Journal of Visualized Experiments

2D-HELS MS Seq: a General LC-MS Based Method for Direct and de novo Sequencing of RNA Mixtures with Different Nucleotide Modifications

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE61281R3
Full Title:	2D-HELS MS Seq: a General LC-MS Based Method for Direct and de novo Sequencing of RNA Mixtures with Different Nucleotide Modifications
Section/Category:	JoVE Biochemistry
Keywords:	Mass spectrometry-based sequencing, direct RNA sequencing, 2-dimensional mass-retention time ladders, hydrophobic end-labeling strategy, RNA modification sequencing, single-base precision.
Corresponding Author:	Shenglong Zhang New York Institute of Technology New York, UNITED STATES
Corresponding Author's Institution:	New York Institute of Technology
Corresponding Author E-Mail:	szhang21@nyit.edu
Order of Authors:	Ning Zhang, Ph.D.
	Shundi Shi
	Barney Yoo
	Xiaohong Yuan
	Wenjia Li
	Shenglong Zhang
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	New York, NY, 10027, USA

TITLE:

2D-HELS MS Seq: a General LC-MS Based Method for Direct and de novo Sequencing of RNA Mixtures with Different Nucleotide Modifications

AUTHORS AND AFFILIATIONS:

Ning Zhang[1,2], Shundi Shi[3], Barney Yoo[4], Xiaohong Yuan[1], Wenjia Li[3], Shenglong Zhang[1]

1. Department of Biological and Chemical Sciences, New York Institute of Technology, New York, NY, USA
2. Department of Chemical Engineering, Columbia University, New York, NY, USA
3. Department of Computer Science, New York Institute of Technology, New York, NY, USA
4. Department of Chemistry, Hunter College, City University of New York, New York, NY, USA

EMAIL ADDRESSES OF CO-AUTHORS:

Ning Zhang (nzhang07@nyit.edu)
Shundi Shi (ss526@columbia.edu)
Barney Yoo (by104@hunter.cuny.edu)
Xiaohong Yuan (xyuan04@nyit.edu)
Wenjia Li (wli20@nyit.edu)

CORRESPONDING AUTHOR:

Shenglong Zhang (szhang21@nyit.edu)

KEYWORDS:

Mass spectrometry-based sequencing, direct RNA sequencing, 2-dimensional mass-retention time ladders, hydrophobic end-labeling strategy, RNA modification sequencing, single-base precision.

SUMMARY:

Here, we describe a detailed protocol for an LC-MS based sequencing method that can be used as a direct method to sequence short RNA (<35 nt per run) without a cDNA intermediate, and as a general method to sequence different nucleotide modifications in a single study at single-base precision.

ABSTRACT:

Mass spectrometry (MS)-based sequencing approaches have been used to directly sequence RNA without the need for a cDNA intermediate. However, they were rarely applied as a de novo RNA sequencing method but used mainly as a tool that can help quality assurance for confirming known sequences of purified single-stranded RNA samples. Recently we reported a direct RNA sequencing method by integrating a 2-dimensional mass-retention time hydrophobic end-labeling strategy into MS-based sequencing (2D-HELS MS Seq). This method is capable of accurately sequencing single-stranded RNA as well as mixtures up to 12 distinct RNA sequences. In addition to the four canonical ribonucleotides (A, C, G, and U), the method has the capacity to

sequence RNA oligos containing modified nucleotides. This is possible because the modified nucleobase either has an intrinsically unique mass that can help to identify and locate it in the RNA sequence or can be converted into a product with a unique mass. Here in this study, we have used RNA, incorporating two representative modified nucleotides, pseudouridine (Ψ) and 5-methylcytosine (m^5C), to illustrate the application of the method for the de novo sequencing of a single RNA oligo, as well as mixture of RNA oligos, each with a different sequence and/or modified nucleotides. The procedures and protocols described here in sequencing these model RNAs will be applicable to other short RNA samples (<35 nt) when using a standard high-resolution LC-MS system. In the future with the development of more robust algorithms and with better instruments, this method could allow sequencing of more complex biological samples.

INTRODUCTION:

Mass spectrometry (MS)-based sequencing methods, including top-down MS and tandem MS[1-4], have been developed for direct sequencing of RNA. However, in situ fragmentation techniques for effectively generating high-quality RNA ladders in mass spectrometers are currently not available for sequencing[5,6]. Furthermore, it is very complicated to analyze the traditional one dimensional (1D) MS data even for de novo sequencing of a purified singe-stranded RNA, and it would be even more challenging for MS sequencing of mixed RNA samples[7,8]. Therefore, a two-dimensional (2D) liquid chromatography (LC)-MS-based RNA sequencing method has been developed and 2D mass-retention time (t_R) ladders are produced to replace 1D mass ladders, making it much easier to identify ladder components needed for de novo sequencing of RNAs[8]. However, the 2D LC-MS-based RNA sequencing method is mainly limited to purified synthetic short RNA as it cannot read a complete sequence solely based on one single ladder, but has to rely on two co-existing adjacent ladders (5´ and 3´ ladder)[8]. More specifically, this approach requires paired-end reads for reading terminal nucleobases. This becomes more complicated in MS sequencing of RNA mixtures because confusion is raised on which ladder fragment belongs to which ladder for the unknown samples.

To overcome the abovementioned barriers in the MS-based sequencing approaches and broaden their applications in direct RNA sequencing, two issues must be addressed, including: 1) how to generate a high-quality mass ladder that can be used to read a complete sequence, from the first nucleotide to the last in the RNA strand; and 2) how to effectively identify each RNA/mass ladder in a complex MS dataset. We have introduced a hydrophobic end labeling strategy (HELS) into the MS-based sequencing, and successfully addressed these two issues by adding a hydrophobic tag at either 5´ and/or 3´ end of the RNAs to be sequenced[9]. This new strategy enables reading a complete RNA sequence from one ladder of a RNA strand without pair-end reading from the other ladder of the RNA, and allows MS sequencing of RNA mixtures with multiple different strands[9]. By adding a tag at the 5´ and/or 3´ end of the RNA, the labeled ladder fragments display a significant delay of t_R, which can help distinguish the two mass ladders from each other and also from the noisy low mass region. The mass-t_R shift caused by adding the hydrophobic tag facilitates mass ladder identification and simplifies data analysis for sequence generation. Also, addition of the hydrophobic tag can help to identify the terminal base due to the mass increase caused by the tag, thus allowing reads for the complete sequence from a single ladder; no paired end reads are required. As a result, we have previously demonstrated the successful sequencing of a complex mixture of up to 12 RNA distinct strands without the use of any advanced sequencing algorithm[9], which opens the door for de novo MS sequencing of RNA containing both canonical and modified nucleotides and makes it more feasible for the sequencing of mixed and more complex RNA samples. In fact, using the 2D-HELS MS Seq, we have even successfully sequenced a mixed population of tRNA samples[10], and we are actively expanding its application to other complex RNA samples.

To facilitate 2D-HELS MS Seq to directly sequence a broader range of RNA samples, here we will focus on the technical aspects of this sequencing approach and will cover all the essential steps needed when using it to directly sequence RNA samples. Specific examples will be used to illustrate the sequencing technique, including synthetic single strand RNAs, RNA mixture of multiple distinct RNA strands, and modified RNAs containing both canonical and modified nucleotides such as pseudouridine (ψ) and 5-methylcytosine (m^{5}C). Since RNAs are all made of phosphodiester bonds, all different kinds of RNAs can be acid hydrolyzed to generate an ideal sequence ladder for 2D-HELS MS Seq under optimal conditions[8,9,11]. However, detection of all the ladder fragments of an RNA is instrument dependent. On a standard high-resolution LC-MS (40K), the minimal loading amount for sequencing purified short RNA sample (<35 nt) is 100 pmol per run. However, more material is required (up to 400 pmol per RNA sample) when additional experiments must be conducted (e.g., to distinguish isomeric base modifications that share identical masses). The protocol used in sequencing the model synthetic modified RNAs will also be applicable to sequencing broader RNA samples, including biological RNA samples with unknown base modifications. However, a larger sample amount, such as 1000 pmol for sequencing tRNA (~76 nt) using a standard LC-MS instrument, is required to sequence the complete tRNA with all the modifications, and an advanced algorithm needs to be developed for its de novo sequencing[10].

PROTOCOL:

1. Design RNA oligonucleotides

1.1. Design synthetic RNA oligonucleotides with different lengths (19 nt, 20 nt and 21 nt), including one (RNA #6) with both canonical and modified nucleotides. ψ is employed as a model for non-mass-altering modifications, which is challenging for MS sequencing because it has an identical mass to U. m^{5}C is chosen as a model for mass-altering modifications to demonstrate the robustness of the approach.

RNA #1: 5´-HO-CGCAUCUGACUGACCAAAA-OH-3´
RNA #2: 5´-HO-AUAGCCCAGUCAGUCUACGC-OH-3´
RNA #3: 5´-HO-AAACCGUUACCAUUACUGAG-OH-3´
RNA #4: 5´-HO-GCGUACAUCUUCCCCUUUAU-OH-3´
RNA #5: 5´-HO-GCGGAUUUAGCUCAGUUGGGA-OH-3´
RNA #6: 5´-HO-AAACCGUψACCAUUAm^{5}CUGAG-OH-3´

1.2. Dissolve each synthetic RNA in nuclease-free DEPC-treated water (expressed as DEPC-treated H_2O unless otherwise indicated) to obtain 100 μM RNA stock solution. Aliquot and store at -20 °C.

1.3. To avoid possible RNA sample degradation, use RNase-free experimental supplies including DEPC-treated water, microcentrifuge tubes, and pipette tips. Frequently wipe down surfaces of lab supplies by RNase elimination wipes.

2. Label the 3´-end of RNAs with biotin

2.1. Two-step reaction protocol (adenylation and ligation)

2.1.1. Add 1 µL of 10x adenylation reaction buffer (50 mM sodium acetate, pH 6.0, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM EDTA), 1 µL of 1 mM ATP, 1 µL of 100 µM pCp-biotin, 1 µL of 50 µM *Mth* RNA ligase and 6 µL of DEPC-treated H_2O (a total volume of 10 µL) into an RNase-free thin walled 0.2 mL PCR tube.

NOTE: Store the reagents at -20 °C before the two-step reaction. Thaw the reagents at room temperature and mix well by vortexing and centrifuging before adding to the reaction.

2.1.2. Incubate the reaction at 65 °C for 1 h and inactivate the reaction at 85 °C for 5 min.

2.1.3. Conduct the ligation step containing 10 µL of reaction solution from the previous step, 3 µL of 10x T4 RNA ligase reaction buffer (50 mM Tris-HCl, pH 7.8, 10 mM $MgCl_2$, 1 mM DTT), 1.5 µL of 100 μM RNA sample to be sequenced, 3 µL of anhydrous DMSO to reach 10% (v/v), 1 µL of T4 RNA ligase (10 units/µL), and 11.5 µL of DEPC-treated H_2O (a total volume of 30 μL). Incubate the reaction overnight at 16 °C.

NOTE: Add reaction components at room temperature due to the high freezing point of DMSO (18.45 °C).

2.1.4. Incubate the reaction overnight (16 h) at 16 °C.

2.1.5. Quench and purify the reaction by column purification to remove enzymes and free pCp-biotin. Add 20 μL of DEPC-treated H_2O to the reaction solution to reach a 50 μL sample volume prior to adding the binding buffer.

2.1.6. Add 100 μL of binding buffer to each reaction solution. Add 400 μL of ethanol, mix by pipetting, and transfer the mixture to the column. Centrifuge at 10,000 x *g* for 30 s. Discard the flow-through.

2.1.7. Add 750 µL of DNA Wash Buffer to the column. Centrifuge at 10,000 x *g* and maximum speed for 30 s and 1 minute, respectively.

2.1.8. Transfer the column to a 1.5 mL microcentrifuge tube. Add 15 µL of DEPC-treated H_2O to the column and centrifuge at 10,000 x *g* for 30 s to elute the RNA product.

NOTE: Samples can be stored at -20 °C at this stage until the next step is performed.

2.2. One-step reaction protocol

2.2.1. Perform a one-step labeling reaction containing 2 µL of 150 µM AppCp-biotin, 3 µL of 10x ligase reaction buffer, 1.5 µL of 100 µM RNA sample to be sequenced, 3 µL of anhydrous DMSO to reach 10% (v/v), 1 µL of T4 RNA ligase (10 units/µL), and 19.5 µL of DEPC-treated H_2O (a total volume of 30 µL).

2.2.2. Incubate the reaction overnight (16 h) at 16 °C.

2.2.3. Perform column purification as described above in step 2.1.5.

NOTE: Prepare a separate/exclusive reaction tube for each RNA sample (150 pmol scale of RNA). Label the 5´-end of RNAs with sulfo-Cy3 or Cy3 may be needed (e.g., for bidirectional sequencing). The method is different than that of 3´-biotinylation and is described in a previous publication[9].

3. Capture biotinylated RNA sample on streptavidin beads

3.1. Activate 200 µL of Streptavidin C1 magnet beads by adding 200 µL of 1x B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl). Vortex this solution and place it on a magnet stand for 2 min. Then discard the supernatant.

