I, Siwei Li, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Chemistry.

It is entitled:
High Throughput Automated Comparative Analysis of RNAs Using Isotope Labeling and LC-MS/MS

Student’s name: Siwei Li

This work and its defense approved by:

Committee chair: Patrick Limbach, Ph.D.
Committee member: William Heineman, Ph.D.
Committee member: Pearl Tsang, Ph.D.
High Throughput Automated Comparative Analysis of RNAs
Using Isotope Labeling and LC-MS/MS

A dissertation submitted to the
Graduate School
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy (PhD)

in the Department of Chemistry
of the McMicken College of Arts and Sciences
November 2013
by

Siwei Li

B.S., Biotechnology, Nanjing University, China 2008
November 2013

Committee Chair: Patrick A. Limbach, PhD
Abstract

Mass spectrometry is a powerful technique for the characterization of ribonucleic acids. The focus of this work is developing mass spectrometric methods for high throughput automated analysis of ribonucleic acids (RNAs) by comparative approaches, where only differences within the mass spectral data need to be analyzed.

This method allows sequence or modification information from a previously uncharacterized RNA to be obtained by direct comparison with a reference RNA, whose sequence or modification information is known. This simple and rapid method is enabled by the differential labeling of two RNA samples. One sample, the reference RNA, is labeled with $^{16}$O during enzymatic digestion. The second sample, the candidate or unknown RNA, is labeled with $^{18}$O. By combining the two digests, digestion products that share the same sequence or post-transcriptional modification(s) between the reference and candidate will appear as doublets separated by 2 Da. Sequence or modification differences between the two will generate singlets that can be further characterized to identify how the candidate sequence differs from the reference. I illustrate application of this approach for sequencing individual RNAs and demonstrate how this method can be used to identify sequence-specific differences in RNA modification.

Using CARD approach, ca. 80% of the tRNAs from the bacterium *Citrobacter koseri* can be sequenced using ribonuclease T1 with *Escherichia coli* tRNAs as the reference. During these studies, a sequence error for *Escherichia coli* tRNA-Thr1 was discovered, and the correct sequence for that tRNA was confirmed by this method.

For many applications, the differences between two samples will be minor meaning that much of the mass spectral data will be doublets with only a few singlets. The challenge in
data analysis is to rapidly identify and characterize these singlets. To address this challenge, an algorithm for automated data analysis was developed based on Microsoft Visual Basic for Application (VBA) macro-program, simplifying and automating the examination of LC-MS data for identification of isotopically labeled doublets and singlets. The automated processing steps in this program drastically reduce the need for manual processing of LC-MS data by applying filters to remove thousands of doublets and interferences in the mass spectral data.

In addition to these method and algorithm developments, during the course of comparative analysis of RNA, many predicted tRNA digestion products share identical $m/z$ values yet different sequences. These tRNA sequence isomers can hinder the accuracy of unknown tRNA identification. To overcome this limitation, tandem mass spectrometric method was developed to identify those interference-prone singlets based on unique isotopically labeled fragments in tandem mass spectra.
Acknowledgement

First and foremost I would like to express my sincere gratitude and appreciation to my advisor, Dr. Patrick Limbach, for his expert direction, understanding, and support during my throughout time in University of Cincinnati, especially when I was taking care of my sick parents in China. Thank him for his tolerance and endurance. Special thanks him for his support and guidance in my new proposals and new ideas.

I would like to thank my committee members, Dr. Apryll Stalcup and Dr. Pearl Tsang, for their advice and guidance concerning my annual talk and dissertation. I also have gratitude to Dr. William Heineman for my committee member at my defenses. Special thanks to Dr. Peter Padolik and Dr. Michael Baldwin for their understanding and support when I was taking care of my sick parents. In addition, special thanks to Dr. Balu Addepalli for returning to the lab and saving one of my most important experiments. Very special thanks to the entire chemistry department of University of Cincinnati for their support during my PhD program time.

Very special thanks to the whole Limbach group members, both past and present, for their suggestion and comments in my past annual talks practices. I am so happy to work with all of you. Thanks to Collin Wetzel for his discussions and collaboration with our multiple collaborative research projects.

