MLC-Seq: de novo Sequencing of Full-Length tRNAs and Quantitative Mapping of Multiple RNA Modifications

Shenglong Zhang (szhang21@nyit.edu)  
New York Institute of Technology  https://orcid.org/0000-0001-8003-7484

Xiaohong Yuan  
New York Institute of Technology  https://orcid.org/0000-0002-9165-5197

Yue Su  
New York Institute of Technology

Xudong Zhang  
University of California Riverside  https://orcid.org/0000-0002-5964-0067

Spencer Turkel  
New York Institute of Technology

Shundi Shi  
Columbia University

Xuanting Wang  
Columbia University

Eun-Jin Choi  
University of Texas Medical Branch

Wenzhe Wu  
University of Texas Medical Branch

Haichuan Liu  
University of California San Francisco

Rosa Viner  
Thermo Fisher Scientific (United States)  https://orcid.org/0000-0003-0550-5545

James Russo  
Columbia University

Wenjia Li  
New York Institute of Technology  https://orcid.org/0000-0001-6059-6422

Xiaoyong Bao  
University of Texas Medical Branch

Qi Chen  
University of California, Riverside  https://orcid.org/0000-0001-6353-9589
MLC-Seq: de novo Sequencing of Full-Length tRNAs and Quantitative Mapping of Multiple RNA Modifications

Xiaohong Yuan\textsuperscript{1a}, Yue Su\textsuperscript{1a}, Xudong Zhang\textsuperscript{2}, Spencer J. Turkel\textsuperscript{1a}, Shundi Shi\textsuperscript{3}, Xuanting Wang\textsuperscript{3}, Eun-Jin Choi\textsuperscript{4a}, Wenzhe Wu\textsuperscript{4a}, Haichuan Liu\textsuperscript{5}, Rosa Viner\textsuperscript{5}, James J. Russo\textsuperscript{3}, Wenjia Li\textsuperscript{1b}, Xiaoyong Bao\textsuperscript{4a, 4b, 4c, 4d}, Qi Chen\textsuperscript{2}, Shenglong Zhang\textsuperscript{1a*}

\textsuperscript{1a}Department of Biological and Chemical Sciences, \textsuperscript{1b}Department of Computer Science, New York Institute of Technology, New York, NY 10023, USA
\textsuperscript{2}Division of Biomedical Sciences, School of Medicine, University of California, Riverside, Riverside, CA 92521, USA
\textsuperscript{3}Department of Chemical Engineering, Columbia University, New York, NY 10027, USA
\textsuperscript{4a}Department of Pediatrics, \textsuperscript{4b}Sealy Center for Molecular Medicine, \textsuperscript{4c}the Institute of Translational Science, \textsuperscript{4d}the Institute for Human Infections &Immunity, University of Texas Medical Branch, Galveston, TX 77555, USA
\textsuperscript{5}Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, CA 95134, USA

Abstract

Despite the extensive use of next-generation sequencing of RNA, simultaneous sequencing and quantitative mapping of multiple RNA modifications remain challenging. Herein, we develop MLC-Seq, a mass spectrometry-based direct sequencing method allowing for simultaneously unravelling the RNA sequences and quantitatively mapping different tRNA nucleotide modifications site-specifically. Importantly, MLC-Seq reveals the stoichiometric changes of tRNA modifications upon treatment with the dealkylating enzyme AlkB, and led to the discovery of new nucleotide modifications.
Despite wide application of high-throughput next generation sequencing (NGS), the true sequence of a RNA, i.e., identity and location of each and every nucleotide building block (canonical or modified) within a full-length RNA, remains a mystery\(^1\), mainly because of the lack of a general method to directly sequence any nucleotide (including unknown ones) at single-nucleotide resolution. tRNAs, where >100 different modifications have been discovered, cannot be directly sequenced efficiently. NGS-based methods, which require cDNA synthesis, are generally unable to sequence modifications, and can only be “tailored” to sequence a specific RNA modification type\(^2\). Direct nanopore-based RNA sequencing has been used for sequencing tRNAs\(^3\), but currently suffers from a high error rate and cannot identify or pinpoint diverse tRNA modifications. Also, not all tRNA modifications are modified 100\% of the time, so that site-specific quantification of their stoichiometries is challenging\(^4\).

Mass spectrometry (MS), in particular combined liquid chromatography and tandem mass spectrometry (LC-MS/MS), is recognized as the ‘gold standard’ for RNA modification analysis, because it does not require a cDNA synthesis step and is not limited to specific known modification types. However, LC-MS/MS analysis of RNA modifications is often restrained to the ribonucleoside level. For example, RNA samples may be completely digested, enzymatically or chemically, to single nucleosides for MS analysis\(^5-7\), but location information for the RNA modifications is lost. To obtain RNA sequence information, LC-MS/MS-based mapping methods rely on other well-established complementary methods (e.g., previous analyses or NGS)\(^8\), and can only analyze short RNA oligonucleotides (~17 nt)\(^1\), due to spectral complexity and resulting difficulty in data interpretation\(^6\). Therefore, they are rarely used for de novo (without prior sequence information) sequencing of RNA.