3.2. Wash the beads twice with 200 µL of Solution A (DEPC-treated 0.1 M NaOH and DEPC-treated 0.05 M NaCl) and once in 200 µL of Solution B (DEPC-treated 0.1 M NaCl). For each wash step, vortex the solution and place it on a magnet stand for 2 min, discard the supernatant. Then add 100 µL of 2x B&W buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl).

3.3. Add 1x B&W buffer to the biotinylated RNA sample until volume is 100 µL. Then add this solution to the washed beads stored in 100 µL of 2x B&W buffer. Incubate for 30 min at room temperature on a rocking platform shaker at 100 rpm. Place the tube in a magnet for 2-3 min and discard the supernatant.

3.4. Wash the coated beads 3 times in 1x B&W buffer and measure the final concentration of supernatant in each wash step by Nanodrop for recovery analysis, to confirm that the target RNA molecules remain on the beads.

3.5. Incubate the beads in 10 mM EDTA, pH 8.2 with 95% formamide at 65 °C for 5 min. Keep the tube on the magnet stand for 2 min and collect the supernatant (containing the biotinylated RNAs released from the streptavidin beads) by pipet.

NOTE: This physical separation step prior to acid degradation is only used for sequencing of RNA#1 in **Figure 1c**, and is not mandatory for the 2D-HELS MS Seq since the hydrophobic biotin label can cause the 3´ labeled ladder fragments to have a significantly delayed t_R during LC-MS measurement, which can clearly distinguish the labeled 3´ ladder fragments from the unlabeled 5´ ladder fragments in the 2D mass-t_R plot.

4. Acid hydrolysis of RNA to generate MS ladders for sequencing

4.1. Divide each RNA sample into three equal aliquots. For instance, divide a volume of 15 µL RNA sample into 3 aliquots of 5 µL.

4.2. Add an equal volume of formic acid to achieve 50% (v/v) formic acid in the reaction mixture[8,9].

4.3. Incubate the reaction at 40 °C, with one reaction running for 2 min, one for 5 min, and one for 15 min, respectively.

4.4. Quench the acid degradation by immediately freezing the sample on dry ice.

4.5. Use a centrifugal vacuum concentrator to dry the sample. The sample is typically completely dried within 30 min, and formic acid is removed together with H_2O during the drying process because it has a boiling point (100.8 °C) similar to that of H_2O (100 °C).

4.6. Suspend and combine the dried samples in 20 µL of DEPC-treated H_2O for LC-MS measurement.

NOTE: Samples can be stored at -20 °C at this stage waiting for LC-MS measurement.

5. Convert ψ to CMC-ψ adduct

5.1. Add 80 µL of DEPC-treated H_2O into 1.5 mL RNase-free microcentrifuge tube containing 0.0141 g of N-cyclohexyl-N'-(2-morpholinoethyl)-carbodiimide metho-p-toluenesulfonate (CMC) and 0.07 g of urea. Add 10 µL of 100 µM RNA to be sequenced, 8 µL of 1 M bicine buffer (pH 8.3) and 1.28 µL of 0.5 M EDTA. Add DEPC-treated H_2O to reach a total volume of 160 µL. Final concentrations are 0.17 M CMC, 7 M urea and 4 mM EDTA in 50 mM bicine (pH 8.3)[12].

NOTE: This protocol is applicable to either a single-stranded synthetic RNA or RNA mixtures.

5.2. Divide 160 µL reaction solution into 4 equal aliquots and incubate at 37 °C for 20 min.

NOTE: 50 µL per tube is the maximum reaction volume that can be used in a thermal cycler.

5.3. Quench each reaction with 10 µL of 1.5 M sodium acetate and 0.5 mM EDTA (pH 5.6).

5.4. Perform column purification with 4 parallel spin columns to remove excessive reactants according to the procedure as described in step 2.1.5. Dissolve the purified product in 15 µL of DEPC-treated H_2O in each collection tube.

5.5. Transfer the purified product to RNase-free, thin walled 0.2 mL PCR tubes × 4. Add 20 µL of 0.1 M Na_2CO_3 buffer (pH 10.4) into each 15 µL of purified product and make a final volume of 40 µL by adding DEPC-treated H_2O for each reaction tube (in total 4). Incubate the reaction at 37 °C for 2 h.

5.6. Quench and purify the reaction by column purification with 4 parallel spin columns as described in step 2.1.5. Elute the CMC-ψ converted product to a collection tube each with 15 µL of DEPC-treated H_2O.

5.7. Combine the purified CMC-ψ converted sample from 4 collection tubes into one tube. Perform formic acid degradation 50% (v/v) according to the procedures as described in step 4.1-4.6 to generate MS ladders for sequencing.

6. LC-MS measurement

6.1. Prepare mobile phases. Mobile phase A is 25 mM hexafluoro-2-propanol with 10 mM diisopropylamine in LC-MS grade water; mobile phase B is methanol.

6.2. Resuspend acid-degraded RNA in DEPC-treated H_2O and transfer the sample to LC-MS sample vial for analysis. Each sample injection volume is 20 µL containing 100-400 pmol of RNA.

6.3. Use the following LC conditions: column temperature of 35 °C, a flow rate of 0.3 mL/min; a linear gradient from 2-20% mobile phase B in 15 min followed by a 2 min wash step with 90% mobile phase B.

NOTE: For more hydrophobic end-labels such as Cy3 and sulfo-Cy3 as mentioned in Section 2, a higher percentage of organic solvent may be necessary for sample elution (i.e., a similar gradient can be used but with an increased percentage range of mobile phase B). For instance, 2-38% mobile phase B in 30 min with a 2 min wash step with 90% mobile phase B.

6.4. Separate and analyze sampled on a Q-TOF mass spectrometer coupled to a LC system equipped with an autosampler and a MS HPLC system. The LC column is a 50 mm x 2.1 mm C18 column with a particle size of 1.7 µm. Use the following MS settings: negative ion mode; range, 350 m/z to 3200 m/z; scan rate, 2 spectrum/s; drying gas flow, 17 L/min; drying gas temperature, 250 °C; nebulizer pressure, 30 psig; capillary voltage, 3500 V; and fragmentor voltage, 365 V. Please note that these parameters are specific to the type or model of mass spectrometer being used.

6.5. Acquire data with the acquisition software. Use Molecular Feature Extraction (MFE) workflow to extract compound information including mass, retention time, volume (the MFE abundance for the respective ion species), and quality score, etc. Use the following MFE settings: "centroid data format, small molecules (chromatographic), peak with height ≥ 100, up to a maximum of 1000, quality score ≥ 50".

NOTE: Optimize MFE settings to extract as many potential compounds as possible, up to a maximum of 1000, and with quality scores of ≥ 50.

7. Automate RNA sequence generation by a computational algorithm

NOTE: This is shown only for RNA #1 in **Figure 1c**.

7.1. Sort out MFE extracted compounds in order of high volume (peak intensity) and retention time. Perform data pre-selection via 1) setting retention time from 4 to 10 min to select the RNA fragments labeled by the biotin since the t_Rs of the biotin labeled mass ladder components are shifted to the t_R window (4 min to 10 min), and 2) using an order-of-magnitude higher of input compounds than the number of ladder fragments for algorithm computation to reduce data amount based on volume. For instance, for a 20 nt RNA, 20 labeled mass- t_R ladder components will be required for sequencing of the 20 nt RNA, thus, 200 compounds from MFE data file will be selected based on volume. Please note that the t_R window may be different when a different type or model of mass spectrometer is used.

7.2. Perform data process and sequence generation of RNA #1 using a revised version of a published algorithm[8]. The source codes of the revised algorithm are described previously (https://academic.oup.com/nar/article/47/20/e125/5558343#supplementary-data)[9].

7.3. In addition to automating sequence generation using the algorithm, manually calculate the mass differences between two adjacent ladder components for base calling. All bases in the RNA can be called manually and match with the theoretical ones in the RNA nucleotide and modification database[8], thus the complete sequence of the RNA strand is accurately read out manually and is used to confirm the accuracy of the algorithm-reported sequence read. More structures of RNA modifications can be found in RNA modification databases[13], and their corresponding theoretical masses are obtained by ChemBioDraw. In Tables S1-S6, the ppm mass difference is shown when comparing the observed mass to its theoretical mass for a specific ladder component, and a value less than 10 is considered a good match for each base calling.

8. Sequencing RNA mixtures

8.1. Label a mixture of 5 RNA strands (RNA #1 to #5) at their 3´-ends using a one-step protocol described in step 2.2. In a 150 µL reaction solution, add 15 µL of 10x T4 RNA ligase reaction buffer, 1.5 µL of each RNA strand (100 µM, RNA #1 to #5, respectively, in total 7.5 µL), 10 µL of 150 µM A(5´)pp(5´)Cp-TEG-biotin-3´, 15 µL of anhydrous DMSO, 5 µL of T4 RNA ligase (10 units/µL), and 97.5 µL of DEPC-treated H_2O. Equally distribute the reaction solution into 5 aliquots. Each RNase-free microcentrifuge tube contains 30 µL of reaction solution.

8.2. Incubate the reaction overnight (16 h) at 16 °C.

8.3. Perform column purification according to the procedure as described in step 2.1.5 with 5 parallel spin columns. Elute a mixture sample of 3´-biotinylated 5 RNA strands (mixture of RNA #1 to #5) to a collection tube each with 15 µL of DEPC-treated H_2O.

8.4. Combine the purified mixture sample from 5 collection tubes into one tube. Perform formic acid degradation according to the procedure described in Section 4.

8.5. Measure samples by LC-MS as described in Section 6, and analyze the data using the data analysis software with optimized MFE settings to extract data containing mass, retention time and volume as described in step 6.5. The typical processing and base-calling algorithm is not applied due to the significantly increased data complexity resulting from the mixture. All bases in the RNA of the mixed sample are called manually in a way similar to Section 7.3 and match well with the theoretical ones in the RNA nucleotide and modification database[8], thus the complete sequences of all 5 RNA strands in the mixed sample are accurately read out. In **Tables S7–S11**, all the information is listed including observed mass, t_R, volume, quality score and ppm mass difference.

REPRESENTATIVE RESULTS:

Introducing a biotin tag to 3´-end of RNA to produce easily-identifiable mass-t_R ladders. The workflow of 2D-HELS MS Seq approach is demonstrated in **Figure 1a**. The hydrophobic biotin label introduced to the 3´ end of the RNA (see Section 2) increases the masses and t_Rs of the 3´ labeled ladder components when comparing to those of their unlabeled counterparts. Thus, the 3´-ladder curve is shifted up (due to increase in the t_Rs) and shifted to the right (due to increase in masses) in the 2D mass-t_R plot. **Figure 1b** shows the protocol for introducing a biotin tag to the 3´-end of RNA. **Figure 1c** demonstrates separation of the 3´-ladder from the 5´-ladder and other undesired fragments on a 2D mass-t_R plot based on systematic changes in t_R of 3´-biotin-labeled mass-t_R ladders of RNA #1. The 3´-ladder curve alone gives a complete sequence of RNA #1, and the 5´-ladder curve that does not have a t_R shift provides the reverse sequence, but it requires end pairing for reading the terminal base[8]. With this strategy of 2D-HELS, end pairing would not be required as reported before and the entire RNA sequence can be read out completely from only one labeled ladder curve[8]. As such, it is possible to sequence mixed samples containing multiple RNAs: two RNA strands of different lengths (RNA #1 and RNA #2, 19 nt and 20 nt, respectively) with a 5´-biotin label at each RNA (**Figure 1d**).

Converting ψ and its CMC-ψ adduct for 2D-HELS MS Seq. ψ is a difficult nucleotide modification for MS-based sequencing because it has an identical mass as uridine (U). To differentiate these two, we treat the RNA with CMC, which converts a ψ to a CMC-ψ adduct (see Section 5). The adduct has a different mass than U and can be differentiated in the 2D-HELS MS Seq. **Figure 2a** shows the HPLC profile of the crude product of the reaction converting ψ to its CMC-adduct in RNA #6. By integrating their UV peaks, we calculated the percent conversion and 42% ψ is converted to CMC-ψ adduct. After acid degradation and LC-MS measurement, we manually acquired the sequence based on both non-converted ladders and CMC-converted ladders identified from the algorithm-processed data[8,9]. A red curve branches up off of the grey curve starting from ψ at the position 8 in RNA #12 (**Figure 2b**), due to partial conversion of ψ to the CMC-ψ adduct. Because of its mass and hydrophobicity of the CMC, this conversion results in a 252.2076 Dalton increase in mass and a significant increase in t_R for each CMC-ψ adduct-containing ladder component when comparing to its unconverted counterpart. Thus, a dramatic shift starting at the position of 8 can be observed in the 2D mass-t_R plot, indicating that this is a ψ at the position of 8 in the RNA sequence.

