last command of Step 5 is temporary and needs to be done for every new terminal session for that user. It is recommended to add the command to the user’s .bashrc file (or similar), or preferably to have your system administrator add usearch to a global PATH location (see above, where usearch is copied to /usr/local/bin).

6. Test your `mimseq` installation by determining the version and printing the help. Output should look like the screenshots in **Figure 1.** Please ensure you have installed `mimseq` v1.1 or newer.

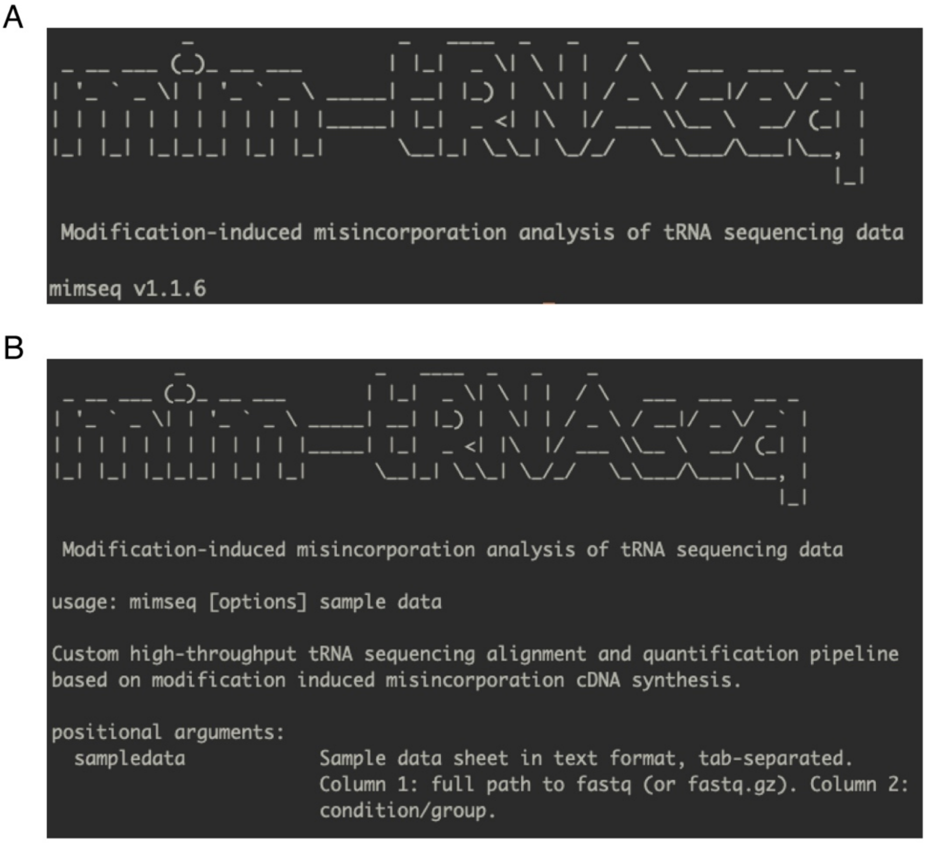
```
    > mimseq --version
    > mimseq --help
```

Preparation of cell or tissue samples for total RNA isolation

Timing: days (depending on the organism, developmental stage, and cell or tissue type)

7. Culture cells in an appropriate medium or collect tissue samples according to the purpose of the experiment.

CRITICAL: Extra care should be taken when working with animals, as post-mortem RNA degradation can occur rapidly in some tissue types (Richter et al., 2022). Dissection should be performed as quickly as possible, and tissue samples should be snap-frozen in liquid nitrogen and stored at −80°C.



```
           _                  _   ____  _   _    _
 _ __ ___ (_)_ __ ___        | |_|  _ \| \ | |  / \   ___  ___  __ _
| '_ ` _ \| | '_ ` _ \ _____| __| |_) |  \| | / _ \ / __|/ _ \/ _` |
| | | | | | | | | | | |_____| |_|  _ <| |\  |/ ___ \\__ \  __/ (_| |
|_| |_| |_|_|_| |_| |_|      \__|_| \_\_| \_/_/   \_\___/\___|\__, |
                                                                |_|

 Modification-induced misincorporation analysis of tRNA sequencing data

mimseq v1.1.6
```

```
           _                  _   ____  _   _    _
 _ __ ___ (_)_ __ ___        | |_|  _ \| \ | |  / \   ___  ___  __ _
| '_ ` _ \| | '_ ` _ \ _____| __| |_) |  \| | / _ \ / __|/ _ \/ _` |
| | | | | | | | | | | |_____| |_|  _ <| |\  |/ ___ \\__ \  __/ (_| |
|_| |_| |_|_|_| |_| |_|      \__|_| \_\_| \_/_/   \_\___/\___|\__, |
                                                                |_|

 Modification-induced misincorporation analysis of tRNA sequencing data

usage: mimseq [options] sample data

Custom high-throughput tRNA sequencing alignment and quantification pipeline
based on modification induced misincorporation cDNA synthesis.

positional arguments:
  sampledata            Sample data sheet in text format, tab-separated.
                        Column 1: full path to fastq (or fastq.gz). Column 2:
                        condition/group.
```

Figure 1 Testing mimseq environment and installation.
(A) Running mimseq --version displays the mimseq logo and version number, which should be higher than v1.1.
(B) mimseq --help displays the help documentation on mimseq parameters.

Oligonucleotide ordering

Timing: 1 h – days (depending on supplier)

8. Order spike-in RNA, 3′ adapter, reverse transcription (RT), and library construction oligonucleotides (*Table S1*).

CRITICAL: Since the library construction protocol is susceptible to the presence of ribonucleases, the spike-in, 3’ adapter, and RT oligonucleotides should be of RNase-free HPLC purity grade.

Note: The *E. coli* tRNA-Lys-UUU tRNA is a suitable spike-in for eukaryotic samples as it differs sufficiently in sequence from eukaryotic tRNAs to avoid read misalignment. Other tRNA sequences can be substituted, provided that care is taken to ensure that no cross-mapping of sample reads to the spike-in reference will occur.

Note: The custom 3’ adapter oligonucleotides should contain a 5’ phosphate, which is necessary for 5’ adenylation, and should be blocked with 3’ dideoxycytidine (or an alternative blocking group) to prevent self-ligation. The invariable 5’ GAT sequence of each adapter ensures an identical sequence context at the tRNA-adapter ligation junction (all mature tRNAs end with a single-stranded 3’ CCA, or a mixture of 3’ CCA and 3’ CC depending on their charging status after periodate oxidation and β-elimination). A 5-nt barcode (I1–I8) is present at positions 4–8. This enables the pooling of up to eight samples before reverse transcription, which reduces sample input requirements and reagent costs.

CRITICAL: The eight barcoded 3’ adapter oligonucleotides in Table S1 were specifically designed to minimize the potential for secondary structure formation, which interferes with ligation[2]. The use of other 3’ adapters may result in decreased ligation efficiency to mature tRNAs and should be carefully evaluated first.

Note: The RT oligonucleotide contains a 5’ phosphate necessary for cDNA circularization by CircLigase followed by an RN dinucleotide, which mitigates potential sequence biases in this reaction[3,4].

CRITICAL: Reverse primers for PCR library construction should contain a unique 6-nt index sequence (denoted with NNNNNN) to discriminate between different libraries loaded on the same flow cell. Indexes should differ by at least 2 nucleotides to avoid demultiplexing errors.

CRITICAL: All primers for library construction PCR should contain a phosphorothioate bond between the last two nucleotides at the 3’ end to prevent their degradation by the KAPA HiFi Polymerase, which has a strong 3’-5’ exonuclease activity. The use of unmodified oligonucleotides results in poor cDNA amplification in step 131.

3’ adapter preadenylation

Timing: 2 h

9. Set up a 3’ adapter preadenylation reaction as in McGlincy and Ingolia[4].

3’ adapter preadenylation reaction

Reagent	Final concentration	Amount
10 × 5′ DNA Adenylation Reaction Buffer	1 ×	2 μL
ATP (1 mM)	0.1 mM	2 μL

Mth RNA Ligase (50 µM)	5 µM	2 µL
Barcoded 3’ adapter (one of I1-I8) (100 µM)	6 µM	1.2 µL
RNase-free water	n/a	12.8 µL
Total	**n/a**	20 L

10. Incubate the reaction at 65°C for 1 h, followed by an incubation at 85°C for 5 min.
11. Add 30 µL RNase-free water and purify the pre-adenylated adapter with the Zymo Oligo Clean & Concentrator kit according to the manufacturer’s protocol. Elute in 6 µL RNase-free water.
12. Preadenylated adapters can be stored at −20°C for several months.

Note: To prepare a larger batch of a pre-adenylated 3’ adapter, set up several reactions and pool them prior to step 11.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
TRIzol Reagent	Thermo Fisher Scientific	Cat#15596026
1-bromo-3-chloropropane	Sigma Aldrich	Cat#B9673
Formamide, deionized	Sigma Aldrich	Cat# F9037
Acrylamide/Bis 19:1, 40% (w/v) solution	Thermo Fisher Scientific	Cat#AM9022
Urea	Sigma Aldrich	Cat#U1250
UltraPure TBE buffer (10 ×)	Thermo Fisher Scientific	Cat#15581044
1,4-Dithiothreitol	Thermo Fisher Scientific	Cat#R0862
Sodium dodecyl sulphate	Sigma Aldrich	Cat#74256
Sodium acetate	Sigma Aldrich	Cat#S7545
Magnesium chloride	Sigma Aldrich	Cat#M2670
Sodium chloride	Sigma Aldrich	Cat# S7653
Potassoum chloride	Sigma Aldrich	Cat#P9541
EDTA	Sigma Aldrich	Cat#E5134
Sodium periodate	Sigma Aldrich	Cat#311448
Sodium tetraborate decahydrate	Sigma Aldrich	Cat#B3545
Ethanol	Sigma Aldrich	Cat#51976
Glycogen	Thermo Fisher	Cat#AM9510
Low Range ssRNA ladder	New England Biolabs	Cat#N0364S
T4 Polynucleotide Kinase	New England Biolabs	Cat#M0201L
T4 RNA ligase 2, truncated KQ (200U/µL)	New England Biolabs	Cat#M0373L

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SUPERase•In RNase Inhibitor (20U/µL)	Thermo Fisher Scientific	Cat#AM2694
TGIRT	InGex	Cat#TGIRT50
CircLigase ssDNA ligase	Lucigen	Cat#CL4115K
KAPA HiFi DNA Polymerase (1U/µL)	Roche	Cat#KK2102
Critical commercial assays		
5´ DNA Adenylation Kit	New England Biolabs	Cat#E2610L
Micro Bio-Spin 30 Columns, RNase-free	Bio-Rad	Cat#326251
Costar Spin-X® Centrifuge Tube Filters, 0.22 µm	Corning Life Sciences	Cat#8160
SYBR Gold Nucleic Acid Gel Stain	Thermo Fischer Scientific	Cat# S11494
DNA Clean & Concentrator-5 kit	Zymo Research	Cat#D4013
Oligo Clean & Concentrator kit	Zymo Research	Cat#D4060
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific	Cat#Q32854
Deposited data		
mim-tRNAseq data from HEK293T and K562 cells	Behrens et al.[1]	GEO: GSE152621
Oligonucleotides		
Oligonucleotides, RNA sequences and primers for library construction	Behrens et al.[1], and this paper; see Table S1	www.doi.org/10.17632/vy8z394gfh.1
Software and algorithms		
mim-tRNAseq v1.1.6	Behrens et al.[1]	GitHub: https://github.com/nedialkova-lab/mim-tRNAseq Zenodo: https://doi.org/10.5281/zenodo.6694873
cutadapt v3.5	Martin[5]	https://cutadapt.readthedocs.io/en/stable/
GSNAP v2019-02-26	Wu and Nacu[6]	http://research-pub.gene.com/gmap/
Samtools v1.14	Li et al.[7]	http://samtools.sourceforge.net/
Bedtools v2.30.0	Quinlan and Hall[8]	https://bedtools.readthedocs.io/en/late
BLAST v2.10.1+	Camacho et al.[9]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
Infernal v1.1.4	Nawrocki and Eddy[10]	http://eddylab.org/infernal/
usearch v10.0.240_i86linux32	Edgar[11]	https://www.drive5.com/usearch/
R/DESeq2 v1.34.0	Love et al.[12]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
R/ComplexHeatmap v2.10.0	Gu et al.[13]	https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html
Python/Biopython v1.79	Cock et al.[14]	https://biopython.org/

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Other		
Low Range ssRNA ladder	New England Biolabs	Cat#N0364S
pBR322 DNA-MspI Digest	New England Biolabs	Cat#N3032S
Disposable homogenizer pestles for 1.5 ml tubes	VWR	Cat#47747-358

Materials and equipment

Denaturing urea polyacrylamide gel, 10%

Reagent	Final concentration	Amount
TBE (10 ×)	1 ×	1 mL
Urea	7 M	4.2 g
Acrylamide/Bis 19:1, 40% (w/v) solution	10%	2.5 mL
RNase-free water	N/A	make up to 10 mL
APS (10%)	0.05%	50 µL
TEMED	0.5%	50 µL
Total	**N/A**	**10 mL**

Prepare freshly before use

2 × RNA loading Buffer

Reagent	Final concentration	Amount
TBE (10 ×)	1 ×	1 mL
Formamide	85%	8.5 mL
SDS (10%)	0.05%	50 µL
EDTA (0.5 M)	0.5 mM	100 µl
RNase-free water	n/a	350 µL
Total	**n/a**	**10 mL**

Store at ambient temperature for up to three months or at -20ºC for up to two years.