LC-MS-based de novo sequencing methods have nevertheless been developed that simultaneously sequence and quantify nucleotide modifications\(^9-12\). These methods rely on a complete set of MS ladders that are produced not by MS/MS, but by exactly one random and unbiased cut on each RNA strand via controlled enzymatic or chemical degradation\(^9,13-15\). The resulting MS ladder fragments retains the nucleotide modifications from the original parental RNAs, allowing identification, quantification and location of each nucleotide modification in a sample. However, each MS ladder must be perfect, i.e., without any missing fragments, in order to read all nucleotides in an RNA strand\(^9,11,16\). These rigorous sample preparation requirements, together with the need for higher sample loading (~1000 pmol)\(^12\) for a tRNA, short read length (<35 nt per run)\(^11\), and low throughput (a single RNA strand or just a few)\(^9,11\), have restricted MS-based de novo sequencing applications to less complex RNA samples, making it difficult to sequence and quantify nucleotide modifications from limited amounts of tRNAs enriched from cells.

To systematically address the drawbacks of previous MS-based methods, we developed a de novo MS ladder complementation sequencing approach (MLC-Seq) (Fig. 1a) that circumvents the perfect ladder requirement for direct MS sequencing, thereby allowing de novo MS sequencing of full-length cellular tRNAs, together with all nucleotide modifications, at single-nucleotide stoichiometric precision. To test the power of the ladder complementation strategy for sequencing of modification-rich tRNAs, we first sequenced a yeast tRNA-Phe sample containing five isoforms (Fig. 1b-c), which have similar sequences but differ in as little as one nucleotide or modification. None of the five isoforms yielded a perfect MS ladder (Fig. 1b) (likely due to
inefficient LC-separation or MS measurement of minor ladder fragments from low-abundance tRNA isoforms), and thus could not be sequenced by previously developed MS sequencing methods. However, after ladder complementation using our MLC-Seq method, the missed MS ladder fragments were corrected to generate a perfect MS ladder, allowing for direct sequencing of all five tRNA isoforms and pinpointing each nucleotide or modification difference quantitatively (Fig. 1c and Fig. S1) (see Methods for details).

MLC-Seq achieves this goal by incorporating a series of systematic and innovative technical advances over previous de novo MS sequencing methods (Extended Data Fig. 1). First, homology search identifies RNA isoforms related to a specific tRNA in a given sample (Extended Data Fig. 2). Second, MassSum, a novel MS-based algorithm, takes advantage of the unique mass of each RNA molecule and site-specific cleavage of RNA phosphodiester bonds under controlled acid degradation conditions, allowing identification and computational isolation of ladder fragments of the RNA in a mixed RNA sample without end-labeling (Extended Data Fig. 3), which was previously required to MS sequence mixed RNA samples. Third, MS ladder complementation allows sequencing of low-abundance RNA samples that cannot provide a perfect MS ladder and thus cannot be sequenced by previous MS sequencing methods (Fig. 1b). Fourth, with the latest Orbitrap MS instrument, MLC-Seq increases read length of de novo MS-based RNA sequencing from ~35 nt to ~80 nt per run, allowing sequencing of the full-length tRNA without T1 digestion, which was required by previous MS sequencing methods, but complicated the sequencing and data analysis. Finally, lowering sample loading ~1000-fold down to ng scale (~20 ng or ~1 pmol for a tRNA), allows sequencing of cellular tRNAs.

MLC-Seq allows direct sequencing of full-length tRNAs without a cDNA intermediate and preserves the sample diversity and modification information in a given tRNA sample that indirect cDNA-based RNA sequencing methods cannot (Fig. 1a). After verifying the sequencing method on yeast tRNA-Phe, we used MLC-Seq to sequence other tRNAs enriched from mouse liver. The identity, position, and stoichiometry of each modification in wild-type tRNA-Gln and tRNA-Glu is shown in Fig. 2a and Extended Data Fig. 4b, respectively. As such, MLC-Seq allows direct and de novo sequencing of full-length tRNAs as well as quantitative mapping of all nucleotide modifications of each specific tRNA enriched from yeast or mice at single-nucleotide resolution, for which no current high-throughput sequencing methods are capable.

Importantly, our method led to the discovery of new RNA modifications, a 3,4-dihydrocytidine (C’) at position 16 of tRNA-Gln (Fig. 2b) and a 2-Ox-G (G’) at position 17 of tRNA-Glu (Extended Data Fig. 4c), which have not been reported before. MLC-Seq showed that when reading the 5' ladder of the tRNA-Gln in the two dimensional (2D) mass-retention time (tR) plot, a new ladder started to branch out at position 16, and the mass differences of ladder fragments between position 16 and 15 are 307.0276 Da and 308.0340 Da, indicating a C’ co-existing with D (dihydrouridine) at position 16 (D/C’: 79% vs. 21%). We propose a structure for this newly discovered nucleotide C’ (Fig. 2bII), which contains two more hydrogens in positions 3 and 4 of C, based on reported structural stability analysis of the isomeric C’17. Similarly, a new sequence branch indicates a G’ co-existing with G at position 17 of tRNA-Glu (G/G’: 69% vs. 31%, AlkB-treated) (Extended Data Fig. 4c).
Notably, to differentiate isomeric methylations such as $m^1A$ and $m^6A$, we leveraged the
demethylation specificity of AlkB on its target methylation types (e.g., $m^1A$, $m^1G$, and $m^3C$)\textsuperscript{18,19} and further successfully distinguished these methylations according to their sensitivities to AlkB
(Fig. 2 and Extended Data Fig. 4); this experiment has also in turn, demonstrated the accuracy of
MLC-seq to track the stoichiometric changes of AlkB-sensitive methylated nucleotides in
various tRNA samples such as tRNA-Gln and tRNA-Glu extracted from mouse liver. MLC-Seq
results pinpoint stoichiometric changes of methylations site-specifically in tRNA-Gln after AlkB
treatment. In wild type tRNA-Gln, $m^1A$ at position 57 of tRNA-Gln occurred at a frequency of
100% (no canonical A co-existing). However, 99% of the $m^1A$ at position 57 was demethylated
into A after AlkB-treatment. Reading the 3′ ladder of the wild-type tRNA in the 2D mass-tR
plot indicates an $m^1A$ (100%) at position 57 of the tRNA (Fig. 2c). However, after AlkB-treatment,
MLC-Seq showed that when reading the 3′-ladder, a ladder started to branch out at position 57,
and that 99% A was observed at this position with only 1% $m^1A$ remaining. In wild type tRNA-
Glu, $m^1A$ at position 57 occurred at a frequency of 68% (vs 32% A), but all $m^1A$ demethylated
into A after AlkB treatment (Fig. S2-3). Similarly, in wild type tRNA-Gln, only $m^1G$ was found
at position 9 (no co-existing canonical G). However, only 45% of the $m^1G$ was demethylated
into G after AlkB-treatment (Fig. S4-5).