Sequencing RNA mixtures. A mixture of five different RNA strands are sequenced by 2D-HELS MS Seq approach with 3´-end labeling (see Section 8). The concern for sequencing mixed RNAs is that multiple ladder curves may overlap with each other when they all share the same starting points (the hydrophobic tag in the 2D mass-t_R plot). However, base calling is made one by one, each based on a mass difference between two adjacent ladder fragments in the MFE data, and we can make the correct base-calling as long as each mass difference matches well with one of the theoretical masses of canonical or modified nucleotides in the data pool[8,9]. In the analysis of the multiplexed RNA samples, the typical processing and base-calling algorithm used in **Figure 1** and **Figure 2** is not used mainly due to the significantly increased data complexity resulting from the mixture. These sequences are base-called manually via calculating the mass difference between two adjacent mass ladder fragments, and comparing it to the theoretical mass of the nucleotide in the data pool[9]. The matched one with a mass PPM <10 is chosen to report the identity at this position. With this one by one manual calculation for base-calling, all sequences in the mixture are accurately sequenced. OriginLab software is used to re-construct a 2D mass-t_R graph, in which the t_Rs are normalized arbitrarily for better visualizing five different RNA sequences (**Figure 3**). Without the normalization, the letter codes (i.e., A, C, G, U, or modifications like ψ) for sequences of 5 RNA would be crowded all together (**Figure S1**), and could not be visualized as easily as in **Figure 3**. Similarly, the sequencing results demonstrate that the approach is not just limited to sequence purified single-stranded RNAs, but more importantly, RNA mixtures with multiple RNA strands. Algorithms are current under development to automate the process of base-calling and sequence generation.

FIGURE AND TABLE LEGENDS:

Figure 1. 2D-HELS MS Seq of representative RNA samples. (**a**) Workflow for 2D-HELS MS Seq. The major steps include 1) hydrophobic tag labeling of RNA to be sequenced, 2) acid hydrolysis, 3) LC-MS measurement, 4) extract and analyze MFE data and 5) sequence generation via algorithms or manual calculation. (**b**) Protocol for introducing a biotin tag to the 3´-end of RNA. (**c**) Separation of the 3´-ladder from the 5´-ladder and other undesired fragments in a mass-retention time (t_R) plot based on systematic changes in t_Rs of 3´-biotin-labeled mass-t_R ladders of RNA #1 (19 nt). The sequences are de novo and automatically read out directly by a computational base-calling algorithm[9]. (**d**) Simultaneous sequencing of 5´-biotin labeled RNA #1 and RNA #2, 19 nt and 20 nt, respectively. Methods for introducing a biotin tag to the 5´ end of RNA are different than that of 3´-biotinylation, and can be found in the previous published protocol[9]. The 5´ end of two RNAs (RNA#1 and RNA#2) are biotinylated and their 5´ biotinylated ladders can be easily identified in the 2D mass-t_R plot after LC-MS. Both 5´ biotinylated ladders are easily separated from their unlabeled 3´ ladders, because the biotinylated ladder components have the larger t_R

shifts due to the hydrophobicity of the biotin, while unlabeled ladder components are in the lower t_R region. Although the 5´ladders and 3´ ladders co-exist, they do not interfere the sequence interpretation of two mixed RNA strands. Each sequences of these two RNAs are manually acquired from 5´ biotinylated ladders based on the computational algorithm-processed data[8,9]. This figure has been modified from Zhang et al.[9].

Figure 2. Converting pseudouridine (ψ) and its adduct for 2D-HELS MS Seq. (**a**) HPLC profile of the crude product of the reaction converting ψ to its CMC adduct in a 20 nt RNA (RNA #6) that contains only one ψ. (**b**) Sequencing of a ψ-containing RNA #6. The conversion of the ψ to the CMC-ψ adducts (ψ*) results in a 252.2076 Dalton increase in mass and a significant increase in t_R because of its mass and hydrophobicity of the CMC. Thus, a dramatic shift starting at the position of 8 can be observed in the mass-t_R curve, indicating that this is a ψ at the position of 8 in the RNA sequence. The sequences are manually acquired based on the computational algorithm-processed data[8,9]. This figure has been modified from Zhang et al.[9].

Figure 3. Sequencing RNA mixtures containing 5 distinct RNAs. A biotin is used to label RNAs at the 3´-end before 2D-HELS MS Seq, and t_Rs are normalized for better visualization. For each sequence, the starting t_R values are normalized to start at 7 min intervals. The absolute differences between the starting t_R value and subsequent t_R remain unchanged for each of the 5 RNAs, and thus it is easier to visualize each of them in one picture. All the base-callings are performed by manually calculating the mass differences of two adjacent ladder components and matching them with the theoretical ones in the RNA nucleotide and modification database[8]; Plots for Figure 3 are re-constructed using OriginLab based on the manual base-calling and sequencing data (see Section of Sequencing RNA mixtures in Representative Results). The 2D mass-t_R figure of the five mixed RNAs without the t_R normalization is shown in Figure S1.

DISCUSSION:

Unlike tandem-based MS fragmentation, highly controlled acidic hydrolysis is used in the sequencing approach to fragment the RNA before analysis with a mass spectrometer[14-19]. As a result, each acid-degraded fragment product can be detected by the instrument forming the equivalent of a sequencing ladder. Under optimal conditions, this method creates an “ideal” sequence ladder from RNA via, on average, once per molecule site-specific RNA cleavage exclusively at the phosphodiester bonds[8-10]. After each degraded fragment is measured by the mass spectrometer in a single run, the mass difference between two adjacent ladder fragments corresponds to the exact mass of the nucleotide at that position. Each RNA modification either has an intrinsic unique mass that can help to identify and locate it in the RNA or can be converted to one with a unique mass. Thus, in theory, this method can report the identity and location of both canonical and modified nucleotides for de novo and direct sequencing of any RNA. However, different sequence ladders may overlap with each other, complicating MS data analysis and making it difficult for RNA sequencing by MS in practice.

One of the benefits of the 3´hydrophobic tag is that it overcomes the major challenge in any fragmentation method (i.e., that every RNA molecule is cleaved into two fragments): one containing the original 5´ end, the other containing the original 3´ end of the RNA. Therefore, each cleavage event produces two fragments, producing two ladders—one measured from the 5′ side, and the other from the 3′ side. There is always ambiguity in figuring out which peak belongs to which ladder. This becomes more problematic in a mixture of several different RNAs, due to generation of a large number of overlapping sequence ladders. However, since all ladder fragments from the 3′ ends are labeled with a hydrophobic tag, they exhibit much later t_Rs (**Figure 1a**). As a result, we can obtain clear and unambiguous ladders in the 2D mass-t_R data exclusively derived from the 3′ end of RNA. Notably, we are optimizing approaches to selectively tag either on the 5′ or 3′ end of any RNA using different chemical conjugation methods[9], which can provide the sequence information twice when reading from both 5′ and 3′ directions, and thus further improving the accuracy of sequencing.

For de novo sequencing of unknown RNA samples, especially for complex biological samples, a general and robust algorithm is required to process a massive amount of LC-MS data for sequence generation in an accurate and efficient manner, which has recently become available via other published work[10]. Although these algorithms have been used for sequencing more complicated samples[9,10], in this manuscript, we performed manual base calling for sequence generation unless indicated otherwise. We aim at covering all the key steps in the 2D-HELS MS Seq, and would like to illustrate the process during which even without using additional sequencing algorithms, we can still manually read out sequences of the RNA to be sequenced. For better visualization and for easier to find ladder fragments needed for sequencing in the 2D mass-t_R plot, the MFE files of each LC-MS run are processed by a revised version of a published algorithm[8] before reading their sequences, unless indicated otherwise. The published algorithm cannot be used directly to read out the sequences from the LC-MS data, but we can still use part of its function to process the data: hierarchically clustering mass adducts to augment compound intensity of ladder components and to reduce the data complexity, especially in the crucial regions where sequence reads are generated[8,9].

One of the crucial steps during sample preparation for 2D-HELS MS Seq is to improve the efficiency of labeling the end of RNA with a hydrophobic tag. A high labeling efficiency can help to reduce the amount of RNA sample needed for generating MS signals that sequence data rely on. In order to increase the labeling efficiency, we employ new labeling strategies, including using activated AppCp-biotin to avoid the adenylation step when labeling the 3′-end of the RNA. The yield of one-step reaction for labelling the 3′-end of a 19 nt RNA with biotin (see step 2.2) can be improved from 60% to ~95%[9]. With the efficient labeling, we are able to sequence a mixed sample containing up to 12 distinct RNAs as previously described[9]. In this manuscript we use a mixture of 5 RNAs as a representative example to illustrate the sequencing process; we also detect all ladder fragments needed for accurate sequencing and read out the complete sequences of each of the 5 RNA strands in the mixture. Better labeling efficiency not only helps to minimize the sample loading amount, but also helps to significantly reduce the data complexity during the downstream data analysis for sequence generation. Novel reactions are currently under development to achieve quantitative yield in labeling RNAs on both 5′ and 3′ends.

When sequencing RNA #1 as shown in **Figure 1c**, streptavidin capture and release are used to physically separate biotinylated RNA #1 prior to acid degradation (see Section 3). This helps to remove a small portion of unlabeled RNA, and subsequently make it easier to visually identify the labeled mass ladders in the 2D mass-t_R plot. However, the physical separation is not a mandatory step because the biotinylated RNA ladder fragments have delayed/larger t_Rs due to the hydrophobicity from the biotin tag, when compared to their unlabeled counterparts. In addition, base calling does not rely on physical separation, but relies on the mass differences of adjacent mass ladder components, thus, the right base calling can be achieved as long as the mass differences of two adjacent ladder components match well with the corresponding masses of a particular nucleotide or modification in the RNA nucleotide and modification datebase[8]. The computational algorithm is currently under development for automating base-calling and sequence generation.

The MFE settings that help to export the original LC-MS data (in the file type of .d) into spreadsheet files are highly crucial to the data processing and subsequent sequence generation (see Section 6.5). For instance, we tested the MFE setting “peak with height” in a range from 100 to 1000 and noticed that setting of 100 can provide us with 2-fold more compounds than those of setting 1000. In order to avoid missing any ladder components, we can adjust the MFE setting during the sequencing workflow. This setting is likely dependent on mass resolution, the amount of mass ladder fragments and data complexity. In addition, it is important to use the centroid dataset and chromatographic type setting for small molecules. The quality score can be varied from 50 % to 100% based on the data quality.

The LC-MS instruments we use in the study has an upper mass resolution of ~40K, limiting the method to only sequencing RNA of less than 35 bases long. However, the exact read length of this method is instrument-dependent; more advanced instruments with higher resolving power may lead to longer read length. Similarly, the throughput, that is how many RNA can be sequenced in a single LC-MS run, remains to be explored, although we manually sequenced a mixture of RNA sample up to 12 distinct RNA strands even without the use of algorithm[9]. With the current workflow, ~100 pmol short RNA (<35 nt) is required for each LC-MS run. The loading amount increases when additional experiments are needed: for differentiating isomeric nucleotide modifications, and typically up to 400 pmol is required. For sequencing specific tRNA like tRNA^{Phe}, ~1000 pmol sample may be needed for its sequencing and modification analysis. However, we expect required sample loading amounts will be decreased on LC-MS instruments with greater sensitivity. With improvements in sample labeling efficiency, sequencing algorithm, instrument sensitivity and resolution, we expect our method to be applicable to a wider range of RNA samples, especially those with various RNA modifications.

ACKNOWLEDGMENTS:

The authors acknowledge the R21 grant from NIH (1R21HG009576) to S. Z. and W. L. and New York Institute of Technology (NYIT) Institutional Support for Research and Creativity grants to S. Z., which supported this work. The authors would like to thank PhD student Xuanting Wang (Columbia University) for assisting in figure-making, and thank Prof. Michael Hadjiargyrou (NYIT), Prof. Jingyue Ju (Columbia University), Drs. James Russo, Shiv Kumar, Xiaoxu Li, Steffen Jockusch, and other members of the Ju lab (Columbia University), Dr. Yongdong Wang (Cerno Bioscience), Meina Aziz (NYIT), and Wenhao Ni (NYIT) for helpful discussions and suggestions for our manuscript.

DISCLOSURES:

The authors have filed a provisional patent related to the technology discussed in this manuscript.

REFERENCES:

1 Addepalli, B., Venus, S., Thakur, P., Limbach, P. A. Novel ribonuclease activity of cusativin from Cucumis sativus for mapping nucleoside modifications in RNA. *Analytical and Bioanalytical Chemistry.* **409** (24), 5645-5654 (2017).

2 Gao, H., Liu, Y., Rumley, M., Yuan, H., Mao, B. Sequence confirmation of chemically modified RNAs using exonuclease digestion and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry.* **23** (21), 3423-3430 (2009).

3 McLuckey, S. A., Van Berkel, G. J., Glish, G. L. Tandem mass spectrometry of small, multiply charged oligonucleotides. *Journal of The American Society for Mass Spectrometry.* **3** (1), 60-70 (1992).

4 Fountain, K. J., Gilar, M., Gebler, J. C. Analysis of native and chemically modified oligonucleotides by tandem ion-pair reversed-phase high-performance liquid chromatography/electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry.* **17** (7), 646-653 (2003).