RNA gel extraction buffer

Reagent	Final concentration	Amount
Sodium acetate, pH=4.5 (3 M)	300 mM	1 mL
SDS (10%)	0.25%	250 µL
EDTA (0.5 M)	1 mM	200 µl
RNase-free water	n/a	9.55 mL
Total	**n/a**	**10 mL**

Aliquot in small batches to minimize RNase contamination risk. Store at ambient temperature for up to a year.

TGIRT reaction buffer

Reagent	Final concentration	Amount
Tris-HCl, pH=8.3 (1 M)	50 mM	50 µL
KCl (3 M)	75 mM	25 µL
$MgCl_2$ (1 M)	3 mM	3 µL
RNase-free water	n/a	922 µL
Total	**n/a**	**1 mL**

Aliquot and store at -20°C for up to two years.

Sodium periodate solution

Reagent	Final concentration	Amount
Sodium periodate	500 mM	0.106 g
RNase-free water	n/a	1 mL
Total	**n/a**	**1 mL**

Prepare freshly before use.

Sodium tetraborate solution

Reagent	Final concentration	Amount
Sodium tetraborate decahydrate, pH=9.5	100 mM	0.38 g
RNase-free water	n/a	10 mL
Total	**n/a**	**10 mL**

Prepare freshly right before use. Dissolve in 7 mL water, adjust pH to 9.5 with 5M NaOH, and make up to 10 mL with water.

CRITICAL: To minimize the risk of RNase contamination, wear gloves during all experiments and change them frequently. Freshly purified Milli-Q water (Millipore) is sufficiently free of RNases for all steps of the protocol. Avoid using DEPC as it can inhibit some enzymatic reactions. Use RNase-free plasticware, low binding microtubes, and low-retention long filter pipette tips in all steps. Do not autoclave plasticware or buffers. RNases can be removed from glassware used for storing larger volumes of buffers (e.g., 1× TBE) by baking at 180°C–220°C for several hours.

CRITICAL: Sodium periodate, sodium tetraborate, acrylamide in solution, TRIzol, 1-bromo-3-chloropropane, formamide, DTT, and SDS are toxic and harmful. Wear appropriate personal protective equipment (goggles, gloves, lab coat) and follow institutional guidelines for handling and disposal.

Alternatives: TRIzol and acrylamide/bis (19:1, 40%) can be replaced with equivalent reagents from other commercial sources.

Alternatives: Handcast denaturing gels can be substituted with commercially available pre-cast ones (e.g., 10% Novex™ TBE-Urea Gels from Thermo Fisher Scientific).

Step-by-step method details

Total RNA isolation under mild acidic conditions

Timing: 2–3 h

This section describes total RNA isolation from cultured mammalian cells under conditions that preserve tRNA charging.

CRITICAL: Working at low pH and maintaining samples at <4°C is essential for preserving tRNA charging. Avoid processing more than 8 samples at a time to decrease handling time.

Alternatives: If quantitation of tRNA charging is not of interest for the research question, total RNA can be extracted with the standard TRIzol protocol or other protocols or commercial kits that efficiently recover RNAs of <200 nt in length. In this case, proceed directly to step 46.

Note: RNA yield varies substantially depending on sample source, organism and cell type, and growth conditions. Commonly used human cell lines typically yield 15–30 µg of total RNA per 1 × 10^6 cultured human cells with TRIzol-based protocols.

Note: In our hands, RNA isolation with column-based commercial kits can lead to significant sample loss and poor recovery of small RNAs. We recommend avoiding the use of such kits, particularly when starting with low cell numbers.

1. Carefully aspirate all culture medium.
2. Add 1 mL of TRIzol Reagent per 1 × 10^5–1 × 10^7 cells directly to the culture dish.
3. Pipet the lysate up and down several times to homogenize and transfer to a 1.5 mL microfuge tube.
4. Snap-freeze samples on dry ice or in liquid nitrogen.

Pause point: Samples can be stored at −80°C for several months.

5. Thaw samples at room temperature (20°C–26°C).
6. Incubate at room temperature for 5 min to dissociate nucleoprotein complexes.
7. Add 0.2 mL of 1-bromo-3-chloropropane (BCP) per 1 mL of TRIzol Reagent used for lysis, then securely cap the tube.
8. Vortex briefly and incubate for 2 min at room temperature (20°C–26°C).
9. Centrifuge the sample for 15 min at 12,000 × g at 4°C in a pre-chilled centrifuge.
10. Place tubes on ice. Transfer the aqueous phase containing the RNA to a new 1.5 mL microfuge tube.
11. Add 0.5 mL of BCP per 1 mL of TRIzol Reagent used for lysis, then securely cap the tube. Vortex briefly.
12. Centrifuge the sample for 5 min at 12,000 × g at 4°C.
13. Place samples on ice. Transfer the aqueous phase containing the RNA to a new 2 mL microfuge tube.
14. Precipitate RNA by adding the following amounts per 1 mL of TRIzol Reagent used for lysis:
    a. 25 µg glycogen.
    b. 100 µL 3 M sodium acetate (pH=4.5).
    c. 1.25 mL of ice-cold 100% ethanol.
15. Vortex well and incubate at −20°C for at least 30 min.

Pause point: Samples can be stored at −20°C indefinitely.

CRITICAL: The use of ethanol instead of isopropanol for RNA precipitation improves the recovery of RNAs <200 nt in length.

16. Centrifuge for 20 min at 12,000 × g at 4°C.
17. Carefully remove the supernatant with a 1-mL pipette tip.
18. Add 1 mL of ice-cold 80% ethanol containing 50 mM sodium acetate, pH=4.5.
19. Vortex the sample briefly, then centrifuge for 5 min at 7,500 × g at 4°C.
20. Carefully remove the supernatant with a 1-mL pipette tip.
21. Briefly spin down and remove all remaining liquid with a 10-µL pipette tip.
22. Air-dry the pellet for 2–3 min.
23. Resuspend in 30 µL 50 mM sodium acetate (pH=4.5), 1 mM EDTA per 1 mL of TRIzol Reagent used for lysis.
24. Measure RNA concentration on a NanoDrop or an equivalent UV spectrophotometer.

Pause point: Aliquot and store at −80°C for up to six months. Avoid repeated freeze-thaw cycles.

Note: We recommend assessing the integrity of the extracted total RNA on a TapeStation system or by electrophoresis on a 10% denaturing polyacrylamide gel (steps 56–65). The tRNA cluster should be clearly visible at 60–100 nt and there shouldn't be any smearing that is indicative of RNA degradation.

Periodate oxidation and β-elimination

Timing: 4–6 h

The 3' ends of tRNAs that carry an amino acid are protected from periodate oxidation. When followed by β-elimination, this treatment leads to the removal of the 3′ nucleotide in uncharged tRNAs. The proportion of transcripts that end with 3′-CCA versus those that end with 3′-CC after periodate oxidation and β-elimination can thus be used to quantify the fraction of aminoacylated tRNA molecules[15,16].

CRITICAL: This part of the protocol requires total RNA isolated under mild acidic conditions as input.

Note: The starting concentration of RNA in the oxidation reaction should be 0.25–1 µg/µL.

25. Assemble the oxidation reaction in 1.5-mL microfuge tubes.

Oxidation reaction

Reagent	Final concentration	Amount
10 µg total RNA	0.5 µg/µL	X µL
10 mM sodium acetate, pH=4.5	N/A	20 - X µL
500 mM sodium periodate (freshly prepared)	50 mM	2.2 µL
Total	**N/A**	**22 µL**

26. Incubate for 30 min at 22°C in a Thermoblock.

Note: During the incubation time, prepare one Micro Bio Spin-30 column per sample according to the manufacturer's protocol.

27. Stop reaction by adding 2.4 µL 1 M glucose. Incubate for 5 min at 22°C in a Thermoblock.

28. Clean up reactions with Micro Bio Spin-30 columns according to the manufacturer's protocol. The recovered volume will be ~25 µL.
29. To each tube, add:
    a. 65 µL RNase-free water.
    b. 10 µL 3 M sodium acetate (pH=4.5).
    c. 25 µg glycogen.
    d. 300 µL ice-cold 100% ethanol.
30. Vortex well and incubate at −20°C for at least 30 min.

Pause point: Samples can be stored at –20°C indefinitely.

31. Precipitate RNA by centrifuging for 30 min at 16,000 × g at 4°C.
32. Carefully remove the supernatant with a 1-mL filter pipette tip. Briefly spin down and remove all residual liquid with a 10-µL filter pipette tip.
33. Resuspend pellets in 20 µL RNase-free water.
34. To perform β-elimination, mix the RNA with 30 µL freshly prepared 100 mM sodium tetraborate (pH=9.5).
35. Incubate at 45°C for 90 min in a Thermoblock.

Note: During the incubation time, prepare one Micro Bio Spin-30 column per sample according to the manufacturer's protocol.

36. Clean up reactions with Micro Bio Spin-30 columns according to the manufacturer's protocol. The recovered volume will be ~50 µL.
37. To each tube, add:
    a. 35 µL RNase-free water.
    b. 10 µL 3 M sodium acetate (pH=4.5).
    c. 25 µg glycogen.
    d. 300 µL ice-cold 100% ethanol.
38. Vortex well and incubate at −20°C for at least 30 min.

Pause point: Samples can be stored at –20°C indefinitely.

39. Precipitate RNA by centrifuging for 30 min at 16,000 × g at 4°C.
40. Carefully remove the supernatant with a 1-mL filter pipette tip. Briefly spin down and remove all residual liquid with a 10-µL filter pipette tip.

41. Air-dry pellet for 2–3 min.
42. Resuspend pellets in 15 µL RNase-free water.
43. Measure RNA concentration on a NanoDrop or an equivalent spectrophotometer.

CRITICAL: The column purification and RNA precipitation steps are necessary to prevent carry-over of sodium periodate and sodium tetraborate. These reagents can inhibit the enzymatic removal of the RNA 3′ phosphate resulting from β-elimination with T4 polynucleotide kinase (PNK) in steps 49–56. This will lead to an underrepresentation of reads from uncharged tRNA in the final sequencing library and thus an overestimation of charged tRNA fractions.

Spike-in addition and tRNA deacylation

Timing: 1 h

The spike-in RNA added in this step is intended to be used as an internal control for 3′-CCA quantitation. This is achieved by adding two synthetic tRNAs that differ by a single nucleotide at the 3' end (*E.coli* tRNA-Lys-UUU-CCA and tRNA-Lys-UUU-CC; for sequences, see Table S1) in a 3:1 ratio to total RNA.

Alternatives: Spike-in addition can be omitted if measurements of charging fractions are not performed.

Note: Following library construction and analysis by `mimseq`, the relative quantities of each spike-in RNA can be checked in the CCA plots to ensure the expected 3:1 ratio is recovered.

Note: We have successfully constructed tRNA sequencing libraries starting with 0.5 µg of total RNA. If input amounts are not limiting, we recommend starting with 2.5–5 µg of total RNA.

44. Dilute 2.5 µg total RNA in 18 µL RNase-free water in a 1.5-mL tube.
45. Add 0.75 µL of synthetic E.coli tRNA-Lys-UUU-CCA (7.5 ng/µL) and 0.75 µL of synthetic E.coli tRNA-Lys-UUU-CC (2.5 ng/µL).

Note: If starting with RNA that has been subjected to periodate oxidation and β-elimination, the next step is not necessary; proceed directly to step 49.

46. Add 1.5 µL of 1 M Tris-HCl (pH=9.0) and 0.5 µL SUPERase·In.

47. Mix and incubate at 37°C for 45 min to deacylate tRNA.
48. Proceed to step 49.

Note: This step ensures the presence of a free 3' OH group in tRNAs, which is a prerequisite for efficient ligation of a 3' adapter in step 50. Although most tRNAs in total RNA samples isolated using standard protocols are likely to be deacylated, the stability of the acyl linkage can vary by an order of magnitude for different aminoacyl-tRNAs[17].

RNA 3' dephosphorylation

Timing: 1.5 h

This step entails the enzymatic removal of RNA 3′ phosphates resulting from β-elimination.

Note: We also perform this reaction when starting with total RNA that has not been subjected to oxidation and β-elimination. This serves to ensure that tRNAs with 3' ends cleaved during stress[18] or as a result of ribosome-associated quality control[19] are also represented in sequencing libraries.

49. Assemble the following reaction in a 1.5-mL microfuge tube:

RNA 3′ end dephosphorylation reaction

Reagent	Final concentration	Amount
Total RNA	0.025–0.1 µg/µL	X µL
10 × T4 PNK buffer	1 ×	10 µL
T4 PNK (10 U/µL)	5 U/µL	1 µL
SUPERase·In (20 U/µL)	5 U/µL	0.5 µL
RNase-free water	N/A	88.5 -X µL
Total	**N/A**	**100 µL**

50. Incubate at 37°C for 45 min.
51. To each tube, add:
    a. 10 µL 3 M sodium acetate (pH=4.5).
    b. 25 µg glycogen.
    c. 300 µL ice-cold 100% ethanol.
52. Vortex well and incubate at −20°C for at least 30 min.

Pause point: Samples can be stored at –20°C indefinitely.