Finally, I would like to thank my parents, Renqing Li and Xinhua Luo, for their love, understanding and support to me as well as my dreams. My mother’s positive and optimistic attitude to her sickness makes me stronger to face all the challenges in my life. Without her understanding and support, I would not have been able to complete my PhD study.
Table of Contents

Abstract .................................................................................................................................................. ii
Acknowledgements ................................................................................................................................... v
List of Figures .......................................................................................................................................... 4
List of Tables ........................................................................................................................................... 9

Chapter 1 Introduction .............................................................................................................................. 11
1.1 Introduction ........................................................................................................................................ 11
1.2 RNA and post-transcriptional modification ..................................................................................... 13
1.3 Methods for characterization of RNA ............................................................................................ 16
1.4 Bioinformatics tools for data analysis of mass spectral data from nucleic acids .................................. 25
1.5 Background for dissertation research ............................................................................................ 29
1.6 Conclusion .......................................................................................................................................... 32

Chapter 2 Method for Comparative Analysis of Ribonucleic Acids Using Isotope Labeling and Mass Spectrometry ..................................................................................................................... 33
2.1 Introduction ........................................................................................................................................ 33
2.2 Experimental ..................................................................................................................................... 33
2.3 Results and discussion ..................................................................................................................... 37
2.4 Conclusion .......................................................................................................................................... 48
Chapter 3 Mass Spectrometry Sequencing of Transfer Ribonucleic Acids by the Comparative Analysis of RNA Digests (CARD) Approach .......... 50

3.1 Introduction ........................................................................................................... 50
3.2 Experimental ........................................................................................................ 51
3.3 Results and discussion ....................................................................................... 52
3.4 Conclusion ............................................................................................................ 73

Chapter 4 A Novel Bioinformatics Tool for Targeted and Untargeted RNA in LC-MS ................................................................................. 75

4.1 Introduction ........................................................................................................... 75
4.2 Experimental ........................................................................................................ 75
4.3 Results and discussion ....................................................................................... 76
4.4 Conclusion ............................................................................................................ 87

Chapter 5 Identification of RNA Sequence Isomer by Isotope Labeling and LC-MS/MS ................................................................................... 88

5.1 Introduction ........................................................................................................... 88
5.2 Experimental ........................................................................................................ 89
5.3 Results and discussion ....................................................................................... 90
5.4 Conclusion ............................................................................................................ 98
List of Figures

Figure 1.1 Structure of RNA .................................................................12
Figure 1.2 Representative structures of modified nucleosides .........................13
Figure 1.3 The structure of domains of tRNA. The crystallographic structure of yeast tRNA\textsuperscript{Phe} is on the left [6]. The cloverleaf secondary structure is in the center. The sequence and secondary structure of the anticodon stem and loop domain from human tRNA\textsuperscript{Lys} is on the right. The figure was adapted from reference [2]. .................................14
Figure 1.4 Overview of mass spectrometric methods for characterization of RNA ......17
Figure 1.5 Strategy for complete characterization of modification in RNA by RNA mass mapping and LC-MS/MS (Figure from reference [38]) ..........................20
Figure 1.6 Nomenclature for RNA fragmentation when performing MS/MS ...............21
Figure 1.7 A. Comparative analysis of modified nucleosides of tRNA\textsuperscript{Sec} by LC-MS; B. Comparative analysis of sequence-specific digestion products of tRNA\textsuperscript{Sec} by MALDI-MS (Figure from reference [37]) .........................................................24
Figure 1.8 Flowchart for MS analysis of RNA. The types of software that are available for the processing of mass spectra of RNAs are shown in underlined bold type. (Figure was adapted from reference [34]) .........................................................25
Figure 1.9 The endonuclease cleavage mechanism. The 3'-cyclic phosphate is an intermediate in enzymatic digestion process. After oxygen from water is incorporated into phosphate group, a stable isotope label from oxygen can be available in the 3'-phosphate (II) .......29
Figure 1.0 General overview of \textsuperscript{18}O labeling and MS approach for quantitation of RNA (Figure from reference [62]) ..............................................................30
Figure 2.1 Schematic outline of comparative sequencing by isotope-labeling and LC-MS where \textit{E. coli} serves as the reference organism and \textit{C. koseri} serves as the candidate (unknown) to be sequenced. tRNA endonuclease digestion products that are equivalent between organisms will appear as doublets (separated by 2 Da) in the mass spectral data; digestion products that are different between the two organisms will appear as a singlet ..........38
Figure 2.2 Singlets detected from comparative sequencing of tRNA\(^{G^3}\) from \(C.\) \(koseri\) using tRNA\(^{G^3}\) of \(E.\) \(coli\) as the reference sequence. (a) Singlet U[s\(^{4}\)U]AACAAAGp (m/z 1469.6, 2- charge) arising from purified \(E.\) \(coli\) tRNA\(^{G^3}\) after labeling \(C.\) \(koseri\) with \(^{18}\)O.  (b) The same singlet after labeling \(E.\) \(coli\) with \(^{18}\)O.  The +1 increase in the m/z isotopic envelope confirms the singlet arises from \(E.\) \(coli\).  The low abundance peaks at < m/z 1470.6 arise from the use of < 100% \(^{18}\)O-labeled water.  (c) Singlet U[s\(^{4}\)U]AGp (m/z 660.0, 2- charge) from purified \(C.\) \(koseri\) tRNA\(^{G^3}\) after labeling \(C.\) \(koseri\) with \(^{18}\)O.  (d) The same singlet after labeling \(E.\) \(coli\) with \(^{18}\)O.  The -1 decrease in the m/z isotopic envelope confirms the singlet arises from \(C.\) \(koseri\).