5 Taucher, M., Breuker, K. Characterization of modified RNA by top-down mass spectrometry. *Angewandte Chemie International Edition in English.* **51** (45), 11289-11292 (2012).

6 Kellner, S., Burhenne, J., Helm, M. Detection of RNA modifications. *RNA Biology.* **7** (2), 237-247 (2010).

7 Thomas, B., Akoulitchev, A. V. Mass spectrometry of RNA. *Trends in Biochemical Sciences.* **31** (3), 173-181 (2006).

8 Bjorkbom, A. et al. Bidirectional direct sequencing of noncanonical RNA by two-dimensional analysis of mass chromatograms. *Journal of the American Chemical Society.* **137** (45), 14430-14438 (2015).

9 Zhang, N. et al. A general LC-MS-based RNA sequencing method for direct analysis of multiple-base modifications in RNA mixtures. *Nucleic Acids Research.* **47** (20), e125 (2019).

10 Zhang, N. et al. 2D-HELS-AA MS Seq: Direct sequencing of tRNA reveals its different isoforms and multiple dynamic base modifications. *BioRxiv* (2019).

11 Bahr, U., Aygun, H., Karas, M. Sequencing of Single and Double Stranded RNA Oligonucleotides by Acid Hydrolysis and MALDI Mass Spectrometry. *Analytical Chemistry.* **81** (8), 3173-3179 (2009).

12 Bakin, A., Ofengand, J. Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique. *Biochemistry.* **32** (37), 9754-9762 (1993).

13 Cantara, W. A. et al. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Research.* **39** (Database issue), D195-201 (2011).

14 Thuring, K., Schmid, K., Keller, P., Helm, M. LC-MS Analysis of Methylated RNA. *Methods in Molecular Biology.* **1562**, 3-18 (2017).

15 Bahr, U., Aygun, H., Karas, M. Sequencing of single and double stranded RNA oligonucleotides by acid hydrolysis and MALDI mass spectrometry. *Analytical Chemistry.* **81** (8), 3173-3179 (2009).

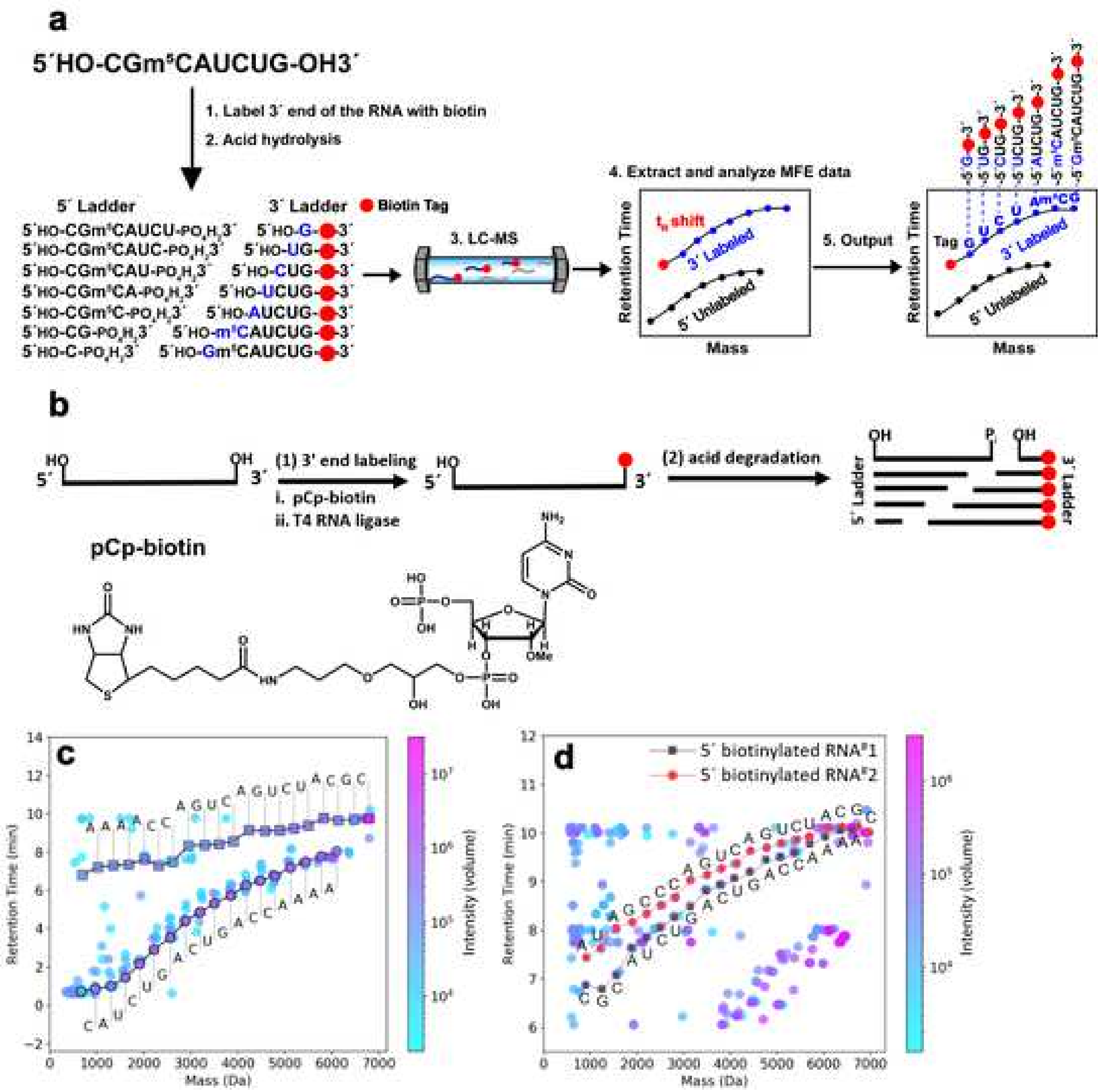
16 Hahner, S. et al. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) of endonuclease digests of RNA. *Nucleic Acids Research.* **25** (10), 1957-1964 (1997).

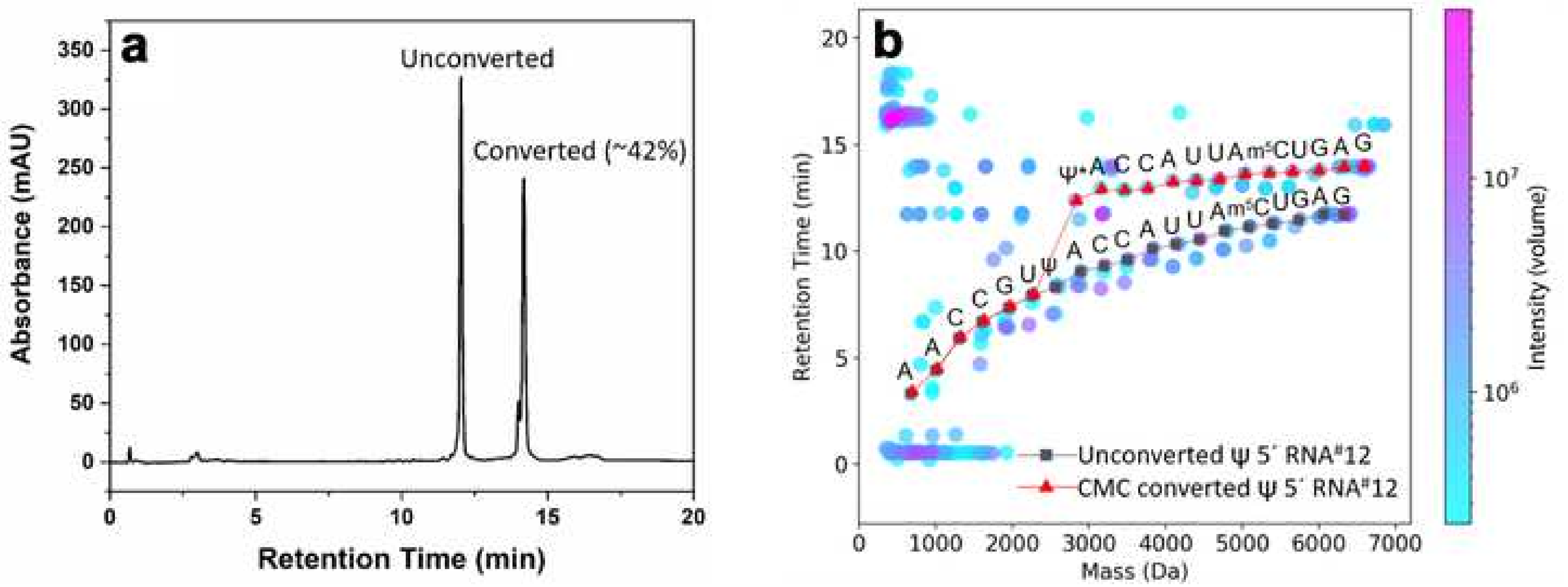
17 Tolson, D. A., Nicholson, N. H. Sequencing RNA by a combination of exonuclease digestion and

uridine specific chemical cleavage using MALDI-TOF. *Nucleic Acids Research.* **26** (2), 446-451 (1998).

18 Smirnov, I. P. et al. Sequencing oligonucleotides by exonuclease digestion and delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Analytical Biochemistry.* **238** (1), 19-25 (1996).

19 Gupta, R. C., Randerath, K. Use of specific endonuclease cleavage in RNA sequencing. *Nucleic Acids Research.* **4** (6), 1957-1978 (1977).


a
5´HO-CGm⁵CAUCUG-OH3´
1. Label 3´ end of the RNA with biotin
2. Acid hydrolysis
5´ Ladder
3´ Ladder
● Biotin Tag
5´HO-CGm⁵CAUCU-PO4H2 3´
5´HO-CGm⁵CAUC-PO4H2 3´
5´HO-CGm⁵CAU-PO4H2 3´
5´HO-CGm⁵CA-PO4H2 3´
5´HO-CGm⁵C-PO4H2 3´
5´HO-CG-PO4H2 3´
5´HO-C-PO4H2 3´
5´HO-G-●-3´
5´HO-UG-●-3´
5´HO-CUG-●-3´
5´HO-UCUG-●-3´
5´HO-AUCUG-●-3´
5´HO-m⁵CAUCUG-●-3´
5´HO-Gm⁵CAUCUG-●-3´
3. LC-MS
4. Extract and analyze MFE data
tR shift
3´ Labeled
5´ Unlabeled
Retention Time
Mass
5. Output
Tag
G U C U A m⁵C G
-5´G-●-3´
-5´UG-●-3´
-5´CUG-●-3´
-5´UCUG-●-3´
-5´AUCUG-●-3´
-5´m⁵CAUCUG-●-3´
-5´Gm⁵CAUCUG-●-3´
b
HO 5´
OH 3´
(1) 3' end labeling
i. pCp-biotin
ii. T4 RNA ligase
HO 5´
3´
(2) acid degradation
OH
P
OH
5´ Ladder
3´ Ladder
pCp-biotin
NH2
N
O
HO
O=P—O
OH
H
H
H
OMe
O
HN
NH
S
HN
O
O
O—P=O
OH
OH
c
Retention Time (min)
Mass (Da)
Intensity (volume)
14
12
10
8
6
4
2
0
−2
0
1000
2000
3000
4000
5000
6000
7000
10⁷
10⁶
10⁵
10⁴
A A A A C C A G U C A G U C U A C G C
C A U C U G A C U G A C C A A A A
d
5´ biotinylated RNA#1
5´ biotinylated RNA#2
Retention Time (min)
Mass (Da)
Intensity (volume)
12
11
10
9
8
7
6
0
1000
2000
3000
4000
5000
6000
7000
10⁶
10⁵
10⁴
A G C C C A G U C A G U C U A C G C
C G C A U C U G A C U G A C C A A A A


a
Unconverted
Converted (~42%)
Absorbance (mAU)
350
300
250
200
150
100
50
0
0
5
10
15
20
Retention Time (min)
b
Retention Time (min)
20
15
10
5
0
0
1000
2000
3000
4000
5000
6000
7000
Mass (Da)
A A C C G U ψ
ψ* A C C A U U A m5C U G A G
A C C A U U A m5C U G A G
Unconverted ψ 5´ RNA#12
CMC converted ψ 5´ RNA#12
Intensity (volume)
10^7
10^6