53. Precipitate RNA by centrifuging for 30 min at 16,000 × g at 4°C.
54. Carefully remove the supernatant with a 1-mL filter pipette tip. Briefly spin down and remove all residual liquid with a 10-µL filter pipette tip.
55. Resuspend pellets in 5 µL RNase-free water.
56. Add 5 µL 2 × RNA loading buffer without dyes.

CRITICAL: Commonly used dyes such as xylene cyanol or bromophenol blue can co-migrate with RNA or DNA fragments of interest. Their co-purification with nucleic acids after gel extraction may interfere with downstream enzymatic reactions. Therefore, we recommend that these dyes be present only in the loading buffer added to marker samples or in empty wells.

Purification of tRNA from total RNA by gel size selection

Timing: 30 min–1 h

Transfer RNAs are 70–90 nt in length and run as a discrete cluster on denaturing 10% polyacrylamide gels. Gel size selection of 60–100 nt RNAs is a cost-effective approach to purify mature, intact tRNAs from total RNA preparations and to separate them from other highly abundant short RNAs with a similar size (e.g., 5S rRNA, 5.8S rRNA, snRNAs) or potential tRNA fragments and degradation intermediates, which would co-elute in small RNA fractions obtained by column-based commercial kits.

Note: The presence of RNA transcripts with substantially fewer RT-blocking modifications that tRNAs may result in their preferential use as templates for cDNA synthesis in steps 102–107.

57. Briefly soak glass plates in 1% SDS and rinse with RNase-free water.
58. Cast a 10% denaturing polyacrylamide gel and let it polymerize for 30 min to 1 h.
59. Place gel in an appropriate tank and pre-run at 20 mA in 1 × TBE for at least 30 min.

Note: Pre-running is critical for heating up urea-containing polyacrylamide gels, which aids RNA denaturation, and for removing excess urea from the wells, which can otherwise lead to band distortion.

60. Prepare gel size marker by mixing 1 µL NEB Low Range ssRNA ladder with 4 µL RNase-free water and 5 µL 2 × RNA loading buffer supplemented with xylene cyanol (0.005%) and bromophenol blue (0.01%).
61. Denature ladder and samples from step 30 at 90°C for 3 min and place immediately on ice.

62. Turn off the gel tank power supply.
63. Rinse gel wells to remove excess urea immediately before loading.
64. Load ladder and samples on gel and run at 20 mA until the bromophenol blue from the marker sample reaches the bottom of the gel.
65. Place gel in an RNAse-free container with SYBR Gold (1:10 000) in 1× TBE for 3 min on a platform shaker.

Note: Longer staining periods do not improve signal substantially and may result in fuzzy bands and sample loss.

66. Place gel on a clean foil on top of a blue light transilluminator to visualize RNA (for an example, see **Figure 2A**).

CRITICAL: Visualization of total RNA on a denaturing gel is a useful indicator of RNA quality. Transfer RNAs, 5S rRNA, and 5.8S rRNA should be visible as discrete bands. Band smearing is indicative of RNA degradation, which can occur during sample collection and/or RNA isolation.

67. Use a clean scalpel blade to cut out the gel region corresponding to tRNA (~60–100 nt).
68. Place the gel fragment in a 1.5-mL microfuge tube.
69. Elute RNA from gel slice (steps 70–83, resuspend the pellet in 8 µL RNAse-free water).

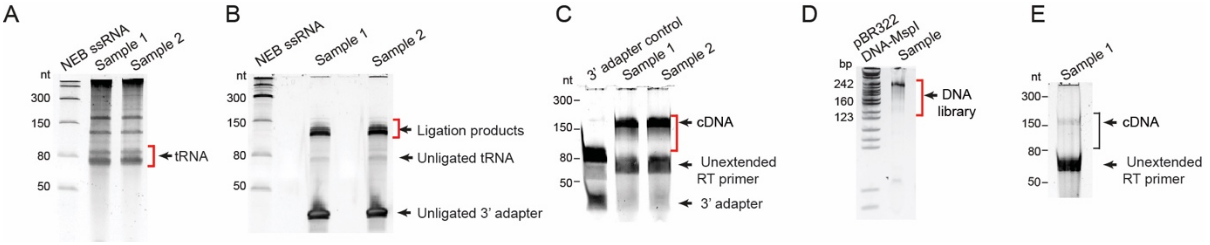


Figure 2 Typical gel images of key steps during mim-tRNAseq library construction.
Gel regions to be excised are indicated by a red bracket.
(A) Total RNA after Step 66.
(B) Adapter-ligated tRNA after Step 92.
(C) cDNA after successful reverse transcription in steps 105–113.
(D) Library DNA after step 135.
(E) cDNA after suboptimal reverse transcription in steps 105–113.

Elution of RNA from gel slices

Timing: 30 min - overnight

Inefficient recovery of RNA from gel slices is a major source of sample loss in sequencing library construction workflows that require gel size selection. We optimized this step to increase the yield of eluted RNA from ~20% in the classical "crush and soak" method[20] to ~70%–80%.

70. Crush gel slice with a disposable 1.5-mL tube pestle.
71. Add 400 μL RNA gel extraction buffer.
72. Incubate tubes at 65°C for 10 min and 2000 rpm in a Thermoblock.

Alternatives: Gel slurry can also be incubated at 65°C in a water bath. In this case, mix slurry by inverting tubes several times every 2–3 min.

73. Snap-freeze gel slurry on dry ice or in liquid nitrogen.
74. Thaw gel slurry at 65°C for 5 min and 2000 rpm in a Thermoblock.
75. Elute RNA overnight at room temperature with gentle mixing on a rotating wheel.

CRITICAL: Omitting the freeze-thaw cycle substantially decreases RNA elution efficiency.

Alternatives: An incubation time of ~2 h at room temperature in the final step is sufficient to achieve an elution efficiency of 40%–50%.

76. Remove gel pieces by centrifuging slurry through a SpinX filter at 10,000 × g for 30 s.
77. Transfer flow-through to a new 1.5-mL microfuge tube.
78. Add 25 μg glycogen and 1 mL ice-cold 100% ethanol.
79. Vortex well and incubate for at least 30 min at −20°C.

Pause point: Samples can be stored at –20°C indefinitely.

80. Precipitate RNA by centrifuging for 30 min at 16,000 × g at 4°C.
81. Carefully remove the supernatant with a 1-mL filter pipette tip. Spin down briefly and remove all residual liquid with a 10-μL tip.
82. Air-dry pellets for 2–3 min and resuspend in the indicated volume of RNase-free water.
83. Measure RNA concentration on a Nanodrop or an equivalent spectrophotometer.

Pause point: Samples can be stored at –80°C for several months.

Adapter ligation to tRNA 3' ends

Timing: 4 h

This step adds a barcoded 3’ adapter oligonucleotide (Table S1) to tRNAs to serve as a priming site for cDNA synthesis by TGIRT.

CRITICAL: Select a distinct barcoded adapter (I1 to I8) for each sample that will be pooled prior to reverse transcription.

CRITICAL: Use block design[21] to minimize the potential for differences between samples to be confounded by technical variation. For example, do not pool all replicates of a control condition in RT reaction 1 and all replicates of a treatment condition in RT reaction 2.

84. Assemble the 3’ adapter ligation reaction.

3’ adapter ligation reaction

Reagent	Final concentration	Amount
Gel-purified tRNA (100 ng)	0.2 μM	5 μL
Pre-adenylated 3′ adapter (20 μM)	1 μM	1 μL
50% PEG-8000	12.5%	10 μL
10 × T4 RNA ligase buffer	1 ×	2 μL
T4 RNA ligase 2, truncated KQ (200 U/μL)	10 U/μL	1 μL
SUPERase·In (20 U/μL)	1 U/μL	1 μL
Total	**N/A**	**20 μL**

CRITICAL: The 50% PEG-8000 stock solution is very viscous, which impacts pipetting accuracy. Inaccurate pipetting can decrease 3’ adapter ligation efficiency, which strongly depends on the concentration of PEG-8000[3]. For best results, equilibrate the 50% PEG-8000 stock solution at room temperature and pipet slowly and/or or use wide bore pipette tips. PEG-8000 should be added to each reaction tube separately rather than being included in a master mix.

85. Mix each reaction by carefully pipetting the entire volume up and down several times.
86. Incubate at 25°C for 3 h.
87. Pool reactions from samples with different barcodes and purify adapter-ligated tRNA with Zymo Oligo Clean & Concentrator kit according to the manufacturer’s protocol.
88. Elute in 10 μL RNase-free water.
89. Add 10 μL 2 × RNA loading buffer without dyes.
90. Prepare a 10% denaturing polyacrylamide gel and a ladder sample as in steps 57–60.

91. Denature ladder and samples from step 88 at 90°C for 3 min and place immediately on ice.
92. Separate ligation products from excess adapter by running samples on the gel as in steps 62–65.
93. Cut out bands corresponding to adapter-ligated tRNA (~110–125 nt, for example see **Figure 2B**) and put each gel fragment in a clean, low-biding RNase-free tube.
94. Recover adapter-ligated tRNA from gel slices as in steps 70–83.

CRITICAL: Reverse transcription by TGIRT is inhibited by trace amounts of ethanol. Ensure that all liquid is removed after precipitating adapter-ligated tRNA.

95. Resuspend pellets in 2.5 µL RNase-free water per sample pooled in step 87.
96. Measure RNA concentration on a Nanodrop or an equivalent spectrophotometer.

Pause point: Samples can be stored at −80°C for several months.

Alternatives: The gel purification step to remove excess 3′ adapter (steps 92–94) can be replaced by 5′-deadenylase treatment followed by digestion of the unligated adapter with the 5′–3′ ssDNA exonuclease RecJ[22]. This alternative substantially increases the cost of library construction and is very inefficient in our hands.

Primer-dependent reverse transcription of tRNA pools with TGIRT

Timing: 30 min – overnight

In this step, adapter-ligated tRNA pools are used as templates for primer-dependent cDNA synthesis by TGIRT, an engineered version of a bacterial group II intron-encoded reverse transcriptase[23,24], under reaction conditions that we optimized to favor tRNA modification readthrough.

CRITICAL: TGIRT cannot be substituted with commercially available reverse transcriptases of retroviral origin (e.g., ProtoScript II, SuperScript IV) since these enzymes have a much lower readthrough efficiency of modified sites in tRNAs.

Alternatives: Other group II intron- or retroelement-encoded enzymes such as MarathonRT[25] and an engineered version of the *B. mori* R2 retroelement reverse transcriptase[26] have also been reported to efficiently read through RNA modifications[27,28]. These enzymes are not commercially available at this time, and protocols for their use may need to be further optimized for tRNA templates.

97. Prepare 100 mM DTT fresh from powder (0.0154 g/mL in RNase-free water).
98. Pre-warm a Thermocycler block at 82°C.
99. Assemble the primer hybridization reaction in a low-binding PCR tube:

RT reaction: primer hybridization

Reagent	Final concentration	Amount
Adapter-ligated tRNA (100 ng)	N/A	X µL
RT primer (1.25 µM)	N/A	2 µL
RNase-free water	N/A	10 – X µL
Total	**N/A**	**12 µL**

Note: Including a control reaction with only unligated 3′ adapter as a template can be useful to gauge RT efficiency, as the RT primer binds to the last 15 nucleotides of the 34 nt-long adapters. The product of this control reaction can be used as a marker during the size selection of cDNA products on the gel in step 113.

100. Denature at 82°C for 2 min and incubate at room temperature for 5 min.
101. In the meantime, pre-warm a Thermocycler block at 42°C (with lid temperature at 48°C).
102. Assemble the RT reaction by mixing the adapter-ligated tRNA:RT primer duplexes with the following components:

RT reaction

Reagent	Final concentration	Amount
Adapter-ligated tRNA:RT primer	125 nM	12 µL
5 × TGIRT Reaction Buffer	1 ×	4 µL
DTT (100 mM; freshly prepared)	0.5 mM	1 µL
SUPERase·In (20 U/µL)	1 U/µL	1 µL
TGIRT-III (10 µM)	500 nM	1 µL
Total	**N/A**	**19 µL**

103. Mix each reaction by pipetting the entire volume up and down several times.
104. Incubate at 42°C for 10 min in a Thermocycler.
105. Add 1 µL 25 mM dNTPs to each tube.
106. Mix each reaction by pipetting the entire volume up and down several times.

107. Incubate at 42°C for 16 h in Thermocycler.

CRITICAL: Incubation times as short as 1 h can yield a substantial proportion of full-length cDNA products from tRNA pools with lower modification frequency and complexity, e.g., from budding or fission yeast. An extended reaction time, however, is particularly important for the efficient recovery of cDNA from tRNAs with multiple RT-blocking modifications, which is characteristic of many human tRNAs. The integrity of tRNA templates is not compromised under these reaction conditions[1].

CRITICAL: Efficient cDNA synthesis from endogenously modified tRNA requires a molar excess of TGIRT, so we highly recommend adhering to the template:TGIRT molar ratios described here.

CRITICAL: TGIRT is inhibited by trace amounts of ethanol. Ensure that all residual liquid is removed after precipitation of adapter-ligated tRNA in step 94.