Figure 2.3 Collision-induced dissociation mass spectrum of the singlet U[s\(^{4}\)U]AACAAAGp from \(E.\) \(coli\).  ....................................................................................................................................................................................41

Figure 2.4 Collision-induced dissociation mass spectrum of the singlet U[s\(^{4}\)U]AGp from \(C.\) \(koseri\).  ....................................................................................................................................................................................42

Figure 2.5 Collision-induced dissociation mass spectrum of the singlet CAAAGp from \(C.\) \(koseri\).  ....................................................................................................................................................................................43

Figure 2.6 Representative singlets detected from comparative analysis of total tRNAs from \(E.\) \(coli\) miaA-\(\Delta\) mutant using total tRNAs of \(E.\) \(coli\) MG1655 as the reference. (a) Singlet ACU[Q]UA[ms\(^{2}\)i6A]\(\Psi\)CUGp (m/z 1365.2, 3- charge) arising from \(E.\) \(coli\) MG1655 after labeling miaA-\(\Delta\) mutant with \(^{18}\)O.  (b) The same singlet after labeling \(E.\) \(coli\) MG1655 with \(^{18}\)O.  The +0.66 increase in the m/z isotopic envelope, corresponding to 2 Da at the -3 charge state, confirms the singlet arises from \(E.\) \(coli\) MG1655.  (c) Collision-induced dissociation mass spectrum of the singlet ACU[Q]UA[ms\(^{2}\)i6A]\(\Psi\)CUGp from \(E.\) \(coli\) MG1655.  The major peak corresponds to a loss of 114 Da from the molecular ion.  This loss is consistent with loss of the 2-methylthio-6-isopentenyl group during CID.  A listing of all m/z values used for sequence assignment can be found in Table 2.5.........................44

Figure 2.7. Representative singlets detected from comparative analysis of total tRNAs from \(E.\) \(coli\) miaA-\(\Delta\) mutant using total tRNAs of \(E.\) \(coli\) MG1655 as the reference. (a) Singlet ACU[Q]UAAA[\(\Psi\)]CUGp (m/z 1327.2, 3- charge) from miaA-\(\Delta\) mutant after labeling miaA-\(\Delta\) mutant with \(^{18}\)O.  (b) The same singlet after labeling \(E.\) \(coli\) MG1655 with \(^{18}\)O.  The -0.66 increase in the m/z isotopic envelope, corresponding to 2 Da at the -3 charge state, confirms the singlet arises from miaA-\(\Delta\) mutant.  (c) Collision-induced dissociation mass spectrum of the singlet ACU[Q]UAAA[\(\Psi\)]CUGp from miaA-\(\Delta\) mutant.  The major peak corresponds to a loss of 175 Da from the molecular ion.  The identity of this fragment ion was not determined in this work.  A listing of all m/z values used for sequence assignment can be found in Table 2.6.........................45
Figure 3.1. Schematic outline of comparative sequencing by isotope-labeling and LC-MS where *E. coli* serves as the reference organism and *C. koseri* serves as the candidate (unknown) to be sequenced. tRNA endonuclease digestion products that are equivalent between organisms will appear as doublets (separated by 2 Da) in the mass spectral data; digestion products that are different between the two organisms will appear as a singlet.