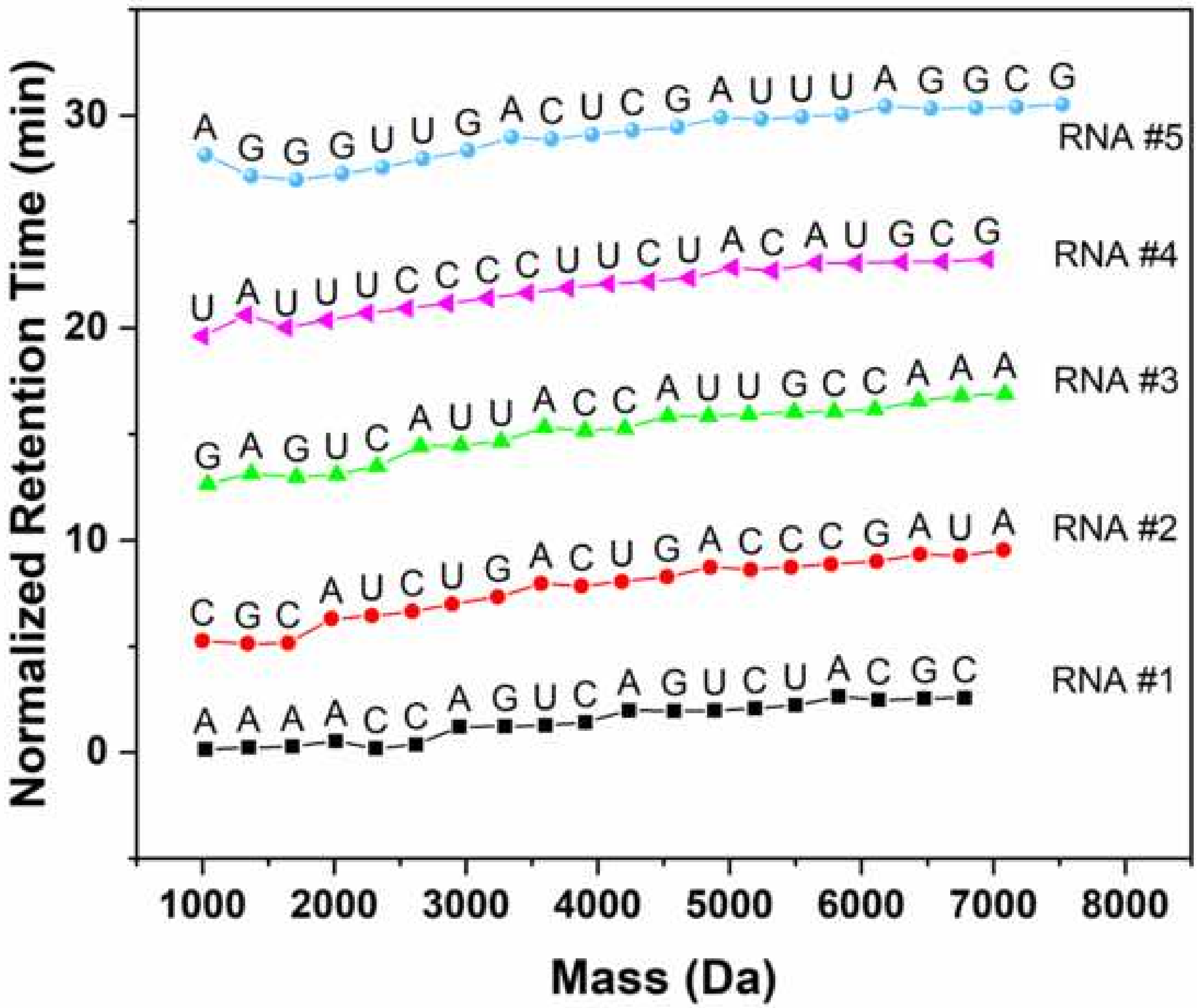

Normalized Retention Time (min)
0
10
20
30
1000
2000
3000
4000
5000
6000
7000
8000
Mass (Da)
A G G G U U G A C U C G A U U U A G G C G
RNA #5
U A U U U C C C C U U C U A C A U G C G
RNA #4
G A G U C A U U A C C A U U G C C A A A
RNA #3
C G C A U C U G A C U G A C C C G A U A
RNA #2
A A A A C C A G U C A G U C U A C G C
RNA #1

Table of Materials

Name of Material/ Equipment	**Company**	**Catalog Number**	**Comments/Description**
5' DNA Adenylation kit	New England Biolabs	E2610S	50uM concentration
6550 Q-TOF mass spectrometer	Agilent Technologies	5991-2116EN	Coupled to a 1290 Infinity LC system
A(5´)pp(5´)Cp-TEG-biotin-3´	ChemGenes	91718	HPLC purified
ATPγS	Sigma-Aldrich	11162306001	Lithium salt
Bicine	Sigma-Aldrich	B8660	BioXtra, ≥99% (titration)
Biotin maleimide	Vector Laboratories	SP-1501	Long arm
C18 column	Waters	186003532	50 mm × 2.1 mm Xbridge C18 column with a particle size o
Centrifugal Vacuum Concentrator	Labconco	Refrig 115v/60hz 7310022	Labconco CentriVap
ChemBioDraw	PerkinElmer	ChemDraw Prime	Generate a chemical structure and property data of structu
CMC (N-cyclohexyl-N’-(2-morpholinoethyl)-c	Sigma-Aldrich	2491-17-0	95% Purifiy
Cyanine3 maleimide (Cy3)	Lumiprobe	11080	Water insoluble
DEPC-treated water	Thermo Fisher Scientific	AM9906	Autoclaved, certified nuclease-free
Diisopropylamine (DIPA)	Thermo Fisher Scientific	108-18-9	99% Alfa Aesar
DMSO	Sigma-Aldrich	276855	Anhydrous dimethyl sulfoxide, 99.9%
EDTA	Sigma-Aldrich	E6758	Anhydrous, crystalline, BioReagent, suitable for cell culture
Formic acid	Merck	64-18-6	98-100%, ACS reag, Ph Eur
Hexafluoro-2-propanol (HFIP)	Thermo Fisher Scientific	920-66-1	99% Acros Organics
LC-MS sample vials	Thermo Fisher Scientific	C4000-11	Plastic screw thread vials
LC-MS vial caps	Thermo Fisher Scientific	C5000-54A	Autosampler vial screw thread caps
Na_2CO_3 buffer	Sigma-Aldrich	88975	BioUltra, >0.1 M Na_2CO_3, >0.2 M $NaHCO_3$
Oligo Clean & Concentrator	Zymo Research	D4060	Spin column
OriginLab	OriginLab	OriginPro	Data analysis and graphing software
pCp-biotin	TriLink BioTechnologies	NU-1706-BIO	20 ul (1 mM)
RNA #1--#6	Integrated DNA Technologies	Custom RNA oligos	19nt-21nt single-stranded RNAs, used without further puri
Rocking platform shaker	VWR	Orbital Shaker Standard 1000	Speed Range 40 to 300 rpm
Streptavidin magnetic beads	Thermo Fisher Scientific	88816	Binding approx. 55ug biotinylated rabbit IgG per mg of bea
Sulfonated Cyanine3 maleimide	Lumiprobe	11380	Water soluble
T4 DNA ligase 1	New England Biolabs	M0202S	400 units/uL
T4 polynucleotide kinase	Sigma-Aldrich	T4PNK-RO	From phage T4 am N81 pse T1 infected Escherichia coli BB
Tris-HCl buffer	Sigma-Aldrich	T6455	Tris-HCl Buffer, pH 10, 10×, Antigen Retriever
Urea	Sigma-Aldrich	81871	Urea for synthesis. CAS No. 57-13-6, EC Number 200-315-5

f 1.7 µm

ures & fragments

e

fication

ds

.

Point-by-Point Response to Editorial and Reviewers' Comments:

Editorial comments:
General:
1. *Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues*.
Response: Thank you very much for the reminder. We have thoroughly read through the manuscript to correct all the spelling and grammar mistakes.

2. *Please include email addresses for all authors in the manuscript itself.*
Response: As per your suggestion, the email addresses for all authors have been added.

3. *Please include at least 6 key words or phrases.*
Response: As per your suggestion, we have included 6 keywords.

4. *Please reduce the length of the Summary to 10-50 words*.
Response: As per your suggestion, the length of the Summary has been reduced to 50 words.

Protocol:
1. *There is a 10-page limit for the Protocol, but there is a 2.75 page limit for filmable content. Please highlight 2.75 pages or less of the Protocol (including headers and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader*.
Response: As per your suggestion, we have highlighted the essential steps in yellow for the video.

2. *The protocol seems mostly fine as-is, but as you make edits, please ensure you answer "how" questions, i.e., how is each step performed? Alternatively, add references to published material specifying how to perform the protocol action. If revisions cause a step to have more than 2-3 actions and 4 sentences per step, please split into separate steps or substeps*.
Response: As per your suggestion, we have listed experimental details in how is each step performed. In case a step includes more than 2-3 actions, we split them into separate steps, such as steps from 2.1.1 to 2.1.4.

Figures:
1. *Please remove the embedded figures from the manuscript*.
Response: We removed the embedded figures from the manuscript, and will submit each figure in JPG file during on-line submission.

References:
1. *Please do not abbreviate journal titles*.
Response: As per your suggestion, we now have full name of journal titles in References.

Table of Materials:

1. *Please ensure the Table of Materials has information on all materials and equipment used, especially those mentioned in the Protocol*.
Response: As per your suggestion, we have provided all the information for all the materials and instrument in the Table of Materials.

Reviewers' comments:
Reviewer #1:
Manuscript Summary:
The authors present a protocol based on an interesting recent paper that describes an improved method to analyze short RNA oligonucleotides by mass spectrometry without tandem MS fragmentation. The method entails partial digestion of purified RNA oligonucleotides and analysis by high resolution LC-MS, where the chromatographic retention time and high resolution mass of hydrolytic fragments are both used to reconstruct the RNA sequence. This version of the method augments that in an earlier publication by adding an enzymatic end-labeling scheme and by incorporating chemical modifications to accommodate isobaric RNA modifications.

Response: We'd like to thank the reviewer's positive comments and feedback.

Major Concerns:
In the abstract and summary, the authors are explicit that the manuscript "demonstrates" and "reports" findings. It appears that the data in Figures 1 and 2 of this manuscript are derived from their recent publication (ref 9), so the authors should be very careful with their language if in fact this manuscript reports only the protocols associated with that prior work, rather than novel results. If this is not the first publication of data, it should not be written as if new claims are being made herein.
Response: We'd like to thank the Reviewer for the comments. As per the suggestions, we have made changes on the expressions in the manuscript accordingly. More specifically, we have made the following revisions:

1) In Summary, "Here we describe a detailed protocol for an LC-MS-based sequencing method, which can be used: 1) as a direct method to sequence short RNA (<35 nt per run) without a cDNA intermediate, and 2) as a general method to sequence different nucleotide modifications in a single study at single-base precision."
2) In Abstract, "Recently we reported a direct RNA sequencing method by integrating a 2-dimensional mass-retention time hydrophobic end-labeling strategy into MS-based sequencing (2D-HELS MS Seq)."

If the data in Figure 3 is to be reviewed as a new result, then some clarifications about this experiment are needed. It is not at all apparent from the theory of the method that data of the kind in Figure 3 - i.e. mixtures of oligonucleotide digests of different length and composition - can in fact be analyzed simultaneously from MS spectra in a single run, as claimed in multiple places, when the sequence and composition of the oligonucleotides are unknown. Indeed, the authors refer to a manual analysis for complex mixtures based on known theoretical masses. Since the authors claim that this protocol can achieve "de novo sequencing," the authors must

clarify whether and to what extent this kind of mixed pool can be analyzed successfully with an unknown mixture.
Additionally, despite being a far more challenging analysis of mixed oligonucleotide digests, the data presented in Figure 3 looks much cleaner than that shown for simpler digests of single oligos (e.g. in Figures 1 and 2), suggesting that it has been substantially cleaned or processed. In order for potential users of the protocol to understand its applicability to real analytical problems, an appropriate presentation of this capability would be to show the figure with a full dataset from which fragments of mixed oligonucleotides were analyzed.

Response: We'd like to be clear that we only use the known theoretical mass for base calling of each nucleobase like A, C, G, U, or nucleotide modifications, and have no prior knowledge of their order/sequence in RNAs when manually reading the sequences of RNA oligos. To reflect the point made by the Reviewer on *de novo* sequencing, we have added two paragraphs in DISCUSSION (in Line 469-495) to explain the theoretical foundation of how to achieve *de novo* sequencing using MS data to generate sequences without any prior sequence knowledge.

For sequencing mixed RNA oligos, we have added a section of "**Sequencing RNA mixtures**" in **REPRESENTATIVE RESULTS** (Line 406-426) to explain how we sequence mixed RNA samples and generate sequences as shown in Figure 3. As stated in Line 419-421: For Figure 3, after manually reading out RNA sequences, "OriginLab software is used to re-construct 2D mass-t_R graph, in which the t_{Rs} are normalized arbitrarily for better visualizing five different RNA sequences (Figure 3)." Also, we added Figure S1 in SI to show their t_R without normalization, in which the letter codes (*i.e.*, A, C, G, U, or modifications like ψ) sequences of 5 RNA strands would crowd together, and cannot be visualized as easy as in Figure 3.

The particular goal of a protocol paper is for others in the community to be able to apply the methodology, and thus it should be clear to readers whether and under what conditions the protocol is appropriate for their own problems. The intro should include a very specific discussion of the scenarios in which it is appropriate to use the protocol, when success or failure should be expected, and what the starting sample requirements might be.