108. Pre-warm a Thermocycler block at 95°C.
109. Add 1 µL 5 M NaOH to each reverse transcription reaction.
110. Incubate at 95°C for 3 min to hydrolyze the RNA template.
111. Add 20 µL 2 × RNA loading buffer to each reaction.
112. Denature cDNA at 95°C for 5 min and place samples immediately on ice.
113. Separate cDNA from unextended RT primer by running reactions alongside NEB Low Range ssRNA ladder (and optionally the 3′ adapter-only control reaction) on a 10% denaturing gel as in steps 57–65.
114. Excise the entire gel region above the unextended RT primer (or adapter-only RT product), which will correspond to all cDNA lengths derived from tRNA reverse transcription (for an example, see **Figure 2C**).
115. Put each gel fragment in a new 1.5 mL microfuge tube.
116. Crush gel slice with a disposable plastic pestle.
117. Add 400 µL 10 mM TE, pH=8.0, and snap-freeze gel slurry on dry ice or in liquid nitrogen.
118. Incubate gel slurry for 1 h at 70°C/2000 rpm in a Thermoblock to elute cDNA.

CRITICAL: Omitting the freeze-thaw cycle substantially decreases cDNA elution efficiency.

119. Remove gel pieces by centrifuging slurry through a SpinX filter at 10,000 × g for 30 s.
120. Transfer flow-through to a new 1.5-mL microfuge tube.

121. Add 25 µg glycogen, 40 µL 3 M NaCl, and 1 mL ice-cold 100% ethanol.
122. Vortex well and incubate for 30 min on ice.

Pause point: Samples can be stored at −20°C indefinitely.

123. Pellet DNA by centrifuging for 30 min at 4°C and 16,000 × g.
124. Carefully remove the supernatant with a 1-mL pipette tip. Briefly spin down and remove all remaining liquid with a 10-µL pipette tip.
125. Resuspend pellets in 5.5 µL water and proceed with cDNA circularization.

Pause point: Samples can be stored at −20°C for several weeks.

Circularization of cDNA

In this step, cDNA is circularized with CircLigase to provide a template for library construction by PCR.

126. Pre-warm a Thermocycler block at 60°C (lid temperature of 65°C).
127. Assemble cDNA circularization reaction in a low-binding PCR tube:

cDNA circularization reaction

Reagent	Final concentration	Amount
Gel-purified cDNA	n.d.	5.5 µL
Betaine (5 M)	1 M	2 µL
10 × CircLigase buffer	1 ×	1 µL
ATP (1 mM)	0.05 mM	0.5 µL
$MnCl_2$ (50 mM)	2.5 mM	0.5 µL
CircLigase ssDNA Ligase (100 U/ µL)	5 U/ µL	0.5 µL
Total	**N/A**	**10 µL**

128. Mix each reaction well by pipetting up and down.
129. Incubate at 60°C for 3 h in a Thermoblock.
130. Incubate at 80°C for 10 min to inactivate the enzyme.

Pause point: Circularized cDNA can be stored at −20°C for several weeks.

CRITICAL: We do not recommend using CircLigase II as it has much lower ssDNA circularization efficiency in comparable library construction protocols[3].

Note: Under the reaction conditions in step 102, TGIRT adds one to three non-templated adenosines to the 3′ ends of most cDNAs[1], resulting in a nearly identical 5' – 3′ circularization sequence context.

Library construction PCR

Timing: 30 min–1 h

The circularized single-stranded cDNA is used as a template for the construction of a double-stranded DNA library with an appropriate structure for sequencing on Illumina platforms.

CRITICAL: Use reverse library PCR primers with distinct 6-nt indexes to amplify different libraries that will be sequenced on the same flow cell.

131. Set up a 50-μL PCR reaction to construct libraries from circularized cDNA.

Library construction PCR reaction

Reagent	Final concentration	Amount
Circularized cDNA	n.d.	2 μL
dNTPs (10 mM)	0.3 mM	1.5 μL
Forward library construction primer (10 μM)	0.5 μM	2.5 μL
Reverse library construction primer (10 μM)	0.5 μM	2.5 μL
5 × KAPA HiFi GC buffer	1 ×	10 μL
KAPA HiFi DNA Polymerase (1 U/μL)	0.05 U/μL	0.5 μL
PCR-grade water	N/A	31 μL
Total	**N/A**	**50 μL**

CRITICAL: The use of other DNA polymerases or buffers can result in the preferential amplification of cDNAs with a shorter length or lower GC content[29], which introduces bias in tRNA abundance measurements.

Note: Circularized cDNA does not need to be purified prior to PCR amplification if its volume does not exceed 10% of the final PCR reaction volume.

132. Amplify for 4–6 cycles at a ramp rate of 3°C/second with the following settings:

PCR cycling conditions

Steps	Temperature	Time	Cycles
Initial Denaturation	95°C	3 min	1
Denaturation	98°C	20 s	4–6 cycles

Annealing	62°C	30 s
Extension	72°C	30 min
Hold	16°C	forever

Note: When performing RT of adapter-ligated tRNA and cDNA circularization according to steps 97–130, we find that 4–5 PCR cycles are sufficient to obtain DNA libraries of 2–5 nM, which is within the optimal concentration range for Illumina sequencing platforms. If more than 6 PCR cycles are necessary to achieve this yield, we recommend increasing the starting amount of circularized cDNA template in the PCR reaction or optimizing the reverse transcription reaction (see problem 1). Performing further PCR amplification will exacerbate bias as small differences in amplification efficiency of DNAs with different length and GC content will accumulate over multiple PCR cycles[29,30].

133. Purify PCR products with Zymo DNA Clean & Concentrator kit according to the manufacturer’s instructions.
134. Elute in 12 µL 10 mM Tris-HCl, pH=8.0.
135. Use 2 µL to measure DNA concentration with the Qubit dsDNA HS kit.

Note: The DNA concentration of samples is too low to be accurately quantified by UV spectrophotometry.

Note: This step efficiently separates the double-stranded DNA libraries (210–225 bp) from PCR buffer components, excess dNTPs, and unused PCR primers (which are ~50 nucleotides and therefore below the retention cut-off). Primer removal will be less efficient if longer oligonucleotides are used for PCR. In this case, libraries should be purified by size selection on a non-denaturing polyacrylamide gel[22]. Prior to electrophoresis, we recommend that DNA clean-up is performed as in step 133 since the presence of the KAPA HiFi GC buffer causes band distortion. For an example of a typical gel image of a DNA library, see **Figure 2D**.

Pause point: Purified DNA can be stored in low-binding tubes at −20°C indefinitely.

136. Assess DNA size and purity by non-denaturing polyacrylamide gel electrophoresis on an 8% non-denatring polyacrylamide gel.

Alternatives: Library size can be analyzed on a High Sensitivity D1000 ScreenTape on an Agilent 2200 TapeStation Nucleic Acid System.

CRITICAL: If purifying libraries by gel size selection, we recommend cutting out fragments of ~150 bp to ~230 bp in length. This will ensure the inclusion of fragments resulting from premature termination of cDNA synthesis at (hyper)modified sites. These fragments may not be visible as discrete bands after SYBR Gold staining of RT or library PCR samples (**Figure 2C** and **Figure 2D**) since they originate from the very few remaining RT roadblocks in a small subset of tRNA transcripts (e.g., ms^{2}t^{6}A37 in human tRNA-Lys-UUU). Failure to include these fragments in the sequencing run, however, will lead to underestimating the abundance of the tRNAs that give rise to them.

Library sequencing on Illumina platforms

The libraries generated with this protocol should be sequenced on an Illumina platform with a single-end run of >=100 bp. The minimal sequencing read length can be determined by adding 12 nt to the longest predicted tRNA transcript in the organism of interest. This accounts for the supplementary sequences added during library construction and required for demultiplexing samples pooled prior to reverse transcription.

Note: Given the high quality scores of single-end reads with a length of 100–150 nucleotides on current Illumina platforms, paired-end sequencing is not required for the analysis of tRNA-derived libraries.

137. Pool and dilute sequencing libraries according to the requirements of your preferred Illumina sequencing provider.

Note: The common range of starting library concentration prior to cluster generation on Illumina sequencing platforms is 1 nM–4 nM.

Note: For NextSeq 550, the loading concentration of tRNA libraries to achieve an optimal cluster density (~220 k/mm2) is 2 pM.

Note: Most organisms encode only a few hundred tRNA transcripts and the libraries generated with this protocol consist almost exclusively of tRNA-derived reads. We, therefore, find that a sequencing depth of 3–5 million reads per sample is sufficient for most analyses. This should be scaled by the number of individual samples pooled prior to reverse transcription, i.e., 40 million reads for a library that contains eight samples, each with a unique barcoded 3′ adapter.

Demultiplexing and-adapter trimming of sequencing reads

Timing: 30 min–1 h

Using fastq files as an input, library demultiplexing and trimming extra sequences at the 3′ and 5′ added during library construction is performed with cutadapt v3.5 (or newer).

138. Prepare a file of barcodes for demultiplexing.

Note: *barcodes.fa* should be a fasta-formatted file with one user-selected header and sequence per barcoded adapter used in library construction. To ensure an efficient match to sequences in reads, we extend the sequence to 10nt (instead of only the 5nt barcode sequence itself). Therefore, each sequence in the file contains the common 5′-GAT sequence + unique barcode + CA-3'. Below is an example of *barcodes.fa*:

```
# Example barcodes.fa file for demultiplexing
>I1
GATATCGTCA
>I2
GATAGCTACA
>I3
GATGCATACA
>I4
GATTCTAGCA
```

Note: The names of the barcodes in *barcodes.fa* can be chosen freely and will be present in the names of the demultiplexed fastq files.

139. Demultiplex fastq files with the following command:

```
# Demultiplex each sample fastq file using barcodes.fa

> for i in *.fastq.gz
      do fn=$(basename $i .fastq.gz)
      cutadapt   --no-indels   -q   30,30   --trimmed-only   -j   10   -a
      file:barcodes.fa   -m   10   -o   $fn'_{name}_trim.fastq.gz'   $i   1>
      $fn'_log.txt'
done
```

Note: This command uses a for loop to process each fastq file similarly. First, we create a filename variable $fn for more appropriate output file names lacking the file type extension and any other uninformative information (the basename command might need to be modified depending on the naming conventions of your files). Cutadapt is then used to quality-trim each read end with a threshold of 30 (-q 30,30) before demultiplexing according to *barcodes.fa*. Additionally, we discard trimmed reads shorter than 10 nt (-m 10) and prohibit indels in matches to barcode sequence (--no-indels).

Note: All reads of >= 100 nt should contain an adapter sequence since mature tRNAs are <=90 nts, and so only reads with a detected adapter which have been trimmed are kept (--trimmed-only).

CRITICAL: Adjust -j to a suitable number of processors for demultiplexing according to your system capabilities.

Note: An output *log.txt* file will be produced for each sample processed above (with 1> $fn'_log.txt'). We highly recommend assessing these files for trimming efficiency. Typically, the vast majority of reads should contain barcodes and adapters (as described above), and so the "Reads with adapters" value reported in the log should reflect that (i.e., at least 80% of reads containing adapters). Low values may suggest issues in library construction, because of e.g., the presence of no-insert libraries due to low RT efficiency.

140. Trim the two additional nucleotides introduced at the 5′ of reads following cDNA circularization.

```
# Trim 2x 5’ random nucleotides

> for i in *trim*.fastq.gz
      do fn=$(basename $i trim.fastq.gz)
      cutadapt -j 10 -m 10 -u 2 -o $fn'_trimFinal.fastq.gz' $i
done
```

Optional: Demultiplexing and trimming are now complete. However, you may wish to rename the *"trimFinal.fastq.gz"* files according to the samples they represent. For this, a table of which samples were multiplexed during library preparation, and which barcodes were used for each is useful to track which final trimmed fastq file represents each sample and/or condition.

Preparing mimseq input files

Timing: 15 min – days for custom references

Optional: Building custom references for `mimseq`:

`mimseq` contains pre-generated input files for multiple reference genomes. These are named using the first letter of the genus and the first three of the species. For example, *Homo* *sapiens* is given as Hsap. For a complete list of available species, please see the `-s/--species` section of the `mimseq` help page (`mimseq --help`) or the corresponding section in the documentation (https://mim-trnaseq.readthedocs.io/en/latest/start.html#pre-built-references). These references also contain the sequence for the synthetic *E. coli* tRNA-Lys-UUU spike-in that is added in step 45 (http://gtrnadb.ucsc.edu/genomes/bacteria/Esch_coli_K_12_MG1655/genes/tRNA-Lys-TTT-1-1.html).

Note: For species without pre-generated references, custom input files can be given. Ideally, the species will have predicted tRNAs on GtRNAdb (http://gtrnadb.ucsc.edu/). This data can be downloaded, and the respective tRNA fasta file and intron information file (*.out* file) can be specified to `mimseq` with `-t` and `-o`, respectively. If available, mitochondrial and/or plastid tRNA sequences (in the case of plant species) from the mitochondrial tRNA database (mitotRNAdb; http://mttrna.bioinf.uni-leipzig.de/mtDataOutput/) and PtRNAdb (http://14.139.61.8/PtRNAdb/index.php) can be given in space-separated format with `-m`. Plastid sequences, in the case of plant species, can also be specified with `-m`.

Note: tRNAScan-SE2.0[31] can be used to predict tRNA genes de novo and to generate the outputs used in `mimseq`.

CRITICAL: The formatting of output files and tRNA names must follow the convention as in GtRNAdb files and other `mimseq` prebuilt references. As an example, see the *S. cerevisiae* sacCer3 reference (https://github.com/nedialkova-lab/mim-tRNAseq/tree/master/mimseq/data/sacCer3-eColitK).

Note: The use of custom input files for species not currently included in `mimseq` has not been extensively tested and might lead to errors at runtime. See troubleshooting problem 3 for advice on how to overcome those.