The full potential of the method’s sequencing read length and throughput remains to be explored,
especially when used with state-of-the-art LC-MS instruments and their built-in capacity for
automation. MLC-Seq, especially MassSum, increases the throughput of \textit{de novo} MS-based
RNA sequencing, allowing sequencing potentially of an unlimited number of RNA sequences in
complex RNA samples, to the extent permitted by MS instruments, and paves the way for large
scale \textit{de novo} MS sequencing of biological samples. It is worth noting that the methods described
here could be used to sequence other modified longer RNAs (e.g., rRNA, snoRNA, snRNA, Y
RNA, and vault RNA) and small non-coding RNAs (e.g., miRNA, piRNA, tsRNA, rsRNA, and
ysRNAs), and as an orthogonal approach to verify RNA sequences or modifications determined
by high-throughput sequencing methods. Thus, MLC-Seq will provide a general sequencing tool
for quantitatively studying RNA modifications, which is urgently needed more than ever,
considering for example that >40 unidentified nucleotide modifications have been reported in
SARS-CoV-2 RNA\textsuperscript{20}, though their identities and functions remain unknown.

\section*{Data and Code Availability}

All MS sequence datasets used in the manuscript are publicly available through the
corresponding repository on Github (https://github.com/rnamodifications/MLC-Seq).

Source codes implemented in Python of all the algorithms described in the manuscript (including
homology search, identifying acid-labile nucleotides, MassSum data separation, GapFill, ladder
complementing) are freely available on Github (https://github.com/rnamodifications/MLC-Seq)
and are distributed under the MIT open source license.

\section*{Acknowledgments}

The authors acknowledge grant support from the USA National Institutes of Health R21
HG009576, R56 HG011099, and RM1 HG011563 (Co-I) to S.Z., R01 HD092431 and R01
ES032024 to Q.C., and R01 AI116812 and R21 AI113771 to X.B.
Conflict of Interest

The authors have filed a patent related to the technology discussed in this manuscript. H.L. and R.V. are employees of Thermo Fisher Scientific, San Jose, CA 95134, USA
### References:

<table>
<thead>
<tr>
<th>Number</th>
<th>Author(s)</th>
<th>Title</th>
<th>Journal/Citation Details</th>
</tr>
</thead>
</table>


**Fig. 1** | MLC-Seq allows direct sequencing and quantitative mapping of all tRNA modifications within their full-length sequence context at single-nucleotide resolution.

a, The general workflow of MLC-Seq of cellular tRNA samples. A given tRNA sample is first divided into two portions; one, serving as a control, is injected directly into LC-MS to provide mass and relative abundance information of all the intact tRNAs (I), while the rest is subjected to the acid-degradation procedure before LC-MS to provide information including mass and related retention time (t<sub>r</sub>) for tRNA ladder fragments (II). Partial nucleotide modifications or editing can be first observed at the intact level and then pinpointed site-specifically at the ladder fragment level. For example, at the intact level from the control sample, partial RNA modification/editing such as a methylation (14 Da difference) and A-G editing (16 Da difference) can be observed directly between the intact masses of the related tRNA isoforms (I). After acid degradation and LC-MS, 5′- and 3′-ladders consisting of ladder fragments with various lengths corresponding to nucleotides from the first to the last position on an RNA molecule, are produced and form sigmoidal curves in a Mass-t<sub>r</sub> plot at the ladder fragment level (II). At this level, the specific site
of the above-mentioned modification or editing can be further pinpointed at single-base
resolution. For example, at position 48 and 67, a second branch appears with a mass difference
of 14 Da and 16 Da between two ladder fragments at the same position (1st nucleotide after the
first branch site in the 2D Mas-tr plot), indicating the partial modification being A/mA and the
partial editing A/G occurring at the respective locations. Stoichiometric quantification of partial
RNA modification/editing can also be calculated based on the relative intensities of the
corresponding ladder fragments in their branches. Finally, combining with LC-MS datasets of
the controlled and acid-degraded portions, novel algorithms have been developed to identify
each tRNA species or isoform, and computationally separate its MS ladders from the LC-MS
data of a given tRNA sample for MLC-Seq (Extended Data Figs. 1-2 and Method Section). This
panel is for illustrative purposes only and does not represent any real tRNA sequence position
and stoichiometry of partial modification/editing. b, Ladder complementation of various related
tRNA isoforms allows direct sequencing that previous MS sequencing methods could not. For
example, a yeast tRNA-Phe containing five isoforms, but none of them has a perfect ladder
required by previous MS-based sequencing methods. However, after ladder complementation
using our MLC-Seq method, the missed MS ladder fragments were corrected to generate a
perfect MS ladder, allowing for direct sequencing of all five tRNA isoforms. All 3´- ladders were
converted to 5´- to keep the positions consistent. Depth on the bottom illustrates the number of
times each nucleotide was read from tRNA isoform ladders (3´ and 5´-ladders). c, Sequences of
tRNA-Phe isoforms, showing 100% coverage and accuracy as compared to the reported
tRNA-Phe reference sequence. For isomeric nucleotides or modifications such as pseudouridine
(ψ) versus U and methylation at different base positions, an extra step is required to differentiate
them following our MS sequencing approach as described previously (Ref. 11).