Response: As per the suggestion**,** we have revised the Introduction (Page 3, Line 101-113) to include a specific discussion about the technical aspect of this protocol:

"Since RNAs are all made of phosphodiester bonds, all different kinds of RNAs can be acid hydrolyzed to generate an ideal sequence ladder for 2D-HELS MS Seq under optimal conditions. However, detection of all the ladder fragments in an RNA is instrument dependent. On a standard high-resolution LC-MS (40K), the minimal loading amount for sequencing purified short RNA sample (<35 nt) is 100 pmol per run. However, more material is required (up to 400 pmol per RNA sample) when additional experiments have to be conducted, *e.g.*, to distinguish isomeric base modifications that share identical masses. This protocol used in sequencing the model synthetic modified RNAs will also be applicable to sequencing broader RNA samples, including biological RNA samples with unknown base modifications. However, a larger sample amount, such as 1000 pmol for sequencing tRNA (~76 nt) using a standard LC-MS instrument, is required for sequencing the complete tRNA with all the modifications, and an advanced algorithm needs to be developed for its *de novo* sequencing[10].”

Minor Concerns:
1) *The manuscript needs proofreading for grammar*.
Response: Thanks for the reminder. We have performed proofreading and corrected all the grammar mistakes.

2) *In the summary, the authors claim they can "quantify" RNA mixtures. No discussion of quantification appears in the protocol, and it is unclear from a theoretical perspective how quantification of chemically distinct species could be achieved by this method alone. The claim should therefore be dropped or somehow justified.*
Response: As per the suggestion, we have deleted the language related to quantification in the summary, as it was not included in the current version of the protocol. However, we carried out the quantification of stoichiometry/percentage of modified RNA in our previous publication (More details can be found in Zhang N. et al. *Nucleic Acids Res*, **2019**, 47: e125).

3) *The protocol should be clear about where the specified enzyme reaction buffers are of a commercial formulation*.
Response: As per the suggestion, we have added the compositions of enzyme reaction buffers in the protocol (Page 4, Line 144-145): “10× adenylation reaction buffer (50 mM sodium acetate, pH 6.0, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM EDTA).”

4) *Step 2.1.5: specify the source of the column used here*.
Response: The source of the column has been added in the protocol (Page 4, Line 165-166): “Provided by Oligo Clean & Concentrator.”

5) *Step 3.1: the buffer composition is not specified*.
Response: The composition of 1× B&W buffer has been added in the protocol (Page 5, Line 196-197): “1× B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl).”

6) *Step 6.4: more information about the instrument are likely needed here*.
Response: The information about the LC-MS instrument for our sample analysis has been added to Section 6.4 (Page 6, Line 295-297):
“Samples are separated and analyzed on a 6550 Q-TOF mass spectrometer coupled to a 1290 Infinity LC system equipped with a Micro AS autosampler and Surveyor MS Pump Plus HPLC system. LC column is a 50 mm × 2.1 mm Xbridge C18 column with a particle size of 1.7 μm.”

7) Step 7.1: *the language is very unclear in this step, and this may need more explanation*.
Response: We have made revisions to explain more in Step 7.1 (Page 8, Line 315-322):
“Sort out MFE extracted compounds in order of high volume (peak intensity) and retention time. Perform data pre-selection via 1) setting retention time from 4 to 10 min to select the RNA fragments labeled by the biotin since the tRs of the biotin labeled mass ladder components are shifted to the tR window (4 min to 10 min), and 2) using an order-of-magnitude higher of input compounds than the number of ladder fragments for algorithm computation to reduce data

amount based on volume. For instance, for a 20 nt RNA, 20 labeled mass- t_R ladder components will be required for sequencing of the 20 nt RNA, thus, 200 compounds from MFE data file will be selected based on volume."

8) *Step 7.2: if special software is needed to carry out the protocol, it may be necessary to provide links to the source code deposited in an online repository accessible to readers*.
Response: The link for the algorithm's source code has been provided in Section 7.2: "(https://academic.oup.com/nar/article/47/20/e125/5558343#supplementary-data)"

Reviewer #2:
Manuscript Summary:
In this manuscript, Zhang et al. described a novel sequencing method that can not only carry out RNA sequencing without making cDNA intermediates but also identify RNA modification for mixed RNAs. The method integrates a 2-dimensional mass-retention time hydrophobic end-labeling strategy into MS-based sequencing (2D-HELS MS Seq), generating a promising strategy to reveal RNA modification for biological samples in the near future.
Major Concerns:
No major issues were found.

Response: We'd like to thank the reviewer's positive comments and feedback.

Minor Concerns:
Lines 44-46, please recognize the sentence.
Response: The sentence has been revised (Page 2, Line 50-54) to: "The procedures and protocols described here in sequencing these model RNAs will be applicable to other short RNA samples (<35 nt) when using a standard high-resolution LC-MS system. In the future with the development of more robust algorithms and with better instruments, we anticipate that this method will allow the sequencing of more complex biological samples."

Line 51: please change "are currently lacking" to "are not available"?
Response: As per the suggestion, we made the change in the manuscript (Page 2, Line 59-60).

Line 58, please move reference 8 to the end of the sentence. Eliminate "sequencing" in " mainly limited to sequencing purified synthetic short RNA".
Response: The ref 8 has been moved to the end of the sentence. The "sequencing" in " mainly limited to sequencing purified synthetic short RNA" has been eliminated (Page 2, Line 66-67).

Line 59, please change "it cannot read a complete sequence from one single ladder solely" to "it cannot read a complete sequence solely based on one single ladder". Change " but have to" to "but has to". Eliminate "fore reading a complete sequence" in Line 60.
Response: All the above-mentioned changes have been made in the manuscript (Page 2, Line 67-68).

Line 71, change "to be sequences" to "to be sequenced".

Response: We'd like to thank the reviewer for pointing out the error, and we have corrected grammar mistake (Page 2, Line 79).

Line 112. Please eliminate the comma between nuclease-free and deionized.
Response: The comma had been removed and we modified this sentence to "nuclease-free DEPC-treated water" (Page 4, Line 132).

Reorganize the sentence in lines 115-116.
Response: The sentence has been revised (Page 4, Line 136-138) to: "To avoid possible RNA sample degradation, use RNase-free experimental supplies including DEPC-treated water, microcentrifuge tubes, and pipette tips. Frequently wipe down surfaces of lab supplies by RNase elimination wipes."

Reorganize the sentence of 2.1.1.
Response: The sentence has been revised (Page 4, Line 144-147) to: "Add 1 µL of 10× adenylation reaction buffer (50 mM sodium acetate, pH 6.0, 10 mM MgCl2, 5 mM DTT, 0.1 mM EDTA), 1 µL of 1 mM ATP, 1 µL of 100 µM pCp-biotin, 1 µL of 50 µM Mth RNA ligase and 6 µL of DEPC-treated H2O (a total volume of 10 µL) into an RNase-free thin walled 0.2 mL PCR tube."

Line 190, eliminate "needed".
Response: The word "needed" has been eliminated.

Line 318, move 8 towards the end of the sentence.
Response: The ref 8 has been moved to the end of the sentence.

Line 413, we used
Response: The sentence has been revised (Page 13, Line 528-529) to: "streptavidin capture and release are used to physically separate biotinylated RNA #1 prior to acid degradation (see Section 3)."

Reviewer #3:
Zhang et al. reports the development of a general LC-MS-based method for direct and de novo sequencing of RNA mixtures containing different nucleotide modifications. The method integrates a 2-dimensional mass-retention time hydrophobic end-labeling strategy into MS-based sequencing (2DHELS MS Seq). The authors successfully apply 2D-HELS MS Seq to accurately de novo sequence synthetic single-stranded RNA and RNA mixtures of up to 12 different sequences. Furthermore, authors show that their method can also identify nucleotide modifications using RNAs containing pseudouridine (Ψ) and 5-methylcytosine (m5C) nucleotide modifications as a proof-of-principle. Authors anticipate that in the near future 2D-HELS MS Seq will be applied to de novo sequence and identify known and even unknown nucleotide modifications in more complex RNA samples including the biological RNA samples.

Traditionally, mass spectrometry (MS)-based approaches have been successfully employed to identify known and unknown nucleotide modifications and map their location in the endogenous

RNA of interest. However there has been limited success to accurately de novo sequence RNA and map nucleotide modification using either one-dimensional (1D) MS data or even 2D LC-MS-based RNA sequencing method. The 2D-HELS MS Seq method presented here is relevant and could be beneficial to the researchers in the field to study and map both known and unknown modifications in the RNA. However, there are some issues in its current format. I believe addressing the issues listed below could strengthen the impact of this paper.

Response: We'd like to thank the reviewer for the positive comment and feedback.

Comments:
(1) In the manuscript authors have not addressed how much input RNA is required for 2D-HELS MS Seq? Have the authors tested various amounts of input RNA to accurately sequence RNA and map nucleotide modifications? I think it would be beneficial to mention or comment on this point. Users would also find it beneficial to know the lowest amount of RNA input that can be used for doing 2D-HELS MS Seq.
Response: As per the suggestion on sample loading amount, we have added language in Introduction. More specially, we have made additions as follows (Page 3, Line 101-113): "On a standard high-resolution LC-MS (40K), the minimal loading amount for sequencing purified short RNA sample (<35 nt) is 100 pmol per run. However, more material is required (up to 400 pmol per RNA sample) when additional experiments have to be conducted, e.g., to distinguish isomeric base modifications that share identical masses. This protocol used in sequencing the model synthetic modified RNAs will also be applicable to sequencing broader RNA samples, including biological RNA samples with unknown base modifications. However, a larger sample amount, such as 1000 pmol for sequencing tRNA (~76 nt) using a standard LC-MS instrument, is required for sequencing the complete tRNA with all the modifications, and an advanced algorithm needs to be developed for its *de novo* sequencing[10]."

(2) Related to the previous point, the authors mention that they have successfully sequenced a mixed population of tRNA samples (Ref 10). From Ref 10, I found that "400 μg purified RNase T1 partial digestion and 3´ biotinylation tRNA sample where sequenced by previous method after acid degradation and followed by LC-MS run". This is indeed a very high RNA input requirement. Can the authors comment on this point, especially that R. Ross et al./Methods 107(2016) 73-78 reported 1-5 μg of total tRNA is needed for MS. I believe optimizing the current protocol for using less amount of the input RNA could be helpful to users as the current requirement using 2D-HELS MS Seq for cellular RNA sequencing and mapping is almost prohibitive.
Response: Thanks for the Reviewer for pointing out the error. We used a wrong unit μg there in the *BioRxiv* paper (Ref. 10), which should be pmol. This mistake was corrected on our latest version of the manuscript; the paper with correct information will replace the on-line BioRxiv paper and will get published soon.

(3) Workflow diagram (Figure 1) is not clear and it is hard to follow. I suggest authors should modify the workflow diagram and add the time required for each step of the protocol. For example, try to include the headers from your main text as a flow diagram.
Response: Figure 1a has been updated according to the Reviewer's helpful suggestions. Specifically, we have added the major steps include 1) hydrophobic tag labeling of RNA to be

sequenced, 2) acid hydrolysis, 3) LC-MS measurement, 4) Extract and analyze MFE data and 5) sequence generation via algorithms or manual calculation.

(4) It will be helpful to include the catalog numbers of the reagents/material used in this protocol which will improve transparency and reproducibility. For example, Line 140 (what column?)
Response: The column information for RNA product purification has been added (Line 165-166) “Provided by Oligo Clean & Concentrator, Zymo Research.”
We also have added information for LC-MS instrument and column (Line 295-297) “Samples are separated and analyzed on a 6550 Q-TOF mass spectrometer coupled to a 1290 Infinity LC system equipped with a Micro AS autosampler and Surveyor MS Pump Plus HPLC system. LC column is a 50 mm × 2.1 mm Xbridge C18 column with a particle size of 1.7 µm.”
The sources about ALL chemicals and equipment are provided in a separate excel file “Table of Materials”.

(*5) One suggestion: Lines 399-411. This information about the labeling efficiency should actually be highlighted and included as a separate section in the protocol aimed at optimizing the labeling step, as this is super crucial. If you have 100% labeling efficiency, you don't need high input of RNA, so improving/optimizing the labeling step can be a game changer as the method can also be applied to endogenous cellular RNAs.*
Response: We completely agree that the labeling efficiency is very crucial for our sequencing method. For labeling 3' end more efficiently, we employed one-step protocol (described in 2.2) to replace two-step protocol (described in Section 2.1), and have one paragraph in Discussion (Page 13, Line 513-526) to discuss the labeling efficiency. To reflect the point made by the Reviewer, we have also added the language in Line 525-526: “Novel reactions are currently under development to achieve quantitative yield in labeling RNAs”.

(6) Along the same lines, related to Lines 434-444, has 2D-HELS MS Seq method been cross validated on a different MS platform?
Response: Yes, in addition to monoisotopic masses (exported by Agilent MassHunter) used here in the 2D-HELS MS Seq, we have also performed full-spectral analysis on using MassWork provided by Cerno Biosciences (Las Vegas, USA) to validate our sequencing data.

(7) For practical purposes, the authors are encouraged to point out the steps where the user can safely stop and store the sample at -80 until the next step.
Response: As per the suggestion, we have added this NOTE in Sections 2.1.5 (Page 5, Line 175) and 4.6 (Page 6, Line 243): “Samples can be stored at -20°C at this stage until the next step is performed.”