Note: For species that have pre-built references in `mimseq`, or once the necessary inputs have been generated, the only input file needed for `mimseq` operation is one that specifies the sample data. This is a tab-separated file with two columns; the first specifies the path to the trimmed fastq files from step 3, and the second specifies the condition or treatment group of the sample.

Note: From here on, protocol details are given in reference to running the example data present in the `mimseq` GitHub repository (https://github.com/nedialkova-lab/mim-tRNAseq;

see *mimseq_hek_1.fastq.gz, mimseq_hek_2.fastq.gz, mimseq_k562_1.fastq.gz, mimseq_k562_2.fastq.gz* and *sampleData_HEKvsK562.txt*). These data are a subset of human HEK293T and K562 data generated in the original publication[1].

141. Create and save sample data tab-separated text file.

```
./mimseq_hek_1.fastq.gz       HEK293T
./mimseq_hek_2.fastq.gz       HEK293T
./mimseq_k562_1.fastq.gz      K562
./mimseq_k562_2.fastq.gz      K562
```

Running mimseq

Timing: 15 min–1 h

```
usage: mimseq [options] sample data
```

The mimseq pipeline and its most important parameters are schematically outlined in **Figure 3**.

Running the mimseq command only requires that the sample data input file is specified last. All other parameters can occur in any order.

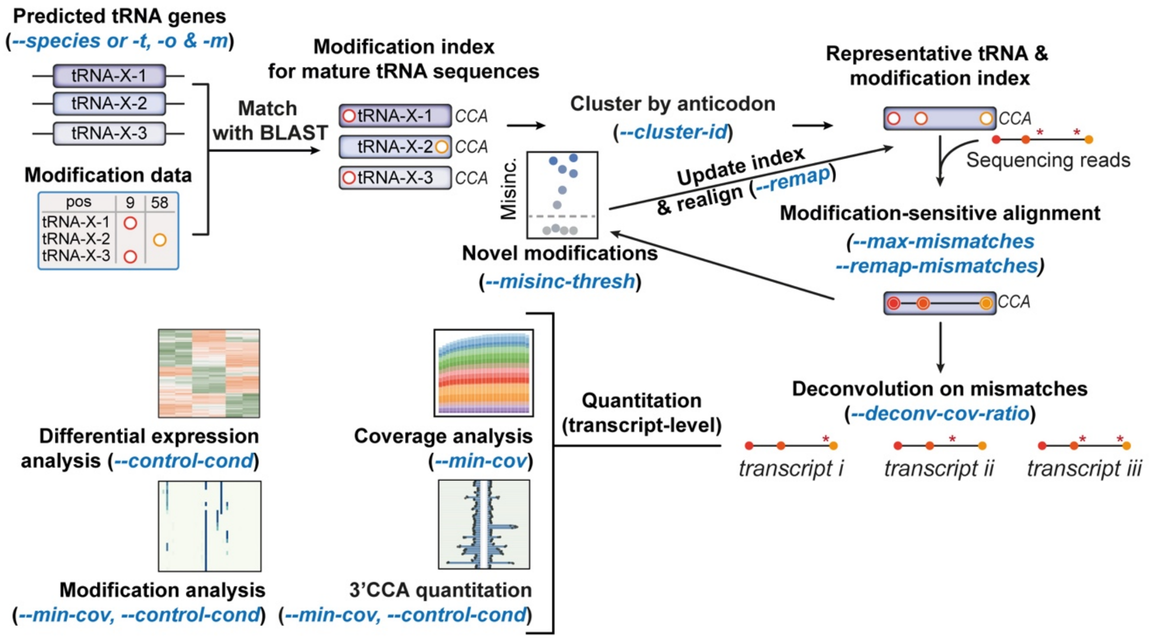

Predicted tRNA genes
(--species or -t, -o & -m)
tRNA-X-1
tRNA-X-2
tRNA-X-3
Modification data
pos
9
58
tRNA-X-1
tRNA-X-2
tRNA-X-3
Match with BLAST
Modification index for mature tRNA sequences
tRNA-X-1 CCA
tRNA-X-2 CCA
tRNA-X-3 CCA
Cluster by anticodon
(--cluster-id)
Misinc.
Novel modifications
(--misinc-thresh)
Update index & realign (--remap)
Representative tRNA & modification index
CCA
Sequencing reads
Modification-sensitive alignment
(--max-mismatches
--remap-mismatches)
CCA
Deconvolution on mismatches
(--deconv-cov-ratio)
transcript i
transcript ii
transcript iii
Quantitation (transcript-level)
Differential expression analysis (--control-cond)
Coverage analysis (--min-cov)
Modification analysis (--min-cov, --control-cond)
3'CCA quantitation (--min-cov, --control-cond)

Figure 3 Schematic of the mimseq computational pipeline
Highlighted are the main customizable parameters and the analysis steps they affect. See text in running mimseq and mimseq --help for detailed parameter descriptions.

142. Customize [optional] parameters.

Note: Listed below are the major customizable parameters that should be considered. This is not an exhaustive list; please see `mimseq` --help for all parameters. Parameters that are required are indicated.

a. -n: Experiment name. Output files and indices will have this prefix. **REQUIRED**.
b. --out-dir: Output directory name. This directory cannot exist. If it does, for example, from a previous `mimseq` run, please remove it and rerun `mimseq`. Default is the current directory.
c. --species or -t, -o and -m: Select species/genome of interest (see https://mim-trnaseq.readthedocs.io/en/latest/start.html#pre-built-references), or specify input files (see preparing `mimseq` input files). **REQUIRED**; either `--species`, or `-t` and `-o`.
d. --control-condition: Specify the condition you would like to use as the control. All comparisons in DESeq2 and differential modification analysis will be reported relative to the control. This must match one of the conditions listed in the sample data file exactly. **REQUIRED**.
e. --threads: Total available processors to use during analysis (particularly alignment). Please scale according to resource availability on your server/computer.
f. --cluster-id: Cluster identity threshold for tRNA clustering, between 0 and 1. This is a genome-specific parameter that needs to be determined for the specific needs of the user. Generally, a good choice will maximize uniquely mapping reads and minimize multi-mapping reads with little compromise to total deconvoluted sequences. In our experience, this value usually lies between 0.90 (e.g., yeast tRNAs) and 0.97 (human tRNAs).

Note: When analyzing your own data for the first time, we recommend performing several `mimseq` runs with –cluster-id set to 0.90, 0.93, 0.95, 0.97, or 1, and otherwise identical settings. Inspect alignment rate plots and deconvoluted transcript fractions printed in the log file to identify the optimal settings for your datasets, biological question, and organism of interest. A --cluster-id 1 run is equivalent to the commonly used tRNA read alignment strategy of collapsing multi-copy tRNA genes into a single reference. In our experience, this results in up to 25% of multi-mapping reads, depending on the organism.

g. --deconv-cov-ratio: Threshold of the required ratio between coverage at 3′ end and mismatch used for deconvolution. Coverage reductions greater than the threshold will result in non-deconvoluted sequences. This should be adjusted according to the general coverage quality of your libraries (see plots in the *cov/* output folder and expected outcomes). Low sequence coverage at 5′ ends with a high --deconv-cov-ratio will result in more non-deconvoluted sequences.

h. --max-mismatches and --remap-mismatches: Controls the proportion of mismatches allowed as a fraction of read length in the first and second round of alignment, respectively (excluding known and predicted modification sites). Due to new modification detection after the first round of alignment, it is generally advisable to reduce --remap-mismatches relative to --max-mismatches to reduce spurious and inaccurate read alignment in the second round. **Setting either parameter higher than 0.1 is not advisable!**

Note: When analyzing your own data for the first time, we recommend performing several `mimseq` runs with --max-mismatches set to 0.075 or 0.1 and --remap-mismatches set to 0.05 or 0.075 and otherwise identical settings. Inspect alignment rate plots and deconvoluted transcript fractions printed in the log file to identify the optimal settings for your datasets, biological question, and organism of interest.

i. --min-cov: Minimum coverage per unique tRNA after deconvolution required for inclusion in coverage plots, modification analysis, and 3′-CCA analysis. This can be a fraction of total mapped reads between 0 and 1 or an integer representing absolute coverage. Note that all clusters are included for differential expression analysis with DESeq2. Default = 0.0005 (0.05% mapped reads).

j. --max-multi: Maximum number of bam files to process simultaneously. Increasing this number reduces processing time but increases total memory usage. Default is 3, maximum is the total number of samples.

Note: Processing too many files at once can cause termination of mim-tRNAseq due to insufficient memory. If mim-tRNAseq fails during coverage calculation, lower this parameter.

k. --misinc-thresh: Required fraction of reads per cluster or transcript containing a given mismatch to pass the novel modification detection threshold. Default is 0.1, i.e., 10% of cluster/transcript aligned reads must contain a given mismatch to call a new modification.

143. Run `mimseq`.

Note: An example of a `mimseq` command is given below using the example data in the GitHub repository. This utilizes the Hsap (hg38) reference with clustering at 97% sequence identity, allows 10% mismatches per read length on round 1 of alignment, and subsequently 7.5% on realignment, and filters all tRNAs with less than 0.05% total mapped reads from plots.

```
> mimseq --species Hsap --cluster-id 0.97 --threads 15 --min-cov 0.0005 -
-max-mismatches 0.075 --control-condition HEK293T -n hg38_test --out-dir
hg38_HEK239vsK562   --max-multi   4   --remap   --remap-mismatches   0.05
sampleData_HEKvsK562.txt
```

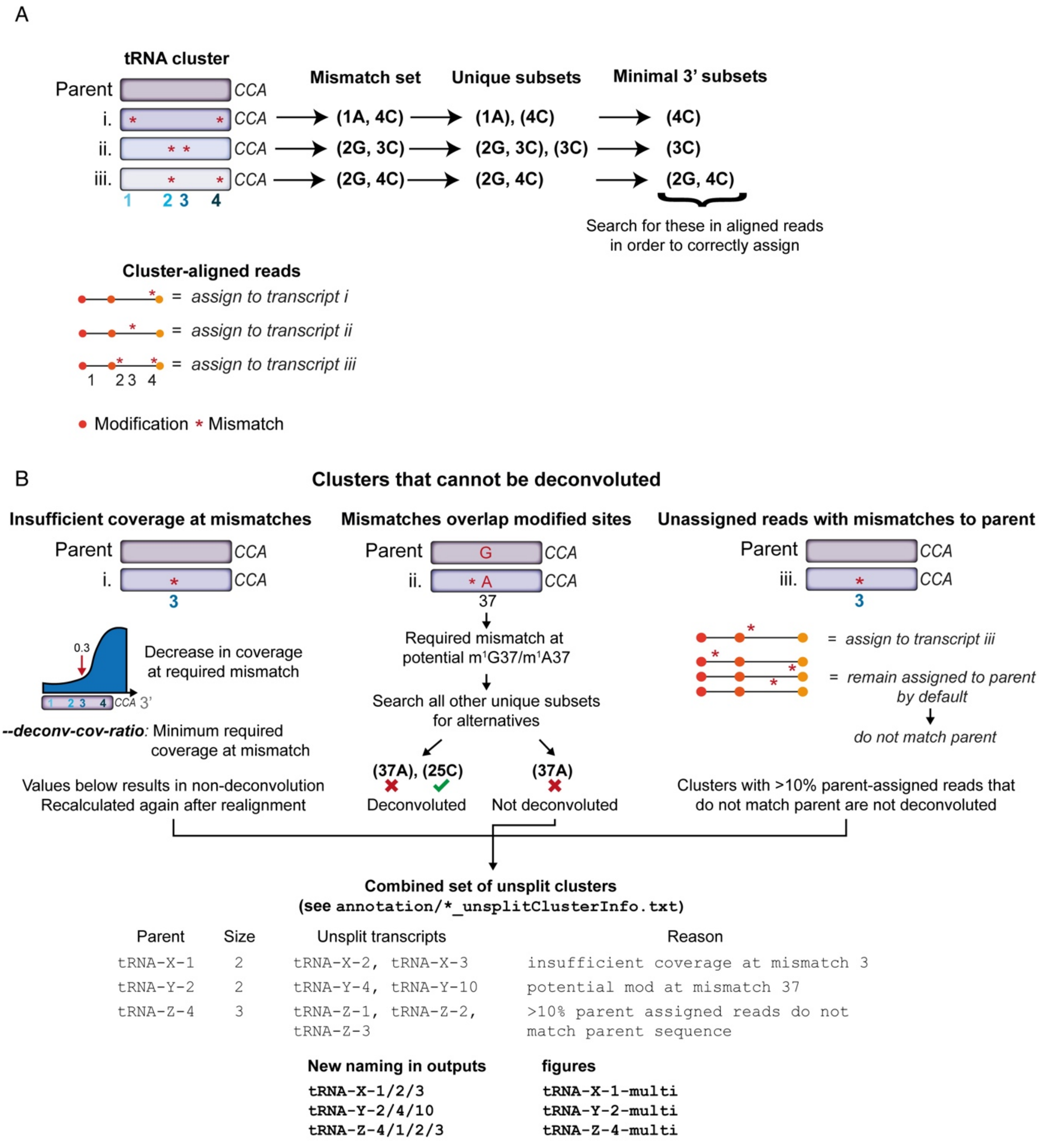
Note: The above command will utilize the example data supplied in the GitHub repository. Please download the four fastq files and *sampleData_HEKvsK562.txt* file if you want to reproduce this analysis.