*The missing ladder fragments at positions 32 and 34 are due to methylations on the 2´-hydroxyl
group of Cm and Gm that block acid degradation.
**Fig. 2**  
**a**, MLC-Seq of tRNA-Gln extracted from mouse liver. I, MLC-Seq results for wildtype tRNA-Gln, showing all the RNA modifications within their full-length tRNA sequence context as well as the stoichiometry of each modification site-specifically. II, Identity, position, and stoichiometry of each modification in wild-type tRNA-Gln. mU: a methylated U, but not Um (2′-O-methyluridine). III, Stoichiometric changes of 2 methylated nucleotides after AlkB-treatment. The percentage of m'G at position 9 was reduced from 100% to 55% (with 45% G co-existing at this position), and the percentage of m'A at position 57 was reduced from 100% to 1% (with 99% A co-existing at this position), respectively. 

**b**, MLC-Seq leading to the discovery of a novel nucleotide of C' (3,4-dihydrocytidine) at position 16 of tRNA-Gln. I, Structure and unit mass of nucleotide D in MLC-Seq of RNA (exact unit mass: 308.0410 Da). II, Proposed structure of a newly discovered nucleotide C' (exact unit mass: 307.0569 Da) at position 16, which contains two more hydrogens at positions 3 and 4 of nucleotide C. III, MLC-Seq showing that when
reading the 5' ladder of the tRNA in the 2D mass-tr plot, a new ladder started to branch out at position 16, and the mass differences of ladder fragments between position 16 and 15 are 307.0276 Da and 308.0340 Da, indicating a C' 3,4-dihydrocytidine co-existing with D (dihydouridine) at position 16. c, MLC-Seq results pinpoint stoichiometric changes of m1A site-specifically at position 57 of tRNA-Gln after AlkB treatment. I, Structure and unit mass of m1A and canonical A in MLC-Seq of RNA. Their exact masses are 343.0682 Da and 329.0525 Da, respectively. II-III, 99% m1A at position 57 was demethylated into A after AlkB-treatment. Reading the 3' ladder of the wild-type tRNA in the 2D mass-tr plot indicates an m1A (100%) at position 57 of the tRNA (II). However, after AlkB-treatment, MLC-Seq shows that when reading the 3'-ladder of the tRNA in the 2D mass-tr plot, a ladder started to branch out at position 57, and that 99% A was observed at this position with only 1% m1A remaining (III).
Extended Data

**Extended Data Fig. 1** | Detailed illustration for “Data Analysis” procedure in Fig. 1. a, Homology Search before acid degradation. From the intact RNA mass level, the sample diversity and sequence information are obtained, showing the number of RNA species in a given sample and if and how they related to a specific tRNA. I-II, Homology Search allows related tRNAs to be identified. After acid degradation, in order to sequence tRNA, we developed b, A complete set of stepwise innovative computational tools/algorithms to synergize MS information at both ladder fragment level and intact RNA level, including: I-II, Homology Search, in this step, isoform1 with masses of 24610.4911 Da before acid degradation, for example, shifted to 24252.3692 Da after acid degradation with a mass difference of 358.1219 Da, indicating that there are acid-labile nucleotide modifications, which was further pinpointed to be a wybutosine (Y) at position 37 of tRNA-Phe (Extended Data Fig. 2). III, For each isoform identified, MassSum algorithm was applied to isolate all its ladder fragments (pairing one in 3’-ladder and one in 5’-ladder) from the complex MS data of the RNA sample (Extended Data Fig. 3). IV, Results using GapFill to find non-paired ladder fragments that are missed by MassSum. After ladder separation, 3’ and 5’-ladder (V) are obtained separately. VI, Ladder Complementation to perfect MS ladders, allowing direct sequencing of full-length tRNA-Phe. (See methods) This figure is for illustrative purposes only.
Extended Data Fig. 2 | MLC-Seq homology search for identification of related RNA isoforms and RNAs with acid labile nucleotide modifications. a, The identification of acid-labile nucleotide Y. At intact mass region (around 24kDa), masses before and after acid degradation were compared, revealing a systematic mass shift of 358.1599 Da. This is consistent with a change of Y to ribose Y’ under acidic conditions. b, Homology Search result before acid degradation. Here C represents that the difference between two related tRNA isoforms is a nucleotide C. A represents that the difference between two related tRNA isoforms is a nucleotide A. Ox represents that the difference between two related tRNA isoforms is an oxygen atom. Their mass differences, 305.0413, 329.0525 and 15.9949, correspond to a unit mass of nucleotide C and A, and an oxygen atom. c, Proposed structural change of Y to Y’ under the acid-degradation conditions. Acid-Deg=Acid-Degradation.
Extended Data Fig. 3 | MassSum strategy and MassSum-based computational data separation. **a**, A purified or mixed RNA starting material is partially acid hydrolyzed under controlled conditions that predominantly generates single-cut fragments. Taking a 9 nt RNA strand as an illustrative example, two ladder fragments are generated as a result of acid-mediated cleavage of the phosphodiester bond between the 1st and 2nd nucleotide of the 9 nt RNA strand. One of them carries the original 5´-end of the RNA strand and has a newly-formed ribonucleotide 3´ (2´)-monophosphate at its 3´-end (denoted F1). The other one carries the original 3´-end of the RNA strand and has a newly-formed hydroxyl at its 5´-end (denoted T8).