(8) I suggest that authors should clearly state limitations and caveats of the 2D-HELS MS Seq method. For example, labeling efficiency and high RNA input requirement.
Response: As per the suggestion, we have added a section to discuss current limitations and aspects need to improve (Page 13, Line 556-564): “With our current workflow, ~100 pmol short RNA (<35 nt) is required for each LC-MS run. The loading amount increases when additional experiments are needed, *e.g.*, for differentiating isomeric nucleotide modifications, and typically up to 400 pmol is required. For sequencing specific tRNA like tRNA^{Phe}, ~1000-2000 pmol sample is needed for its sequencing. However, we expect decreased sample loading requirements

on LC-MS instruments with greater sensitivity. With improvements in sample labeling efficiency, sequencing algorithm, instrument sensitivity and resolution, we expect our method to be applicable to a wider range of RNA samples, especially those with various RNA modifications."

(9) I find many linguistic and grammatical errors along with typos throughout the manuscript. Authors should do a thorough proof-reading of the manuscript to make sure there are no typos and other grammatical errors.

Response: We have proofread the whole manuscript and corrected all the errors in the manuscript.

Supporting Information

2D-HELS MS Seq: A general LC-MS-based method for direct and *de novo* sequencing of RNA mixtures with different nucleotide modifications

AUTHORS AND AFFILIATIONS:

Ning Zhang[1a,2], Shundi Shi[2], Barney Yoo[3], Xiaohong Yuan[1a], Wenjia Li[1b] and Shenglong Zhang[1a, *]

[1a] Department of Biological and Chemical Sciences, New York Institute of Technology, New York, NY, 10023, USA

[1b] Department of Computer Science, New York Institute of Technology, New York, NY, 10023, USA

[2] Department of Chemical Engineering, Columbia University, New York, NY, 10027, USA

[3] Department of Chemistry, Hunter College, City University of New York, New York, NY, 10065, USA

EMAIL ADDRESSES OF CO-AUTHORS:

Ning Zhang (nzhang07@nyit.edu)

Shundi Shi (ss526@columbia.edu)

Barney Yoo (by104@hunter.cuny.edu)

Xiaohong Yuan (xyuan04@nyit.edu)

Wenjia Li (wli20@nyit.edu)

[*]**CORRESPONDING AUTHOR:**

Shenglong Zhang (szhang21@nyit.edu)

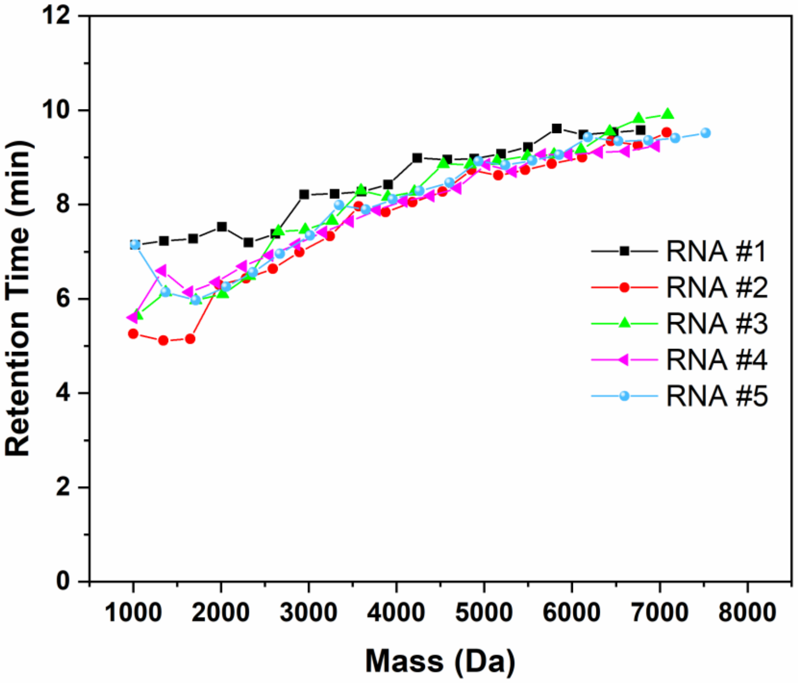


Figure S1. 2D-HELS MS sequencing of 5 mixed RNA strands simultaneously using a biotin tag to label the 3´-ends. Original t_R was displayed without any normalization.

Table S1. LC-MS analysis of 3´-biotin-labeled RNA #1 after streptavidin-aided bead separation followed by subsequent chemical degradation (3´-labeled ladder components of RNA #1, referring to the top curve in Figure 1c).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
19	6781.0733	305.0413	C	6781.0413	9.752	16819442	100	4.72
18	6476.0320	345.0474	G	6475.9924	9.717	247965	84	6.11
17	6130.9846	305.0413	C	6130.9398	9.662	178841	80	7.31
16	5825.9433	329.0525	A	5825.9037	9.782	510096	80	6.80
15	5496.8908	306.0253	U	5496.8566	9.383	262486	99	6.22
14	5190.8655	305.0413	C	5190.8364	9.241	349988	100	5.61
13	4885.8242	306.0253	U	4885.7908	9.135	356118	100	6.84
12	4579.7989	345.0475	G	4579.7738	9.109	386687	100	5.48
11	4234.7514	329.0525	A	4234.7271	9.145	305380	100	5.74
10	3905.6989	305.0413	C	3905.6749	8.575	145505	96	6.14
9	3600.6576	306.0253	U	3600.6373	8.420	195308	100	5.64
8	3294.6323	345.0474	G	3294.6165	8.370	125991	100	4.80
7	2949.5849	329.0525	A	2949.5716	8.339	106993	100	4.51
6	2620.5324	305.0413	C	2620.5193	7.492	90629	100	5.00
5	2315.4911	305.0413	C	2315.4814	7.299	163692	100	4.19
4	2010.4498	329.0525	A	2010.4388	7.625	279963	100	5.47
3	1681.3973	329.0525	A	1681.3891	7.354	183827	100	4.88
2	1352.3448	329.0526	A	1352.3378	7.303	135065	100	5.18
1	1023.2922	329.0525	A	1023.2859	7.219	106700	100	6.16

Output sequence: CGCAUCUGACUGACCAAAA

Table S2. LC-MS analysis of 3´-biotin-labeled RNA #1 after streptavidin-aided bead separation followed by subsequent chemical degradation (5´-unlabeled ladder components of RNA #1, referring to the bottom curve in Figure 1c).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
19	6024.8778	249.0862	A	6024.8483	7.664	14325731	100	4.90
18	5775.7916	329.0525	A	5775.7522	7.701	457844	87	6.82
17	5446.7391	329.0525	A	5446.6965	7.411	417145	100	7.82
16	5117.6866	329.0525	A	5117.6572	7.105	490290	100	5.74
15	4788.6341	305.0413	C	4788.6060	6.685	728135	100	5.87
14	4483.5928	305.0413	C	4483.5657	6.428	481770	100	6.04
13	4178.5515	329.0525	A	4178.5286	6.183	297514	100	5.48
12	3849.4990	345.0475	G	3849.4787	5.653	518403	100	5.27
11	3504.4515	306.0253	U	3504.4331	5.238	614494	100	5.25
10	3198.4262	305.0413	C	3198.4106	4.785	524613	99	4.88
9	2893.3849	329.0525	A	2893.3714	4.341	373933	100	4.67
8	2564.3324	345.0474	G	2564.3219	3.458	509219	100	4.09
7	2219.2850	306.0253	U	2219.2752	2.840	579139	100	4.42
6	1913.2597	305.0413	C	1913.2521	2.081	466058	100	3.97
5	1608.2184	306.0253	U	1608.2123	1.375	372038	80	3.79
4	1302.1931	329.0525	A	1302.1878	0.925	240613	100	4.07
3	973.1406	305.0413	C	973.1367	0.765	208989	100	4.01
2	668.0993	345.0474	G	668.0955	0.652	26061	100	5.69
1	323.0519	305.0413	C	NA*	NA	NA	NA	NA

* NA: Not Analyzed. The 350 Da threshold was set to minimize background ions from the elution buffers. Thus, the masses which are smaller than 350 Da were not detected.

Output sequence: CGCAUCUGACUGACCAAAA

Table S3. LC-MS analysis of 5´-biotin-labeled RNA #1 (5´-labeled ladder components of RNA #1, referring to the bottom ladder curve in black in Figure 1d).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	Theoretical mass	Base mass	Base	MFE mass	t_R	Volume	Quality Score	ppm
19	6600.0415	249.0862	A	6600.0153	10.113	1468018	100	3.97
18	6350.9553	329.0525	A	6350.9006	10.094	139388	80	8.61
17	6021.9028	329.0525	A	6021.8665	9.957	152155	80	6.03
16	5692.8503	329.0525	A	5692.8225	9.806	122377	84	4.88
15	5363.7978	305.0413	C	5363.7567	9.594	255396	100	7.66
14	5058.7565	305.0413	C	5058.7320	9.508	169499	80	4.84
13	4753.7152	329.0525	A	4753.6944	9.449	121869	96	4.38
12	4424.6627	345.0475	G	4424.6389	9.204	222046	100	5.38
11	4079.6152	306.0253	U	4079.5902	9.067	296271	100	6.13
10	3773.5899	305.0413	C	3773.5679	8.937	249085	100	5.83
9	3468.5486	329.0525	A	3468.5308	8.838	185624	100	5.13
8	3139.4961	345.0474	G	3139.4834	8.507	319911	100	4.05
7	2794.4487	306.0253	U	2794.4360	8.288	380189	100	4.54
6	2488.4234	305.0413	C	2488.4134	8.073	317954	100	4.02
5	2183.3821	306.0253	U	2183.3725	7.863	305479	100	4.40
4	1877.3568	329.0525	A	1877.3489	7.642	222446	100	4.21
3	1548.3043	305.0413	C	1548.2982	7.088	361254	100	3.94
2	1243.2630	345.0474	G	1243.2575	6.798	162972	100	4.42
1	898.2156	305.0413	C	898.2105	6.880	88421	100	5.68

Output sequence: CGCAUCUGACUGACCAAAA

Table S4. LC-MS analysis of 5´-biotin-labeled RNA #2 (5´-labeled ladder components of RNA #2, referring to the top ladder curve in red in Figure 1d).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	6898.0505	225.0750	C	6898.0210	10.014	3995416	100	4.28
19	6672.9755	345.0474	G	6673.4755	10.115	92706	80	-74.9
18	6327.9281	305.0413	C	6327.8894	10.117	108088	80	6.12
17	6022.8868	329.0525	A	6022.8313	10.104	133027	100	9.21
16	5693.8343	306.0253	U	5693.7870	9.920	68281	80	8.31
15	5387.8090	305.0413	C	5387.7785	9.850	167081	80	5.66
14	5082.7677	306.0253	U	5082.7314	9.784	170198	100	7.14
13	4776.7424	345.0474	G	4776.7210	9.695	114657	99	4.48
12	4431.6950	329.0526	A	4431.6685	9.629	143358	92	5.98
11	4102.6424	305.0412	C	4102.6199	9.367	245033	100	5.48
10	3797.6012	306.0253	U	3797.5819	9.264	184127	100	5.08
9	3491.5759	345.0475	G	3491.5567	9.131	91691	100	5.50
8	3146.5284	329.0525	A	3146.5054	9.028	187937	100	7.31
7	2817.4759	305.0413	C	2817.4633	8.675	288050	100	4.47
6	2512.4346	305.0413	C	2512.4233	8.509	138698	100	4.50
5	2207.3933	305.0413	C	2207.3835	8.335	192998	100	4.44
4	1902.3520	345.0474	G	1902.3433	8.161	149466	100	4.57
3	1557.3046	329.0525	A	1557.2976	8.042	133349	100	4.49
2	1228.2521	306.0253	U	1228.2455	7.618	188828	100	5.37
1	922.2268	329.0525	A	922.2213	7.434	86674	100	5.96

Output sequence: AUAGCCCAGUCAGUCUACGC

Table S5. LC-MS analysis of a 1 ψ-containing RNA #6 (ψ unconverted ladder components in the 5´ ladder of RNA #6, referring to the bottom ladder curve in black in Figure 2b).