Cluster deconvolution

By clustering tRNAs within an anticodon family by a sequence identity threshold, reads are aligned to a representative cluster parent, which substantially reduces multi-mapping for reads from nearly identical tRNA transcripts[1]. To restore single-transcript resolution for subsequent analyses, we developed a cluster deconvolution algorithm which reassesses aligned reads and reassigns them to tRNA transcripts based on mismatch patterns to the cluster parent sequence.

Note: Since the original publication, we have introduced several major updates to the cluster deconvolution algorithm (**Figure 4A**). Since v0.3, the set of all mismatches between each unique tRNA transcript and the cluster parent are considered, rather than single mismatches as in earlier versions. From the full set of mismatches, unique 3′ subsets are determined, and aligned reads are searched for these mismatches to reassign them to unique transcripts within a cluster. This theoretically allows the distinction and deconvolution of all unique tRNA transcripts. When analyzing tRNA pools from an organism that encode a high number of highly similar tRNA transcripts, such as mammals, a small proportion of clusters, or particular transcripts within a cluster, cannot be deconvoluted due to three main reasons (**Figure 4B**):

- A small subset of clusters may still exhibit 3′ coverage bias due to modifications that induce stops to RT, such as the rare ms^{2}t^{6}A/ms^{2}i^{6}A [1]. Complete deconvolution of reads for such clusters might not be possible if coverage is significantly lower at mismatches required for deconvolution than at the 3′ end of the tRNA. To overcome this, the --deconv-cov-ratio parameter can be used to set a threshold for this difference in coverage (see running mimseq). Coverage at positions required for deconvolution not passing this threshold will result in the transcript being marked as not deconvoluted.


A
tRNA cluster
Parent
CCA
i.
ii.
iii.
1 2 3 4
Mismatch set
Unique subsets
Minimal 3′ subsets
(1A, 4C)
(1A), (4C)
(4C)
(2G, 3C)
(2G, 3C), (3C)
(3C)
(2G, 4C)
(2G, 4C)
(2G, 4C)
Search for these in aligned reads in order to correctly assign
Cluster-aligned reads
= assign to transcript i
= assign to transcript ii
= assign to transcript iii
1 2 3 4
● Modification * Mismatch
B
Clusters that cannot be deconvoluted
Insufficient coverage at mismatches
Parent
i.
3
0.3
Decrease in coverage at required mismatch
--deconv-cov-ratio: Minimum required coverage at mismatch
Values below results in non-deconvolution
Recalculated again after realignment
Mismatches overlap modified sites
Parent
G
ii.
* A
37
Required mismatch at potential m¹G37/m¹A37
Search all other unique subsets for alternatives
(37A), (25C)
(37A)
Deconvoluted
Not deconvoluted
Unassigned reads with mismatches to parent
Parent
iii.
3
= assign to transcript iii
= remain assigned to parent by default
do not match parent
Clusters with >10% parent-assigned reads that do not match parent are not deconvoluted
Combined set of unsplit clusters
(see annotation/*_unsplitClusterInfo.txt)

Parent	Size	Unsplit transcripts	Reason
tRNA-X-1	2	tRNA-X-2, tRNA-X-3	insufficient coverage at mismatch 3
tRNA-Y-2	2	tRNA-Y-4, tRNA-Y-10	potential mod at mismatch 37
tRNA-Z-4	3	tRNA-Z-1, tRNA-Z-2, tRNA-Z-3	>10% parent assigned reads do not match parent sequence

New naming in outputs	figures
tRNA-X-1/2/3	tRNA-X-1-multi
tRNA-Y-2/4/10	tRNA-Y-2-multi
tRNA-Z-4/1/2/3	tRNA-Z-4-multi

Figure 4 Fundamental principles of the new cluster deconvolution algorithm

(A) Schematic representation of deconvolution methodology. For each cluster, the set of mismatches distinguishing each transcript is found. For each set, the minimal unique subsets are found, from which the most 3′ subset is chosen. Reads are assessed individually for these mismatches in order to assign them to a member transcript within a cluster.

(B) Schematic representation of conditions when clusters or transcripts cannot be deconvoluted. Either coverage at a required mismatch is too low (left), the mismatch is a potentially modified site (middle), or >10% parent assigned reads contain mismatches to the parent.

- Some tRNA transcripts are only distinguishable from the parent by positions that might also be modified sites (for example, G26 or A58). In these cases, it is impossible to tell if a mismatch in a read is due to mismatches between cluster members or if it is due to misincorporations at the modified nucleotide. Such tRNA transcripts (and the parent of the cluster) are also labeled as not deconvoluted.
- Thirdly, reads that cannot be assigned to a transcript within a cluster are by default left assigned to the parent sequence. In very rare cases, these parent-assigned reads also contain mismatches that do not pass our modification calling criteria, which indicates that they might not originate from the parent transcript either. If 10% or more parent-assigned reads contain such mismatches, the entire cluster is not deconvoluted.

Transcripts that are not deconvoluted are grouped and their count and modification data are aggregated. These are then renamed to provide information on which transcripts remain clustered, and are treated similarly to other single transcripts for differential expression analysis, modification profiling, and other downstream analyses (**Figure 4B**). For further information on non-deconvoluted transcripts and clusters, please refer to the *annotation/*unsplitClusterInfo.txt* output file.

Expected outcomes

There are a number of expected outcomes from the `mimseq` computational pipeline that can be used as quality control (QC) for your experiment or to guide further optimizations to the library construction protocol and/or the parameters used for running `mimseq`. Below are examples and details of these outputs. Please refer to the package documentation for a description of the full set of `mimseq` outputs (https://mim-trnaseq.readthedocs.io/en/latest/output.html).

Alignment (see *align/* folder)

An important QC step is ensuring that the majority of sequencing reads are aligned uniquely to the reference of choice, with minimal multi-mapping reads. To assess this, see *mapping_stats.txt*, and *Remap_alignstats.pdf* if --remap is enabled, or *Primary_alignstats.pdf* otherwise.

Note: If --remap is enabled, each library will have two entries in the text file, the first representing the results from the first round of alignment, while those labeled "** NEW

ALIGNMENT ∗∗” indicate statistics after realignment. Please only assess the new alignment results.

For eukaryotic tRNA libraries constructed with our workflow, there should be >70% uniquely mapped reads and <5% multi-mapped reads (**Figure 5**). Many factors can influence these fractions, including various aspects of library preparation (RNA integrity, 3′ adapter ligation, RT, and cDNA circularization efficiency), the organism and tRNA reference, and the stringency of alignment (mismatch allowance). Please consider all of these aspects when troubleshooting low read mappability.

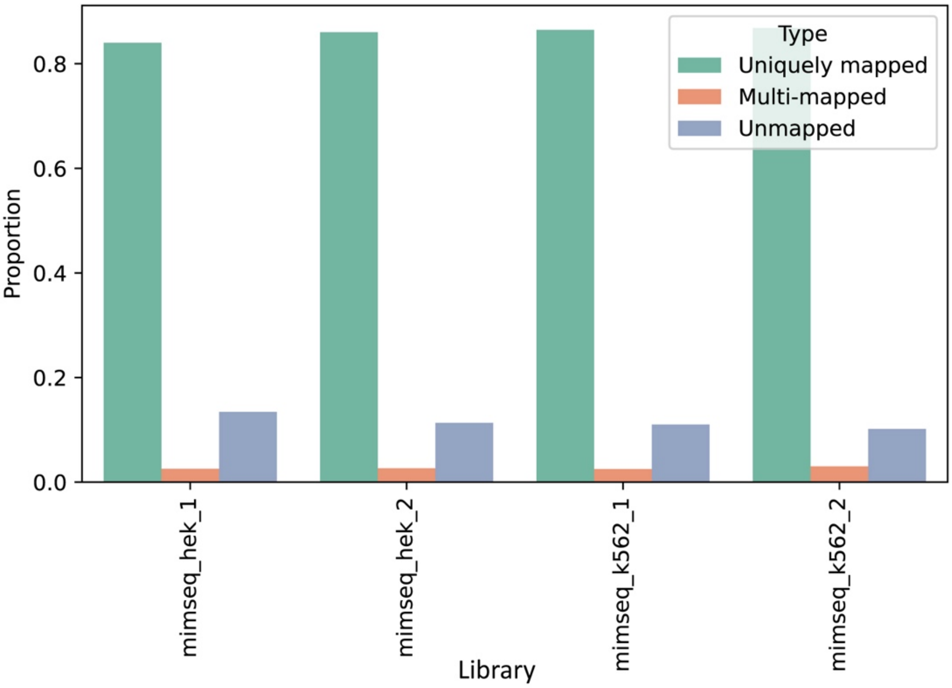


Figure 5 Alignment statistics for sample human dataset
Shown are the uniquely mapping, multi-mapped, and unmapped read proportions per library after realignment (if enabled with --remap as in step 142 - 143).

Coverage and full-length transcript proportions (see *cov/* and *mods/* folders)

Our optimized RT protocol (steps 97–107) reduces RT stops at modified tRNA sites by promoting misincorporation. This should lead to a high proportion of full-length reads and reduced 3′ end bias in libraries. An easy way to assess this is by looking at the metagene plots for total normalized coverage per library scaled to the 3′ end in *cov/coverage_byaa_norm_scaled.pdf* (**Figure 6A**). For the test data, there is consistently more than 62.5% cumulative coverage at tRNA 5′ ends, indicating efficient readthrough of modifications and a majority of full-length transcripts. Full-length transcript proportions can also be assessed by looking at *mods/RtstopTable.csv*. Here, you should see a high level of reads that stop (proportion column) for each tRNA transcript close to the 5′ end (canon_pos value close to 1, or 0 for tRNA-His). This is also visually represented in the “RT stops” heatmaps for each condition (top plot in *mods/*comb_heatmap.pdf*; **Figure 6B**).

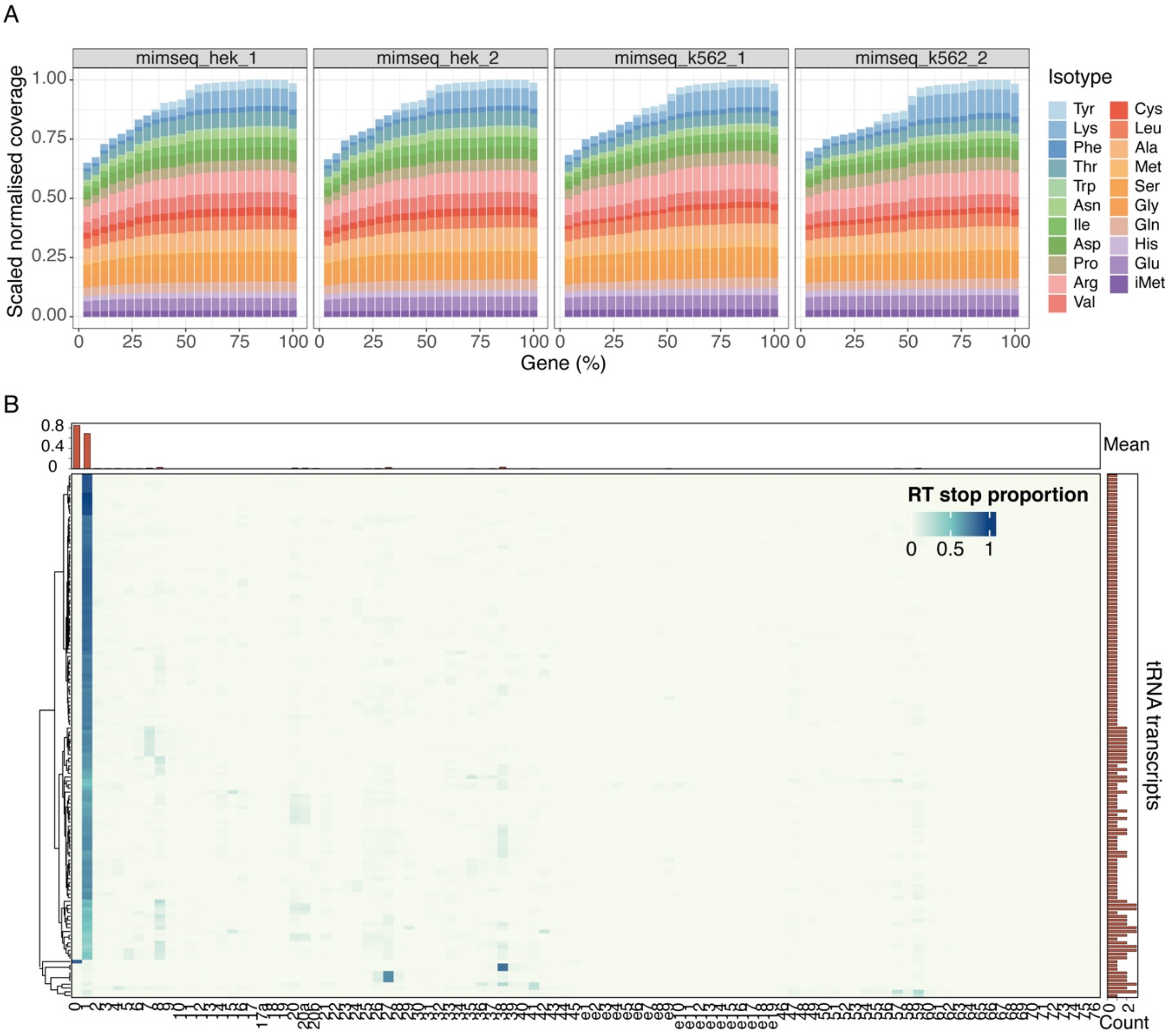


Figure 6 Quality control for tRNA coverage and full-length transcripts

(A) Metagene plots of coverage per nuclear-encoded tRNA isotype in each library specified in the sample data input file. Coverage is normalized to total mapped reads and scaled to the second last bin. See cov/coverage_byaa_norm_scaled.pdf.