Figure 3.2. Procedure for generating anticipated tRNA sequences from the candidate organism. (a) The reference (RNA) sequence and the candidate gene (DNA) sequence are aligned. (b) Where base identity is conserved between the two sequences, known post-transcriptional modifications present on the reference tRNA are placed onto the candidate tRNA sequence to generate the final, predicted candidate tRNA sequence. These sequences can then be digested in silico with the RNase of choice to predict the minimum number of singlets that should appear in the mass spectrum.

Figure 3.3 Modified nucleosides in *Citrobacter koseri* total tRNAs. 1, dihydrouridine (D). 2, pseudouridine (Ψ). 3, 5-carboxymethoxyuridine (cmo5U). 4, 3-(3-amino-3-carboxypropyl)uridine (acp3U). 5, 2-thiocytidine (s'C). 6, 5-methyluridine (m5U). 7, (unknown). 8, 2'-O-methylguanosine (Gm). 9, 1-methylguanosine (m1G). 10, N6-threonylcarbamoyladenosine (t6A). 11, 2-methyladenosine (m2A). 12, N6-methyladenosine (m6A). *, 2-methylthio-N6-isopentenyladenosine (ms2i6A)(data not shown in UV chromatogram).

Figure 3.4. Mass spectra corresponding to a detected singlet when (a) *C. koseri* is labeled with 18O and (b) *E. coli* is labeled with 18O. The singlet U[s4U]AACAAAGp (m/z 1469.6, 2-charge) arises from the *E. coli* tRNA-Cys(GCA) as confirmed by the +1 increase in the m/z isotopic envelope after 18O-labeling of *E. coli*.

Figure 3.5 Extracted ion chromatograms corresponding to the anticipated doubly- and triply-charged ions of AC[s3C][U][mnm5U][t6A]AGp, a singlet from tRNA-Arg(UCU), at 5 μg and 20 μg of *E. coli* total tRNA after RNase T1 digestion loaded on column. The XIC response at 31.5 min corresponds to a doubly-charged ion consistent with a base composition of C2U2A2Gp + methyl, which is not the anticipated singlet.

Figure 3.6. Mass spectra corresponding to the anticipated singlet of [m7G]UCCCCCAGp from *C. koseri* tRNA-Arg(UCU) when (a) *C. koseri* is labeled with 18O and (b) *E. coli* is labeled with 18O. In this case, the doubly charged ion is characterized by a doublet rather than the expected singlet.

Figure 3.7. (a) Sequence of tRNA-Thr1 (GGU) obtained from the tRNAdb 2009 database (http://trnadb.bioinf.uni-leipzig.de/).[11] (b) Sequence alignment of tRNA-Thr1 (GGU) genes from *E. coli* K12 W3110 strain and *C. koseri*. The two genes are identical. (c) The corrected sequence for *E. coli* tRNA-Thr1 (GGU).

Figure 3.8 Mass spectrum corresponding to the anticipated singlet [DUGp] from *C. koseri* tRNA-Asp(GUC). A number of interfering ions, likely originating from the ubiquitous digestion products CCGp (m/z 972.1), (CU)Gp (m/z 973.1), and UUGp (m/z 974.1), are present and can interfere with detection of the DUGp singlet.
Figure 3.9. Mass spectra corresponding to anticipated singlet from *C. koseri* tRNA-Leu(CAG) when (a) *C. koseri* is labeled with \(^{18}\text{O}\) and (b) *E. coli* is labeled with \(^{18}\text{O}\). The tRNA-Leu(CAG) singlet is predicted at \(m/z\) 1566.2 (2- charge) when *C. koseri* is labeled with \(^{16}\text{O}\) (E. coli with \(^{18}\text{O}\)). Interfering doublets arising from tRNA-Gly(CCC) at \(m/z\) 1567.2 (2-charge) and 5S rRNA at \(m/z\) 1566.6 (2- charge) hinder identification of the tRNA-Leu(CAG) singlet because of isotopic overlap when *C. koseri* is labeled with \(^{18}\text{O}\).