**b**, The mass sum of any one-cut fragment pair, *e.g.*, mass sum of F2 and T7 or of F1 and T8, is constant and is equal to the mass of the 9 nt RNA plus the mass of a water molecule. Since the mass sum is unique to each RNA sequence/strand, it can be used to computationally separate all paired fragments of the RNA sequence/strand that were inextricable previously in the complex MS datasets.

**c**, Computational isolation of MS data for all ladder fragments derived/degraded from the most abundant tRNA-Phe isoform (monoisotopic mass: 24252.2692 Da) in both the 5´- and 3´-ladders out of the complex MS data of mixed samples with multiple distinct RNA strands. It is then identified as the 75 nt PheIF1 with MLC-Seq.
Extended Data Fig. 4 | Ladder complementation and MLC-Seq result for tRNA-Glu extracted from mouse liver. 

**a**, Ladder complementation of 6 related tRNA isoforms allows direct sequencing of a wild-type tRNA-Glu sample enriched from mouse liver. Before ladder complementation, each of the six isoforms of tRNA-Glu could not provide a perfect MS ladder and thus could not be sequenced by previous MS methods. After ladder complementation, the missing MS ladder fragments were filled in, allowing for direct sequencing of all 6 tRNA isoforms (Fig. S3). All 3′-ladders were converted to 5′- to keep the positions consistent.

**b, MLC-Seq of tRNA-Glu extracted from mouse liver.**

**I.** MLC-Seq results of wild-type tRNA-Glu, showing all the RNA modifications within their full-length tRNA sequence context as well as the stoichiometry of each modification site-specifically. 

**II.** Identity, position, and stoichiometry of each modification in wild-type tRNA-Glu.

**III.** Stoichiometric changes of m^1^A after AlkB-treatment. There was 68% m^1^A at position 57 (with 32% A) in the wild-type tRNA-Glu. All the
m¹A is demethylated to become A at position 57. c, MLC-Seq leading to the discovery of a novel G’ nucleotide at position 16 of tRNA-Glu. **I, Structure and unit mass of nucleotide G in MLC-Seq of RNA (exact unit mass: 345.0474 Da).** II, Proposed structure of the newly discovered nucleotide G’ (exact unit mass: 346.0314 Da) at position 16, where an -NH is substituted by -O (see partial structure in red). **III, MLC-Seq showing that when reading the 5’ ladder of the tRNA in the 2D mass-tr plot, a new ladder started to branch out at position 16, and the mass differences of ladder fragments between position 16 and 17 are 345.0515 Da and 346.0589 Da, indicating a G co-existing with G’ at position 16.

*The missing ladder fragment at position 53 is due to the methylation on the 2’-hydroxyl group of m⁵Um at position 53 that blocks acid degradation (Fig. S2a).#

At position 17, the ratio presented is based on the dataset that provides the most continuous ladder fragments. However, this ratio is not uniformly supported by all the datasets. m₂C (a dimethylcytidine) at position 48 was determined from the AlkB-treated sample, not the wild-type sample due to the low abundance of ladder fragments for the latter (Fig. S2b).
Methods

Reagents and Chemicals

All chemicals were purchased from commercial sources and used without further purification. tRNA (phenylalanine specific from brewer’s yeast) was obtained from Sigma-Aldrich (St. Louis, Missouri, USA). Formic acid (98–100%) was purchased from Merck (Darmstadt, Germany). 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP, 99%) and N,N-diisopropylethylamine (DIPEA) (99%) were purchased from Thermo Fisher Scientific (Waltham, MA, USA). All other chemicals were obtained from Sigma-Aldrich unless indicated otherwise.

Mouse liver tRNA pulldown

The tRNAs were purified by an affinity pulldown assay combined with gel recovery, with modified protocols from the previous report\(^1\). The total RNA of mouse liver was harvested by TRIzol\textsuperscript{TM} reagent (Invitrogen\textsuperscript{TM} 15596026) as the manufacturer instructed. The concentration of total RNAs solution was adjusted to 2mg/ml by RNase-free water. Small RNAs fraction (<200 nt) were separated by the buffer containing 50%(w/v) PEG 8000 and 0.5M NaCl solution via centrifuging at 12000rpm and 4 °C for 20mins. The supernatant was collected followed by adding 1/10 volume NaAc solution (Invitrogen\textsuperscript{TM}). 1 ml supernatant was added with 3 ml Ethanol, and 5 ul Linear Acrylamide (Invitrogen\textsuperscript{TM}) to precipitate small RNAs (<200 nt) with -20 °C overnight incubation followed by centrifugation at 12000rpm at 4 °C for 20 min. Small RNA (<200 nt) solution was adjusted to 1mg/ml, 1 ml small RNA solution with 6 ul biotinylated probe(100 uM), 26 ul 20× SSC solution (Invitrogen\textsuperscript{TM}) and 15 ul RNase inhibitor (NEB) were incubated at 50°C overnight. 200 ul Streptavidin Sepharose (Cytiva 17511301) was added to the hybridization solution to enrich the biotin labeled probe that captured with the targeted tRNA. After incubation at room temperature for 30mins, the Streptavidin Sepharose was transferred to the 1.5 ml Ultrafree-MC tube (Millipore) and washed by 0.5× saline-sodium citrate (SSC) buffer, the washing step was repeated five times. 500 ul nuclease-free water was added to the MC tube and incubated at 70°C for 15 min followed by centrifugation at 2500g at room temperature for 1 min to elute the RNAs that are complementary to the biotinylated probe. The eluent was collected followed by adding 1/10 volume NaAc solution (Invitrogen\textsuperscript{TM}). 1 ml eluent was added with 3 ml Ethanol, and 5 ul Linear Acrylamide (Invitrogen\textsuperscript{TM}) to precipitate RNAs with -20 °C overnight incubation followed by centrifugation at 12000rpm at 4 °C for 20 min. Nuclease-free water was added to dissolve the RNAs pallets. RNAs were loaded into the 7M Urea-PAGE gel to run electrophoresis, the main tRNA band was recovered from the PAGE gel as previously described\(^2\) to obtain enriched tRNAs for MLC-Seq. The DNA probes for pull down experiments were synthesized by IDT and the sequences were listed below:

- tRNA-Glu pulldown probe: 5´-Biotin-CTAACCACTAGACCACCAGGGA.
- tRNA-Gln pulldown probe: 5´-Biotin-TGGAGGTTCACCAGGAGGAG.

Northern Blot

A Northern Blot was performed as previously described\(^3\) to validate capture of tRNAs (as described above). RNA was separated by 10% Urea-PAGE gel stained with SYBR Gold, and immediately imaged, then transferred to positively charged Nylon membranes (Roche), and UV crosslinked with an energy of 0.12 J. Membranes were pre-hybridized with DIG Easy Hyb solution (Roche) for 1h at 42 °C. To detect tRNAs, membranes were incubated overnight (12–16h) at 42 °C with DIG-labelled oligonucleotides probes synthesized by Integrated DNA
Technologies (IDT). The membranes were washed twice with low stringent buffer (2× SSC with 0.1% (wt/vol) SDS) at 42 °C for 15 min each, rinsed twice with high stringent buffer (0.1× SSC with 0.1% (wt/vol) SDS) for 5 min each, and rinsed in washing buffer (1× SSC) for 10 min. Following the washes, the membranes were transferred into 1× blocking buffer (Roche) and incubated at room temperature for 3 h, after which the Anti-Digoxigenin-AP Fab fragments (Roche,) was added into the blocking buffer at a ratio of 1:10,000 and incubated for an additional 30mins at room temperature. The membranes were washed four times with DIG washing buffer (1× maleic acid buffer, 0.3% Tween-20) for 15 min each, followed by incubation in DIG detection buffer (0.1 M TrisHCl, 0.1 M NaCl, pH 9.5) for 5 min, and then coated with CSPD ready-to-use reagent (Roche), incubated in the dark for 30 min at 37 °C before imaging with ChemiDoc™ MP Imaging System (BIO-RAD). Digoxigenin labeled Northern blot probe for tRNA detection were synthesized by IDT and the sequence was listed below:
tRNA-Glu Northern blot probe: 5´-DIG-CTAACCACTAGACCACCA.
tRNA-Gln Northern blot probe: 5´-DIG-TGGAGGTTCCACCGAGATT.

Treatement of tRNA with AlkB
200 ng tRNA was incubated in a 50 μl reaction mixture containing 50 mM Na-HEPES, pH 8.0 (Alfa Aesar), 75 μM ferrous ammonium sulfate (pH 5.0), 1 mM α-ketoglutaric acid (Sigma Aldrich), 2 mM sodium ascorbate, 50 μg/ml BSA (Sigma-Aldrich), 2.5 μl RNase inhibitor (NEB), and 200 ng AlkB enzyme at 37 °C for 30 min (the recommended mass ratio of AlkB enzyme to RNA is 1:1). The mixture was added to 500 μl TRIzol™ reagent to perform the RNA isolation procedure as per manufacturer instructions.

Workflow of de novo sequencing of tRNA isoform mixtures
A given tRNA sample is firstly divided into two portions; 1/10 is injected directly into LC-MS while the remaining 9/10 is subjected to acid-degradation before LC-MS injection (Fig. 1). After outputting LC-MS datasets of the control portion and acid-degraded portions, novel algorithms we developed (uploaded to Github (https://github.com/rnamodifications/MLC-Seq)) were used to identify each tRNA species or isoform, and computationally separate its MS ladder from the LC-MS data of a given tRNA sample for MLC-Seq.

Since MS-based sequencing techniques rely on a unique mass value for identifying and locating each nucleotide, in the case where modifications have isomers with identical masses but different chemical structures such as Ψ versus U and methylation at different base positions, an extra step is required to differentiate these isomeric nucleotide modifications following our MS sequencing approach as described previously when sequencing a tRNA-Phe sample. To differentiate different methylations in the tRNA-Gln and tRNA-Glu, we included an additional step in the MLC Seq workflow by treating the tRNA-Gln and tRNA-Glu samples with AlkB, which is known to demethylate tRNA-rich modifications such as m3C, m1A, and m1G2.