(B) Global heatmaps of average proportions of stops to RT per canonical tRNA position for each unique tRNA transcript with coverage above 0.005 (as per --min-cov) in tRNA sequencing data from human K562 cells (n = 2).

Note: The --min-cov parameter will influence the filtering of tRNA transcripts with low coverage. In the heatmap below (**Figure 6B**), only filtered transcripts are displayed. If you notice many transcripts in these plots with noisy stop and misincorporation data, please check the value used for --min-cov and the proportion of stops and misincorporations for these transcripts and their counts in the *counts/* output folder. Raising the --min-cov threshold might help filter out these noisy, low coverage transcripts.

Normalized tRNA counts and replicate clustering (see *DESeq2/*)

If your samples include replicates, DESeq2 is automatically run using your desired --control-condition as a basis for comparison. Ideally, replicates would show high similarity to each other with regards to DESeq2 variance stabilizing transformed tRNA count data[12] to ensure high sensitivity for calling differentially expressed transcripts and anticodon pools (*isodecoder/* and *anticodon/* sub-directories, respectively; see *vst-transformedCounts.csv* for transformed count data). Here, it is useful to assess the Euclidean distance between the samples (*qc-sampledists.png*; **Figure 7A**), which should be low between replicates. Replicates should also cluster well, especially on principal component 1 in the PCA plots based on normalized count data for tRNA transcripts (*isodecoder/qc-pca.png*; **Figure 7B**). Poor clustering between samples might indicate biological variation in the samples (e.g., due to heterogeneity in cell-type composition of tissues) or large technical variation during library preparation and sequencing.

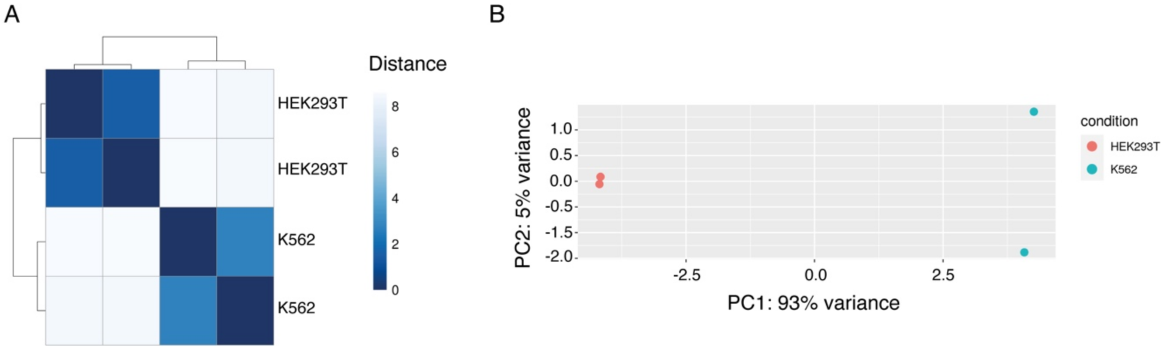


Figure 7 Assessing replicate similarity with variance stabilizing transformed (vst) tRNA count data from DESeq2

(A) Distance matrix representing pairwise Euclidean distance for each pair of samples.

(B) Principal component analysis (PCA) plot using the first two principal components from tRNA isodecoder analysis. Percent variance explained by each principal component is given in axis titles.

CRITICAL: We advise against using spike-in tRNAs for global-scaling data normalization and absolute tRNA quantification. The overall RNA content per cell, as well as the fraction of total RNA composed of tRNA, can vary substantially among different tissues, cell types, and growth conditions due to many biological factors[32]. These differences can unduly influence the number of reads mapping to the spike-in in each library, as the spike-in is added to a fixed amount of total RNA rather than to a fixed number of cells.

Note: If statistical power is low, which may be the case for more heterogeneous samples such as tissues, we recommend increasing the number of biological replicates rather than the sequencing depth[33]. Higher sequencing coverage may still be beneficial in specific cases, e.g., when analyzing misincorporation patterns at modified sites in low-abundance tRNA transcripts.

Limitations

We have successfully used mim-tRNAseq to profile tRNA abundance, aminoacylation, and modification status in samples from a wide range of eukaryotic organisms such as yeast, flies, plants, mouse tissues, and human cell lines. The experimental workflow is applicable to any organism, though it may require some optimization depending on the type and frequency of tRNA modifications present in the sample. We have successfully generated high-quality libraries starting with as little as 0.5 µg of total RNA. The protocol is currently not compatible with ultra-low input samples or single-cell methodologies.

The computational pipeline has a degree of customization to allow the user to fine-tune the analysis to their particular needs and can also be used with tRNA sequencing datasets generated with other tRNA library construction protocols. It is designed for single-end sequencing reads and currently does not work on paired-end ones. However, since mature tRNAs are <= 100 nt in length, paired-end data are unnecessary. Finally, mim-tRNAseq is currently not designed for investigating pre-tRNAs, but this functionality is currently under development.

Troubleshooting

Problem 1

Low yield of full-length cDNA after reverse transcription.

Potential solution

Performing reverse transcription of adapter-ligated tRNA under low-salt conditions and for an extended time is critical for obtaining a high proportion of full-length cDNA molecules. Typically, more than 50% of the RT primer is extended, and the major cDNA products visible on gel are full-length (**Figure 2C**). Efficient reverse transcription is critically important for minimizing amplification bias during library construction PCR.

Poor reverse transcription manifests as the extension of only a small fraction of the RT primer (**Figure 2E**) or the presence of shorter cDNA fragments detectable as dominant and discrete bands on gel[1].There are two potential causes for this: i) ethanol carry-over from template precipitation, which can inhibit reverse transcription, and ii) loss of TGIRT activity upon prolonged storage. We, therefore, recommend purifying adapter-ligated tRNA with the Zymo Oligo Clean & Concentrator after step 96. This can be done routinely if sample input amounts are not limited. In addition, we have observed that TGIRT can lose activity when stored for more than 3 months at −20°C. We recommend −80°C for prolonged enzyme storage while avoiding repeated freeze-thaw cycles.

Problem 2

When running `mimseq`, the step Analyzing misincorporations and stops to RT, and analyzing 3′ ends produces the following:

```
IndexError: Too many levels: Index has only 1 level, not 2
```

Potential solution

This error is often caused by an empty or sparse misincorporation table (*mods/mismatchTable.csv*). This might result from very poor alignment. First, check that adapter trimming has been done correctly by analyzing cutadapt logs (steps 138 – 140). Validate your trimming approach on another dataset, or by manually trimming reads at both ends to ensure some alignment.

Problem 3

When running `mimseq`, errors occur during “tRNA processing (Processing tRNA sequences…)” or shortly after. For example:

```
ID = re.search("tRNAscan-SE ID: (.*?)\).|\((chr.*?)-
",seqIO_dict[seqIO_record].description).groups()
AttributeError: 'NoneType' object has no attribute 'groups'
```

```
tRNA_dict[seq]['anticodon'] = anticodon = re.search('.tR(NA|X)-.?-(.*?)-
', seq).group(2)
AttributeError: 'NoneType' object has no attribute 'group'
```

```
anticodon = seq_parts[4]
IndexError: list index out of range
```

Potential solution

This usually indicates problems with your input reference files (see preparing mimseq input files). You have most likely specified custom input files with -t, -o, and/or -m. If these files are not present on GtRNAdb (http://gtrnadb.ucsc.edu/) for your species of interest, please ensure that the header for each sequence in the fasta file of genomic tRNAs is formatted exactly the same as those from in the `mimseq` pre-built indices. An example can be found here; pay close attention to the order of information, the number of fields separated by spaces, the naming convention for tRNA genes, and the tRNAScan-SE ID given in parentheses that matches the corresponding entry in the out file (specified with -o).

Note: If mitochondrial and/or plastid sequences are specified with -m, these also require specific formatting that is distinct from the nuclear genomic file as they match the format provided by the mitotRNAdb (http://mttrna.bioinf.uni-leipzig.de/mtDataOutput/). If formatting is incorrect, `mimseq` will produce an error similar to the third error above. See an example here for correct formatting; again, pay close attention to the number of fields per sequence header (i.e., 5) and the use of "|" as a field separator. In this case, the first field specifying the ID can be any user-chosen value. The third field is a unique species code, which is unused by `mimseq` but must be present.

Problem 4

Running mimseq fails at the alignment step with a non-zero exit status 9 error during GSNAP alignment:

```
subprocess.CalledProcessError: Command '['gsnap', '--gunzip', '-D',
'hg38_HEK239vsK562/Hsap_tRNAgenome', '-d', 'Hsap_tRNAgenome', '-V',
'hg38_HEK239vsK562/Hsapsnp_index', '-v',
'hg38_diff_modificationSNPs'...]' returned non-zero exit status 9.
```

Potential solution

These errors occur when `mimseq` tries to run GSNAP for read alignment. In this case, the *align.log* file in the *align/* folder can be very useful for debugging. Most commonly, there is an error in the path to the trimmed fatsq files supplied in the sample data file (see preparing mimseq input files). Carefully check these paths to make sure they point to files that exist. In this case, the log file will show something such as the following:

```
Cannot open gzipped file ./mimseq_k562_.fastq.gz
```

Other possibilities include incorrect GSNAP version installations. Please ensure that GSAP version 2019-02-26 is installed by typing "gsnap --version" in the terminal within your `mimseq` environment. With newer versions, an error will be produced due to changes in parameters available in the gsnap command:

```
gsnap.avx512: unrecognized option '--ignore-trim-in-filtering'
```

Problem 5

Samples show low uniquely mapped read proportions in *align/mapping_stats.txt.*

Potential solution

This problem may arise for several reasons. First, reassess your read trimming step and ensure that the correct barcodes and adapter sequences were specified in *barcodes.fa*. Check that the proportions of reads trimmed in the log files are as expected (>80% reads with adapters) and that not too many reads were excluded because they were too short. A high proportion of short reads may indicate poor modification readthrough during RT or too many PCR cycles during library construction, resulting in overamplifying short cDNA fragments. Ensure that RT is performed with templates of high purity and with a fresh enzyme batch, and minimize PCR cycles.

Secondly, check that your alignment and realignment mismatch allowance (Step 142-143) is not too stringent. Try raising these values and assessing how this impacts the alignment statistics for each alignment round. Be cautious not to raise these values too much as this may cause misalignment and spurious modification calling (see *mods/predictedMods .csv*).

Lastly, evaluate if you have contamination of other RNA types in your sample, such as rRNA or snoRNA. This may result from RNA degradation during sample collection and/or RNA isolation (Steps–1 - 12) or from imprecise cutting out of tRNA-containing gel fragments Step 38 (Figure 2A). Align trimmed reads to the full genome of the species of interest and assess areas and gene features with high read coverage to identify potential contaminants and optimize RNA isolation and/or tRNA size selection accordingly.

Problem 6

Spike-in sequences other than *E. coli* tRNA-Lys-UUU (Table S1) are used during library generation and spike-in addition.

Potential solution

`mimseq` reference files can be easily edited to include new sequences of interest. The only requirement is that the formatting of the sequence header is maintained as in other tRNA references.

There are two methods to achieve this:

1. The reference fasta files can be downloaded from GitHub (https://github.com/nedialkova-lab/mim-tRNAseq/tree/master/mimseq/data), edited, and specified to

mimseq with -t. Please note that the corresponding intron information *.out* file also needs to be specified with -o and can also be downloaded from the link above.

2. The reference files included in your local installation of mimseq can be directly modified. To find the location of these, activate your mimseq environment, determine the location of the mimseq executable file, and use this path to find the included *data/* folder. For example:

```
# find mimseq executable
> which mimseq
/home/drew/anaconda3/envs/mimseq/bin/mimseq
```

In this case, the prebuilt references will be found in

```
/home/drew/anaconda3/envs/mimseq/lib/python3.7/site-packages/mimseq/data/
```

The appropriate reference folder can be found here, and the fasta file within can be edited and saved.

This is more problematic if something goes wrong as the mimseq master files would have been permanently changed. However, this method does allow you to specify the reference simply with the --species parameter.

In both cases, if an unspliced intron-containing spike-in is used, there will need to be a matching entry in the corresponding *.out* file with a matching tRNAScan-SE ID number to ensure correct splicing.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Danny D. Nedialkova (nedialkova@mpg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The mim-tRNAseq computational pipeline is available under a GNU public License v3 on GitHub (https://github.com/nedialkova-lab/mim-tRNAseq), Zenodo (https://doi.org/10.5281/zenodo.6694873) and on Bioconda. The accession number for the sequencing data reported in the original publication is GSE152621. Example analyses presented here are based on a subset of the replicate HEK293T and K562 data (GSM4618859, GSM4618860, GSM4618861, GSM4618862) which are available in the GitHub repository. A package description and installation guide are available at https://mim-trnaseq.readthedocs.io/en/latest/. Supplementary Table S1 containing oligonucleotides, RNA sequences and primers for library construction can be found on Mendeley Data (www.doi.org/10.17632/vy8z394gfh.1).