Figure 3.10. Mass spectra corresponding to the anticipated singlet CAAAGp (\(m/z\) 827.5, 2-charge) when (a) *C. koseri* is labeled with \(^{18}\text{O}\) and (b) *E. coli* is labeled with \(^{18}\text{O}\). Although the isotopic envelope shifts by -1 \(m/z\) unit when *C. koseri* is labeled with \(^{16}\text{O}\), interfering RNase T1 digestion products hinder unequivocal assignment of this singlet.

Figure 3.11. (a) Collision-induced dissociation mass spectrum of the singlet U[s4U]AACAAAGp from *E. coli* total tRNAs. (b) Collision-induced dissociation mass spectrum of the doublet [m7G]UCCCCAGp.

Figure 3.12. Calculated codon usage frequencies for *E. coli, C. koseri* and *T. thermophilus*. Comparative sequencing is more effective when the codon usage frequencies of the reference and candidate organisms are similar.

Figure 4.1. Overview of algorithm for identifying isotopically labeled singlet digestion products in LC-MS data.

Figure 4.2. The linear regression line of b value and mass values of oligonucleotides.


Figure 4.5. The representative mass spectrum of false assigned singlet from script.


Figure 4.7. Mass spectrum of CA[i6A]A[Ψ]CCGp from single *C. koseri* total tRNAs.

Figure 4.8. Full sequences of tRNA\(^{\text{Cys}}\) from *C. koseri*.

Figure 5.1. Identification of sequence isomers by tracking isotope labeling using MS/MS. Identical sequences will yield doublet y-type fragment ions, separated by 2 Da due to \(^{16}\text{O}\)- and \(^{18}\text{O}\)- labeling. Different RNA sequence isomer will yield singlet y-type ions.

Figure 5.2. Optimization of isolation width during MS/MS to enable detection of the isotopic doublet present on y-type fragment ions.

Figure 5.3. The sequences of tRNAs-Glu2 and Tyr1. The bold are the sequences of the digestion product sequence isomers ACGp and CAGp generated by RNase T1.
Figure 5.4 Mass spectra of two oligonucleotides ACGp and CAGp

Figure 5.5 Tandem mass spectra illustrating the identification of ACGp in the presence of the interfering sequence isomer, CAGp, using isotope-labeling and tandem mass spectrometry. (a) Collision induced dissociation mass spectrum of oligonucleotides from tRNA-TyrI,GluII with $^{16}$O and tRNA-GluII with $^{18}$O. (b) Collision induced dissociation mass spectrum of multiple oligonucleotides from tRNA-TyrI,GluII with $^{18}$O and tRNA-TyrI with $^{16}$O.

Figure 5.6 Identification of tRNA-Cys CCAAUAGp from E. coli miaA-Δ mutant in presence of interfering sequence isomers CCAAUAGp and interfering digestion product AAUACCUGp (a)(c) Collision induced dissociation mass spectrum of multiple oligonucleotides from E. coli wt with $^{16}$O and miaA-Δ mutant with $^{18}$O. (b)(d) Collision induced dissociation mass spectrum of multiple oligonucleotides from E. coli wt with $^{18}$O and miaA-Δ mutant with $^{16}$O.

Figure 6.1 Comparison of the doublet [m7G]UCCCCAGp of tRNA-Thr(GGU) cultured in natural culture media and isotope enriched media.
List of Tables