Controlled Acid Degradation of tRNA Samples
Formic acid was applied to degrade tRNA samples, including tRNA-Phe (Sigma-Aldrich, St. Louis, Missouri, USA), tRNA-Gln and tRNA-Glu from mouse live (see Sections of tRNA sample preparation), to produce mass ladders according to reported experimental protocols. In brief, we divided each RNA sample solution into three equal aliquots of 5 μl (each in a RNase-free, thin walled 0.2 ml PCR tube) for formic acid degradation, each using 50% (v/v) formic acid
at 40 °C in a PCR machine, with one reaction running for 2 min, one for 5 min, and one for 15 min. The reaction mixtures were each immediately frozen on dry ice at the specified times, followed by centrifugal vacuum concentration (Labconco Co., Kansas City, MO, USA) to dryness, which was typically completed within 30 min. The dried samples of three different time points for each specific tRNA sample were then combined and suspended in 20 μl nuclease-free, deionized water for LC-MS measurement.

**LC-MS Measurement of Intact tRNA and Resulting Acid-Degraded tRNA Samples**

Each combined acid-hydrolyzed tRNA sample was individually analyzed on an Orbitrap Exploris 240 mass spectrometer (ThermoFisher Scientific, Bremen, Germany) coupled to a Vanquish Horizon UHPLC using a DNA Pac reversed-phase (RP) column (2.1 mm x 50 mm, ThermoFisher Scientific, Sunnyvale, California, USA) with 2% HFIP and 0.1% DIPEA as eluent A, and methanol, 0.075% HFIP, and 0.0375% DIPEA as eluent B. A gradient of 20% to 80% B over 6.7 minutes was used for the analysis of intact RNA samples, and from 15% to 35% B over 20 min for acid-degraded samples. The flow rate was 0.2 or 0.4 ml/min, and all separations were performed with the column temperature maintained at 70 °C. Injection volumes were 5–25 μl and sample amounts were 20–200 pmol of tRNA-Phe and 1-10 pmol (or ~20 ng) of tRNA-Glu and tRNA-Gln. tRNA samples were analyzed in negative ion full MS mode from 410 m/z to 3200 m/z with a scan rate of 2 spectra/s at 120k resolution at m/z 200. The data were processed using the Thermo BioPharma Finder 4.0 (ThermoFisher Scientific), and a compound detection workflow with a deconvolution algorithm was used to extract relevant spectral and chromatographic information from the LC-MS experiments as described previously4-7.

**Homology Search**

Once LC-MS data is displayed as a two dimensional (2D) mass-retention time (tR) plot, we start a homology search and match the intact tRNAs in the monoisotopic mass range of >~24k Dalton using an in-house developed algorithm in Python (see Github) to first identify related tRNA isoforms that may share the same ancestral precursor tRNA, but are different in absolute sequence, e.g., in posttranscriptional profiles of nucleotide modifications, editing, and truncations. Mass differences between two intact tRNA isoforms are calculated and matched to the known mass of each nucleotide or nucleotide modification in the database5. For example, known mass differences between intact tRNAs of 14.0157 Da and 329.0525 Da (with parts per million (ppm) difference <10 ppm)8 can be assigned to a methylation (Me-/CH2-) event and an additional A nucleotide, respectively. Therefore, these intact tRNAs are assigned to the same tRNA group and considered as homologous isoforms of a specific tRNA for sequencing together.

The homology search is a non-target pre-selection to group possible related tRNA isoforms together for sequencing. However, only one monoisotopic mass difference of two intact masses is used to identify the tRNA isoforms differing, e.g., by RNA editing/modifications and/or 3′-CCA truncations. Thus, there may be errors when grouping a tRNA isoform that does not belong to this specific tRNA group or vice versa, missing a tRNA isoform when cataloging a group. These errors can be fixed later when sequencing each group of tRNA isoforms, and sequencing results can further verify the inter-connection between isoforms.

**Detection of all acid-labile nucleotide modifications**
Acid-labile nucleotides are identified using another algorithm in Python (see Github). The algorithm analyzes the connections between the compounds (with a monoisotopic mass >24 K Dalton for tRNAs) measured by LC-MS before acid degradation and the LC-MS-detected compounds after acid degradation. For each such a compound pair, if the monoisotopic mass difference can be matched to a known mass difference (calculated from the possible structural change to a specific nucleotide modification during acid hydrolysis), or be matched to the mass difference sum of a subset of different acid-labile nucleotide modifications, the compound pair will be selected and further considered as potentially containing acid-labile nucleotide modifications. In general, if the intact mass of an RNA species does not change after acid degradation, this intact mass will be used for MassSum data separation (See below). Otherwise, an acid-labile nucleotide may be identified by matching to the observed mass difference with the theoretical mass difference caused by acid-mediated structural changes of the nucleotide (See Extended Data Fig. 2).

5´- and 3´-ladder separation
The t_R differences can be used to further computationally separate these two ladders (5´- and 3´-ladders), breaking two adjacent sigmoidal curves into two isolated curves: one for the 3´- and the other for the 5´-ladder. Due to the large number of RNA/fragment compounds, the dividing line between two subsets of 5´- and 3´-ladder fragments is not visually decisive in the 2D mass-t_R plot. Thus, we developed a computational tool (see Github) to separate the 5´- and 3´-fragments. We roughly divide all compounds in each LC-MS data pool into two subgroup areas by circling compounds in the top collective curve of the 2D mass-t_R plot and marking the compounds as 5´-ladder fragment compounds, while the compounds in the bottom curve are marked as 3´-ladder fragment compounds. The purpose of selecting the top area is to include as many 5´-fragment compounds as possible, while minimizing the 3´-fragments in this area. Similarly, the purpose of selecting the bottom area is to include as many 3´-fragment compounds as possible, while minimizing the 5´-fragments in this area. Overlap between two selected ladder subgroups is inevitable, due to limited t_R differences between these two subgroups. The aim of the manual selection step is not to separate the 5´- and 3´-fragments from each other completely, but to serve as two input ladder fragments for MassSum algorithm (See below) to output 5´- and 3´-ladder fragments separately for each tRNA isoform/species.