Acknowledgements

We thank Katrin Strasser for assistance with library construction and the Max Planck Institute of Biochemistry Core Facility for high-throughput sequencing. This work was funded by the Max Planck Society, the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (ERC Starting Grant No. 803825-TransTempoFold), and SFB 1035 (German Research Foundation DFG, Sonderforschungsbereich 1035, Projektnummer 201302640, project B14). A.B. is supported by the International Max Planck Research School for Molecular Life Sciences.

Author contributions

Conceptualization, D.D.N.; Experimental methodology, D.D.N.; Software, A.B.; Writing, A.B., and D.D.N.; Supervision and Funding Acquisition, D.D.N.

Declaration of interests

A.B. and D.D.N. are inventors on a patent application filed by the Max Planck Society pertaining to the mim-tRNAseq technology.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xpro.2022.101579.

References

1. Behrens, A., Rodschinka, G. & Nedialkova, D. D. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol. Cell* **81**, 1–14 (2021).
2. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res.* **40**, (2012).
3. Heyer, E. E., Ozadam, H., Ricci, E. P., Cenik, C. & Moore, M. J. An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. *Nucleic Acids Res.* **43**, e2–e2 (2015).
4. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
5. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
6. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
8. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
9. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
10. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
11. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
12. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
13. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
14. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
15. Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M. & Pan, T. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep.* **6**, 151–157 (2005).
16. Evans, M. E., Clark, W. C., Zheng, G. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic Acids Res.* **45**, e133–e133 (2017).
17. Peacock, J. R. *et al.* Amino acid–dependent stability of the acyl linkage in aminoacyl-tRNA. *RNA* **20**, 758 (2014).
18. Czech, A., Wende, S., Mörl, M., Pan, T. & Ignatova, Z. Reversible and rapid transfer-RNA deactivation as a mechanism of translational repression in stress. *PLoS Genet.* **9**, e1003767 (2013).
19. Yip, M. C. J., Savickas, S., Gygi, S. P. & Shao, S. ELAC1 Repairs tRNAs Cleaved during Ribosome-Associated Quality Control. *Cell Rep.* **30**, 2106-2114.e5 (2020).
20. Green, M. R. & Sambrook, J. Isolation of DNA Fragments from Polyacrylamide Gels by the Crush and Soak Method. *Cold Spring Harb. Protoc.* **2019**, 143–146 (2019).
21. Auer, P. L. & Doerge, R. W. Statistical Design and Analysis of RNA Sequencing Data. *Genetics* **185**, 405–416 (2010).
22. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
23. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).
24. Qin, Y. *et al.* High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22**, 111–128 (2016).
25. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195 (2018).
26. Upton, H. E. *et al.* Low-bias ncRNA libraries using ordered two-template relay: Serial template jumping by a modified retroelement reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
27. Guo, L. T. *et al.* Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J. Mol. Biol.* **432**, 3338–3352 (2020).
28. Gustafsson, H. T. *et al.* Deep sequencing of yeast and mouse tRNAs and tRNA fragments using

OTTR. *bioRxiv* 2022.02.04.479139 (2022) doi:10.1101/2022.02.04.479139.

29. Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. (2012) doi:10.1038/nmeth.1814.

30. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).

31. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).

32. Coate, J. E. & Doyle, J. J. Variation in transcriptome size: are we getting the message? *Chromosoma* **124**, 27–43 (2015).

33. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301–304 (2014).

CHAPTER 4

Concluding remarks

Open-source and community-driven development within the mim-tRNAseq framework

Despite over 60 years since their initial discovery[1], many aspects of tRNA biology remain elusive. Their unique structural and chemical attributes remain difficult barriers to overcome in advancing deep-sequencing technology for their quantitation and analysis. However, their abundance, complex life-cycle, and pivotal role in translation emphasise the need to further understand the fundamental biology of these RNA species and the unique avenues for application in therapy and biotechnology.

This work represents the development of a new state-of-the-art methodology, mim-tRNAseq, for global tRNA transcriptome analysis in eukaryotes. We show that previous barriers to library generation and computational analysis are overcome. Moreover, specific focus is placed on the computational methods largely overlooked until now. This work, therefore, details not only the development and application of mim-tRNAseq in a more traditional research format[2], but also the development of a detailed step-by-step protocol and guide for the use of mim-tRNAseq[3]. Furthermore, highlighted are important updates to the mim-tRNAseq computational package, various sources of documentation, version control and code annotation, and a platform for the community to raise issues and request functionality. These kinds of protocols and extensive documentation and coding practices are generally lacking in all of biological research, contributing to the reproducibility crisis and the lack of continued use and maintenance of many computational tools over time[4].

So far, mim-tRNAseq has seen active use in the community, evidenced by the numerous, productive exchanges on the GitHub repository, which have led to error debugging, new functionality implementation, and multiple technical contributions from users and members of the community. Furthermore, the original publication[2] has been referenced extensively, stressing its applicability and performance as a tool for transcriptome-wide tRNA analysis. This level of interest clearly exemplifies the requirement for innovative tools to guide innovative discovery in biology.

Accuracy and precision of quantitative methods

Although multiple methods for quantitation of tRNA pools exist, the underlying statistical nature of such measurements should always be tested and verified. In this regard, various terminology and statistics can be used for describing the nature of the data and how closely it represents the ground truth of the quantity intended to be measured. *Accuracy* and *precision* are sometimes used synonymously in the literature, although, strictly speaking, these two terms refer to different characteristics of measured quantities[5]. *Accuracy* is the closeness of a measurement to the true value of what it is that should be measured, and relates to systematic error or bias. While *precision* is the reproducibility or closeness of measurements from independent trials under similar conditions, and relates to variable error. Therefore, precise measurements can be inaccurate when they do not reflect the true nature of the quantity. Similarly, a method may more accurately estimate a quantity with more variance between measurements, and may therefore be less precise.

Ideally, both high precision and accuracy are desired, but which value makes a method more valid? One can argue that consistent precision, despite the accuracy, may make a method reliable[6], although its measurements may be reliably incorrect. Precision then has downstream implications for statistical tests. For example, in testing differential expression between tRNA transcripts and anticodon pools, sensitivity can be significantly increased when replicates are more precisely measured, resulting in increased detectability of differential expression at low fold-changes. However, if the underlying measurements of tRNA abundance are inaccurate, then how valid and meaningful are the differential expression results anyway?

Classifying measurements as precise is relatively straightforward; various measures of similarity or distance and clustering can be employed. For expression data, a Euclidean distance matrix, hierarchically clustered heatmap of normalized expression, or a PCA or other dimensionality reduction method can very quickly show the preciseness of measurements and similarity within experimental groups. Such quality control is important to show that the method performs similarly time and time again.

Accuracy, on the other hand, is far more difficult to show. Demonstrating the accuracy, or perhaps estimating the extent of error of measured quantities is certainly essential. Orthogonal methods such as Northern blots can validate individual measurements of RNA abundance from sequencing data. Internal controls and spike-in RNA can help investigators understand sources of bias and show the accuracy of the method on a more global scale. Similarly, for computational methods, *in silico* simulated datasets in which RNA abundances are known can help test computational biases. Moreover, extensive testing of wet-lab approaches by analysis with the same computational tools, or testing of different computational tools by analyzing the same input data, can be invaluable in understating unique sources of bias and drawing conclusions about accuracy of each method.

In chapter 2, several of the above methods were employed; *E. coli* tRNA-Lys-UUU oligonucleotides with variable 3' ends are spiked-in directly after RNA isolation in known concentrations to test biases in all subsequent library generations steps, and in alignment and quantitation at the computational level, Northern blots verified the trends seen in differential expression analysis in various human cell types, tRNA from modification-deficient yeast strains were mixed in known concentrations to test the accuracy of modification abundance estimation, and mim-tRNAseq was extensively tested against other methods, both in the lab and computationally. Such rigorous testing of both accuracy and precision is crucial in method development, and should be discussed extensively along with known limitations and remaining biases to ascertain the validity of the method.

Perspectives and outlook

The utility of the unprecedented resolution, accuracy, and precision of mim-tRNAseq has multiple applications in research and discovery. Utilizing the comprehensive nature of mim-tRNAseq, investigators can be informed when implementing interventions and therapeutic innovations that require a careful understanding of the balance among tRNA anticodon pools.

Indeed, a recent study utilized mim-tRNAseq to assess potential perturbations to tRNA homeostasis following suppressor tRNA (sup-tRNA) administration in a mouse model[7]. The authors show that successful readthrough of the premature termination codon (PTC) by the sup-tRNA mitigated the disease phenotype, while mim-tRNAseq results showed that relative abundances of tRNAs at the anticodon level remain unchanged, and charging efficiency of the parental isodecoder remains highly efficient in the presence or absence of treatment with the sup-tRNA[7].

The mim-tRNAseq package allows the simultaneous analysis of multiple characteristics of the tRNA pool, not only abundance and aminoacylation levels. The study of modifications in all RNA types has seen increased interest in research and biotechnology recently[8–11], with many studies aiming to understand if modifications change upon stress, differentiation, or within disease context, how these changes are elicited, and the effect of such changes. mim-tRNAseq implements transcript-level resolution analysis of modification stoichiometry. We validate these measurements for select modifications, showing their accuracy. Of course, mim-tRNAseq is limited to Waston-Crick face modifications, and validation of modification abundance has not been carried out on an extensive set of modifications either. However, one promising avenue for development would be to extend mim-tRNAseq capability to analyze modifications other than those in tRNA, such as mRNA and rRNA.

In the biotechnology space, many companies are now interested in RNA modifications as targets for human health and disease. Storm therapeutics (www.stormtherapeutics.com), for example, aims to use small molecule inhibitors against RNA-modifying enzymes. In this way, alterations to the epitranscriptome may act as therapeutics agents – a treatment method already demonstrated for myeloid leukemia[12]. Storm therapeutics is currently using high-throughput mass-spectrometry for the detection and quantitation of RNA modifications in different samples[13]. However, issues of sensitivity, cost, and difficulty in retaining sequence context information with mass spectrometry-based methods may implicate more sensitive, cheaper, and higher resolution sequencing-based methods such as mim-tRNAseq as effective complementary methods to aid in the examination of the epitranscriptome.

Single-cell RNA sequencing (scRNA-seq) presents another exciting area of research that is currently under active development, particularly with regards to computational analysis[14]. In this approach, RNA from single cells is sequenced to generate gene expression profiles from each cell, removing the masking effect of gene expression dynamics in subpopulations of cells typically present in bulk RNA sequencing. This technology holds incredible promise for unprecedented spatial and temporal resolution of RNA abundance dynamics at the individual cell level, giving investigators insight into the subtle differences that underlie cell-type cell heterogeneity.

In the future, scRNA-seq can be further improved to overcome some remaining hurdles. Limited cellular RNA from individual cells makes measuring low abundance transcripts challenging, and generally leads to high uncertainty in quantitative measurements[15]. Amplification of such material, for example with PCR, which is common in RNA-seq library generation, adds noise to the data. Moreover, adding more resolution to experiments increases dimensionality of the resulting data matrices, which require more complex and scalable models and analysis frameworks[14,15]. However, as solutions to these challenges are developed, single-cell sequencing for tRNA becomes an enticing prospect to increase the resolution and our understanding of complex regulation of eukaryotic tRNA pools. Although recent studies have begun to use single-cell technology to study tRNA regulation, such as querying chromatin state around tRNA genes in mouse and human tissues with single-cell ATAC-seq (scATAC-seq)[16], single-cell measurements of tRNA abundance have not yet been performed, but surely will provide insights into many unanswered questions regarding tRNA biology and translation control.

References

1. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–257 (1958).
2. Behrens, A., Rodschinka, G. & Nedialkova, D. D. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol. Cell* **81**, 1–14 (2021).
3. Behrens, A. & Nedialkova, D. D. Experimental and computational workflow for the analysis of tRNA pools from eukaryotic cells by mim-tRNAseq. *STAR Protoc.* **3**, 101579 (2022).
4. Turkyilmaz-van der Velden, Y., Dintzner, N. & Teperek, M. Reproducibility Starts from You Today. *Patterns* **1**, (2020).
5. Eisenhart, C. Expression of the uncertainties of final results. *Science (80-.).* **160**, 1201–1204 (1968).
6. Stallings, W. M. & Gillmore, G. M. A note on "accuracy" and "precision". *J. Educ. Meas.* **8**, 127–129 (1971).
7. Wang, J. *et al.* AAV-delivered suppressor tRNA overcomes a nonsense mutation in mice. *Nat. 2022 6047905* **604**, 343–348 (2022).
8. Frye, M., Harada, B. T., Behm, M. & He, C. RNA modifications modulate gene expression during development. *Science (80-.).* **361**, 1346–1349 (2018).
9. Suzuki, T. The expanding world of tRNA modifications and their disease relevance. *Nat. Rev. Mol. Cell Biol.* 1–18 (2021) doi:10.1038/s41580-021-00342-0.
10. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).
11. Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nat. Rev. Cancer 2020 206* **20**, 303–322 (2020).
12. Yankova, E. *et al.* Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. *Nature* **593**, 597–601 (2021).
13. Wein, S. *et al.* A computational platform for high-throughput analysis of RNA sequences and modifications by mass spectrometry. *Nat. Commun.* **11**, (2020).
14. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods 2021 187* **18**, 723–732 (2021).
15. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol. 2020 211* **21**, 1–35 (2020).
16. Gao, W., Gallardo-Dodd, C. J. & Kutter, C. Cell type-specific analysis by single-cell profiling identifies a stable mammalian tRNA-mRNA interface and increased translation efficiency in neurons. *Genome Res.* **32**, 97–110 (2022).