Theoretical				**Extracted data file after LC/MS analysis**				**Error**
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	6345.9028	265.0811	G	6345.9217	11.736	41088112	100	-2.98
19	6080.8217	329.0525	A	6080.8255	11.769	2582596	100	-0.62
18	5751.7692	345.0474	G	5751.7749	11.496	2169051	100	-0.99
17	5406.7218	306.0253	U	5406.7209	11.315	2126771	100	0.17
16	5100.6965	319.057	m^{5}C	5100.6941	11.167	1149416	100	0.47
15	4781.6395	329.0525	A	4781.6402	10.970	2692877	100	-0.15
14	4452.5870	306.0253	U	4452.5866	10.566	5448251	100	0.09
13	4146.5617	306.0253	U	4146.5603	10.343	4115258	100	0.34
12	3840.5364	329.0526	A	3840.5352	10.141	2038738	100	0.31
11	3511.4838	305.0413	C	3511.4836	9.610	1167942	100	0.06
10	3206.4425	305.0412	C	3206.4401	9.331	3422282	100	0.75
9	2901.4013	329.0526	A	2901.3988	9.067	2391922	100	0.86
8	2572.3487	306.0253	Unconverted ψ	2572.3468	8.328	4952174	100	0.74
7	2266.3234	306.0253	U	2266.3215	7.944	4534905	100	0.84
6	1960.2981	345.0474	G	1960.2956	7.360	3437270	100	1.28
5	1615.2507	305.0413	C	1615.2481	6.693	4151449	100	1.61
4	1310.2094	305.0413	C	1310.2062	5.915	1289241	87	2.44
3	1005.1681	329.0525	A	1005.1655	4.416	913589	100	2.59
2	676.1156	329.0525	A	676.1140	3.321	748977	100	2.37
1	347.0631	329.0525	A	NA*	NA	NA	NA	NA

* NA: Not Analyzed. The 350 Da threshold was set to minimize background ions from the elution buffers. Thus, the masses which are smaller than 350 Da were not detected.

Output sequence: AAACCGUψACCAUUAm^{5}CUGAG

Table S6. LC-MS analysis of a 1 ψ-containing RNA #6 (ladder components with CMC-converted ψ in the 5´ ladder of RNA #6, referring to the top ladder curve in red in Figure 2b)

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	6597.1025	265.0811	G	6597.1125	13.985	60627484	100	-1.52
19	6332.0214	329.0525	A	6332.0201	13.979	1541470	100	0.21
18	6002.9689	345.0474	G	6002.9756	13.816	2147847	89	-1.12
17	5657.9215	306.0253	U	5657.9243	13.742	2608610	100	-0.49
16	5351.8962	319.057	m^{5}C	5351.8960	13.695	2110248	100	0.04
15	5032.8392	329.0525	A	5032.8400	13.633	1907945	100	-0.16
14	4703.7867	306.0253	U	4703.7861	13.394	4110706	88	0.13
13	4397.7614	306.0253	U	4397.7599	13.320	2867370	100	0.34
12	4091.7361	329.0526	A	4091.7361	13.283	1855682	100	0.00
11	3762.6835	305.0413	C	3762.6830	12.962	2817838	100	0.13
10	3457.6422	305.0412	C	3457.6396	12.878	1149319	100	0.75
9	3152.6010	329.0526	A	3152.5974	12.934	746862	100	1.14
8	2823.5485	557.2251	Converted ψ	2823.5455	12.380	2149383	100	1.06
7	2266.3234	306.0253	U	2266.3213	7.944	4767282	100	0.93
6	1960.2981	345.0474	G	1960.2956	7.360	3433416	100	1.28
5	1615.2507	305.0413	C	1615.2481	6.694	4174772	100	1.61
4	1310.2094	305.0413	C	1310.2071	5.917	806139	87	1.76
3	1005.1681	329.0525	A	1005.1655	4.416	913589	100	2.59
2	676.1156	329.0525	A	676.1140	3.321	743305	100	2.37
1	347.0631	329.0525	A	NA*	NA	NA	NA	NA

* NA: Not Analyzed. The 350 Da threshold was set to minimize background ions from the elution buffers. Thus, the masses which are smaller than 350 Da were not detected.

Output sequence: AAACCGUψACCAUUAm^{5}CUGAG

Table S7. LC-MS analysis of 3´-biotin-labeled RNA #1, showing its ladder components (referring to the ladder curve in black in Figure 3).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
19	6781.0733	305.0413	C	6781.0426	9.576	35286012	100	4.53
18	6476.0320	345.0474	G	6475.9985	9.535	23351	60	5.17
17	6130.9846	305.0413	C	6130.9933	9.473	50125	90	-1.42
16	5825.9433	329.0525	A	5825.9244	9.634	55880	80	3.24
15	5496.8908	306.0253	U	5496.8590	9.218	633795	80	5.79
14	5190.8655	305.0413	C	5190.8470	9.078	849742	100	3.56
13	4885.8242	306.0253	U	4885.7976	8.976	1193120	100	5.44
12	4579.7989	345.0475	G	4579.7742	8.951	1191558	100	5.39
11	4234.7514	329.0525	A	4234.7340	8.989	1196633	100	4.11
10	3905.6989	305.0413	C	3905.6808	8.420	729180	100	4.63
9	3600.6576	306.0253	U	3600.6382	8.275	605689	100	5.39
8	3294.6323	345.0474	G	3294.6179	8.229	935654	100	4.37
7	2949.5849	329.0525	A	2949.5713	8.210	903559	100	4.61
6	2620.5324	305.0413	C	2620.5217	7.376	587699	100	4.08
5	2315.4911	305.0413	C	2315.4825	7.191	700118	100	3.71
4	2010.4498	329.0525	A	2010.4378	7.527	1052796	100	5.97
3	1681.3973	329.0525	A	1681.3901	7.273	714971	100	4.28
2	1352.3448	329.0526	A	1352.3387	7.230	447072	100	4.51
1	1023.2922	329.0525	A	1023.2881	7.148	736463	100	4.01

Output sequence: CGCAUCUGACUGACCAAAA

Table S8. LC-MS analysis of 3´-biotin-labeled RNA #2, showing its ladder components (referring to the ladder curve in red in Figure 3).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	7079.0823	329.2088	A	7079.0513	9.529	34343980	100	4.38
19	6750.0298	306.1667	U	6749.9875	9.259	170073	78	6.27
18	6444.0045	329.2088	A	6443.9653	9.344	934361	97	6.08
17	6114.9519	345.2077	G	6114.9082	9.000	176482	94	7.15
16	5769.9045	305.1828	C	5769.8590	8.867	537259	80	7.89
15	5464.8632	305.1828	C	5464.8338	8.733	381043	100	5.38
14	5159.8219	305.1827	C	5159.7998	8.619	939572	99	4.28
13	4854.7806	329.2088	A	4854.7556	8.734	1104050	100	5.15
12	4525.7281	345.2078	G	4525.7027	8.273	799528	100	5.61
11	4180.6807	306.1667	U	4180.6575	8.047	727253	100	5.55
10	3874.6554	305.1828	C	3874.6361	7.836	1007297	100	4.98
9	3569.6141	329.2087	A	3569.5985	7.960	1323892	100	4.37
8	3240.5616	345.2078	G	3240.5458	7.328	854305	100	4.88
7	2895.5141	306.1668	U	2895.5009	6.991	838944	100	4.56
6	2589.4888	305.1827	C	2589.4785	6.639	1076014	100	3.98
5	2284.4476	306.1668	U	2284.4388	6.433	1085561	100	3.85
4	1978.4223	329.2088	A	1978.4152	6.298	1224106	100	3.59
3	1649.3697	305.1827	C	1649.3632	5.150	443067	100	3.94
2	1344.3284	345.2078	G	1344.3229	5.115	530069	100	4.09
1	999.2810	305.1827	C	999.2764	5.258	300175	100	4.60

Output sequence: AUAGCCCAGUCAGUCUACGC

Table S9. LC-MS analysis of 3´-biotin-labeled RNA #3, showing its ladder components (referring to the ladder curve in green in Figure 3).

Theoretical				**Extracted data file after LC/MS analysis**				**Error**
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	7088.0826	329.0525	A	7088.0479	9.902	18422776	100	4.90
19	6759.0301	329.0525	A	6758.9878	9.816	342458	82	6.26
18	6429.9776	329.0525	A	6429.9401	9.553	297978	100	5.83
17	6100.9251	305.0413	C	6100.8860	9.162	176200	80	6.41
16	5795.8838	305.0413	C	5795.8502	9.059	325811	100	5.80
15	5490.8425	345.0475	G	5490.8084	9.029	561379	99	6.21
14	5145.7950	306.0253	U	5145.7640	8.927	543764	100	6.02
13	4839.7697	306.0253	U	4839.7382	8.852	751511	100	6.51
12	4533.7444	329.0525	A	4533.7170	8.857	916467	100	6.04
11	4204.6919	305.0413	C	4204.6726	8.273	363029	100	4.59
10	3899.6506	305.0413	C	3899.6323	8.164	664338	100	4.69
9	3594.6093	329.0525	A	3594.5912	8.300	1247513	100	5.04
8	3265.5568	306.0253	U	3265.5400	7.653	597972	100	5.14
7	2959.5315	306.0253	U	2959.5186	7.464	985122	100	4.36
6	2653.5062	329.0525	A	2653.4963	7.431	1500526	100	3.73
5	2324.4537	305.0413	C	2324.4444	6.486	663475	100	4.00
4	2019.4124	306.0253	U	2019.4039	6.101	752760	100	4.21
3	1713.3871	345.0474	G	1713.3811	5.973	1299628	100	3.50
2	1368.3397	329.0525	A	1368.3335	6.144	379728	100	4.53
1	1039.2872	345.0474	G	1039.2820	5.644	273139	100	5.00

Output sequence: AAACCGUUACCAUUACUGAG

Table S10. LC-MS analysis of 3´-biotin-labeled RNA #4, showing its ladder components (referring to the ladder curve in pink in Figure 3).

Theoretical				Extracted data file after LC/MS analysis				Error
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
20	6954.9836	345.0475	G	6954.9478	9.243	16978916	100	5.15
19	6609.9361	305.0412	C	6609.8899	9.131	184784	80	6.99
18	6304.8949	345.0475	G	6304.8568	9.109	510790	80	6.04
17	5959.8474	306.0253	U	5959.7956	9.056	393186	90	8.69
16	5653.8221	329.0525	A	5653.7838	9.059	830821	100	6.77
15	5324.7696	305.0413	C	5324.7319	8.701	496925	98	7.08
14	5019.7283	329.0525	A	5019.6982	8.848	1059427	100	6.00
13	4690.6758	306.0253	U	4690.6470	8.345	581020	82	6.14
12	4384.6505	305.0413	C	4384.6245	8.185	852527	100	5.93
11	4079.6092	306.0253	U	4079.5872	8.071	872930	100	5.39
10	3773.5839	306.0253	U	3773.5632	7.884	880358	100	5.49
9	3467.5586	305.0413	C	3467.5339	7.639	168485	97	7.12
8	3162.5173	305.0413	C	3162.4881	7.411	503294	100	9.23
7	2857.4760	305.0413	C	2857.4625	7.156	851140	100	4.72
6	2552.4347	305.0412	C	2552.4231	6.920	1065610	100	4.54
5	2247.3935	306.0253	U	2247.3838	6.690	1189236	100	4.32
4	1941.3682	306.0253	U	1941.3605	6.350	1445336	100	3.97
3	1635.3429	306.0254	U	1635.3384	6.009	22256	85	2.75
2	1329.3175	329.0525	A	1329.3120	6.598	1296266	100	4.14
1	1000.2650	306.0253	U	1000.2606	5.604	422194	100	4.40

Output sequence: GCGUACAUCUUCCCCUUUAU

Table S11. LC-MS analysis of 3´-biotin-labeled RNA #5, showing its ladder components (referring to the ladder curve in light blue in Figure 3).

Theoretical				**Extracted data file after LC/MS analysis**				**Error**
Fragments	**Theoretical mass**	**Base mass**	**Base**	**MFE mass**	**t_R**	**Volume**	**Quality Score**	**ppm**
21	7522.1050	345.0475	G	7522.0681	9.519	21361914	100	4.91
20	7177.0575	305.0413	C	7176.9933	9.405	68800	60	8.95
19	6872.0162	345.0474	G	6871.9775	9.363	252280	88	5.63
18	6526.9688	345.0474	G	6526.9161	9.345	403291	100	8.07
17	6181.9214	329.0526	A	6181.8847	9.425	1246921	100	5.94
16	5852.8688	306.0253	U	5852.8226	9.054	263228	92	7.89
15	5546.8435	306.0253	U	5546.8116	8.935	1204009	100	5.75
14	5240.8182	306.0253	U	5240.7914	8.839	944494	100	5.11
13	4934.7929	329.0525	A	4934.7693	8.917	796848	100	4.78
12	4605.7404	345.0474	G	4605.7119	8.465	673185	100	6.19
11	4260.6930	305.0413	C	4260.6681	8.290	729523	100	5.84
10	3955.6517	306.0253	U	3955.6308	8.107	803678	100	5.28
9	3649.6264	305.0413	C	3649.6084	7.894	1056834	100	4.93
8	3344.5851	329.0525	A	3344.5687	7.990	1336987	100	4.90
7	3015.5326	345.0474	G	3015.5131	7.343	882742	100	6.47
6	2670.4852	306.0253	U	2670.4731	6.959	659989	100	4.53
5	2364.4599	306.0253	U	2364.4502	6.560	845446	100	4.10
4	2058.4346	345.0475	G	2058.4278	6.256	752026	100	3.30
3	1713.3871	345.0474	G	1713.3811	5.973	1299628	100	3.50
2	1368.3397	345.0475	G	1368.3335	6.144	379728	100	4.53
1	1023.2922	329.0525	A	1023.2881	7.148	736463	100	4.01

Output sequence: GCGGAUUUAGCUCAGUUGGGA