MassSum data separation
MassSum is an algorithm in Python (see Github) developed based upon the acid degradation principle presented in Extended Data Fig. 3. Taking advantage of the fact that each fragmented pair from two ladder groups (5´- and 3´-ladders) sums to a constant mass value that is unique to each specific tRNA isoform/species, MassSum can isolate ladder compounds corresponding to a specific tRNA isoform. MassSum simplifies the dataset by grouping MS ladder components into subsets for each tRNA isoform/species based on its unique intact mass. Since the well-controlled acid degradation reaction cleaves RNA oligonucleotides at one specific site of the phosphodiester bond (on average, one cut per RNA^5) the masses of two RNA fragments (Mass _3´-portion_ and Mass _5´-portion_) from the same strand sum to a constant value (Mass_sum).

\[
\text{Mass}_{3´-\text{portion}} + \text{Mass}_{5´-\text{portion}} = \text{Mass}_{\text{intact}} + \text{Mass}_{H_2O} = \text{Mass}_{\text{sum}}
\]  

(Taking advantage of this unique mass sum between the paired 3´- and 5´-ladder fragments (Equation 1), the algorithm chooses two random compounds from the acid-degraded LC-MS
dataset and adds their mass values together, one pair at a time. If the sum of the two selected
compounds is equal to a specific Mass_sum, these two compounds will be selected into the pool for
each RNA accordingly. The process repeats until all compound pairs have been inspected. In the
end, MassSum will cluster the dataset into different groups; each group is a subset that contains
3´- and 5´-ladders of one specific RNA sequence.

**GapFill**

GapFill is another Python-based algorithm (see Github) developed as a complement to
MassSum. Since MassSum handles compounds in pairs, in the case that one ladder fragment is
missing, e.g., in the 5´-ladder, the corresponding single-cut ladder fragment, even if it exists in
the 3´-ladder, will not be separated/called by the MassSum algorithm. In order to extract all
ladder fragments from the complex MS data, a GapFill algorithm was designed to “rescue” any
ladder fragments missed by MassSum separation. GapFill identifies a gap that has ladder
fragments missing between ladder fragment compounds, e.g., two adjacent compounds with
Mass values Mass_{5´-i} and Mass_{5´-j} in the 5´-ladder which were found by MassSum algorithm.
Within the gap, there exist many ladder compounds in the degraded LC-MS dataset, but
presumably none were selected by the MassSum algorithm during data separation. GapFill
iterates over each potential compound in the gap in the original LC-MS dataset, and examines
the mass differences of this compound and the two ending compounds, Mass_{5´-i} and Mass_{5´-j}. If
the mass difference is equal to the sum of one or more nucleotides or modifications in the RNA
modification database, we define it as a connection. If the compound in the gap has connections
with both ending compounds, this compound would be selected into a candidate pool for the later
sequencing process. After iteration, GapFill calculates connections of the compounds in the
candidate pool and assigns weights to them based on the frequency of each connection. The
compounds that contain the highest weights would be the ones chosen to fill in the gap.

**Ladder complementation and generation of RNA sequences**

After MassSum and GapFill, each tRNA isoform has its own set of separate 5´-and 3´-ladders
(not combined). Each ladder (5´- or 3´-) consists of a ladder sequence, and we can determine if
these ladders are perfect (without missing any ladder fragments), which would allow reading of
the full RNA sequence (from the first to the last nucleotide in the sequence). If not, we can
complement ladders from related isoforms in order to obtain a more complete ladder for
sequencing (ideally no missing ladder fragment or as complete as possible). A Python-based
computational algorithm (see Github) was designed to align ladders from related isoforms based
on the position of the ladder fragment in the 5´→3´ direction. For example, we lay out each
tRNA-Phe isoform’s full 5´-ladder, e.g., the 5´-ladder in Fig. 1b, on top of each other vertically;
horizontally, we align the 5´- ladder of each isoform according to the position of each
corresponding ladder fragment, ranging from nucleotide position 1 to nucleotide position 76 for
tRNA-Phe. Ladder complementation can be performed separately on 5´- or 3´-ladders separately
(but not mixed ladders), resulting in one final 5´-ladder or one final 3´-ladder. Additionally, all
3´-ladder fragments can be converted to their corresponding 5´-ladder fragments for each tRNA
isoform based on the MassSum principle. As such, each tRNA isoform could have two 5´-ladder
fragments in each position on the 5´-ladder: one original 5´-ladder fragment, and a second ladder
fragment that was converted from its corresponding 3´-ladder fragment, for reaffirmation and/or
complementation.
Stoichiometric quantification of partial nucleotide modifications/editing

Stoichiometries/percentages of partial nucleotide modifications/editing were quantified according to reported protocols\textsuperscript{4,6,7}. In brief, to accurately determine the ratio of partial modification/editing, from datasets of multiple experimental trials (n \( \geq 3 \)) of a given tRNA sample, 3 pairs of ladder fragments (one in the original ladder, and the other in the branched ladder) were taken among the partial modification position (first nucleotide after the branch point where the partial modification was observed) or otherwise its immediate next ladder fragment in the same ladder in case the ladder fragment at the partial modification position does not exist, \textit{e.g.}, due to methylation on the 2'-hydroxyl group of Cm that blocks acid degradation. Mean ratio and standard deviation were calculated.

References:

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingInformationfinal.docx
- zhangepcflat.pdf