Table 1.1 Software for identification of nucleic acids based on mass spectrometry…….…..26
Table 2.1 Listing of all m/z values used in oligonucleotide sequence assignment………....41
Table 2.2 Listing of all m/z values used in oligonucleotide sequence assignment………....42
Table 2.3 Listing of all m/z values used in oligonucleotide sequence assignment………....42
Table 2.4 RNase T1 singlets detected during comparative analysis of E. coli wt MG1655 tRNAs and miaA-Δ mutant (JW4129-2) tRNAs…………………………………………..46
Table 2.5 Listing of all m/z values used in oligonucleotide sequence assignment of data shown in Figure 2.6c……………………………………………………………….…47
Table 2.6 Listing of all m/z values used in oligonucleotide sequence assignment of data shown in Figure 2.7c………………………………………………………………….48
Table 3.1 Basic Local Alignment Search Tool (BLAST) results for searching 14 randomly chosen E. coli tRNA genes against the C. koseri genome. The value reported is the %sequence homology between the E. coli gene and that found in C. koseri………………53
Table 3.2 Predicted singlets arising during comparative sequencing of E. coli and C. koseri total tRNAs when tRNAs digested using the denoted RNases. Sequence difference identified by bold, underlined text. RNase T1 interference-free singlets are denoted by an asterisk,*…………………………………………….55
Table 3.3 tRNA Sequences of Citrobacter koseri……………………………………..69
Table 3.4 tRNA Sequences of Salmonella enterica……………………………………70
Table 4.1 The complete output from the script after processing of LC-MS data from tRNA-Cys of E. coli and C. koseri. This script identifies m/z values that shift between labeling experiments as well as differences in A and A+2 isotope peak contributions. The direction of the m/z shift can be used to identify the source of the singlet. ………..………..………..………..………..………..………..………..79
Table 4.2 The identification of the predicted singlets of tRNA Cys from E. coli and C. koseri by VBA script…………………………………………………………………………..80
Table 4.3 The complete output from the script after processing of LC-MS data from total tRNAs of E. coli and C. koseri. This script identifies m/z values that shift between labeling experiments as well as differences in A and A+2 isotope peak contributions. The direction of the m/z shift can be used to identify the source of the singlet………………83
Table 4.4 Comparison of the singlets identified using the script outlined in Figure 4 to those that are identified by comparing tRNA gene sequences between two organisms. The bottom singlets were identified from within the LC-MS datasets from RNase T1 digestion of total tRNAs. …………………………………………………………86
Table 5.1 RNA sequence isomers predicted based on RNase T1 digestion of *E. coli* wt and miaA-Δ mutant tRNAs.

Table 5.2 Expected y-type fragment ion *m/z* values for CAA[Ψ]CCGp, CCAUAGp and AAUACCUGp.
Chapter 1 Introduction

1.1 Introduction

Among all known biomolecules within living cells, ribonucleic acids (RNAs) are the only ones that store genetic information and act as catalysts. Because RNAs are post-transcriptionally modified[1] and these modifications have structural and functional significance,[2] approaches that enable the sensitive and accurate determination of RNA sequence and extent of RNA modification are required.

The overall goal of this dissertation is to develop mass spectrometric approaches for high-throughput automated comparative analysis of RNAs. To accomplish this goal, two major aims were developed. First, the mass spectrometric method for high-throughput comparative analysis of RNA must be developed without requiring purification of individual RNAs. Second, an approach for the rapid processing of data from large mass spectral datasets must be developed.

This chapter discusses the significance of characterizing RNA sequence and post-transcriptional modifications, as well as the background of the mass spectrometry techniques that have been used in characterizing and sequencing of RNA. Chapter 2 presents the development of the Comparative Analysis of RNA Digests (CARD) approach using isotope labeling and mass spectrometry. Next, the application of this CARD approach for sequencing of total transfer RNA from *Citrobacter koseri* is presented in Chapter 3. In Chapter 4, a novel bioinformatics tool for identification of both targeted and untargeted RNA in LC-MS datasets is presented. A method developed for identification of RNA sequence isomers
by isotope labeling and tandem mass spectrometry is demonstrated in Chapter 5. Chapter 6 presents the conclusions of my research and the perspectives of future experiments.

Figure 1.1 Structure of RNA
1.2 RNA and post-transcriptional modification

RNA is mainly composed of nucleosides, which consists of a nucleobase bound to a ribose ring via a β-glycosidic linkage, connected by a phosphodiester backbone (Figure 1.1). In general, adenine (A), cytosine (C), guanine (G), and uracil (U) are the four main nucleobases in RNA. Besides the four common nucleosides, over 100 modified nucleosides have been found so far in various RNA species, such as ribosomal RNA (rRNA), transfer RNA (tRNA), messenger RNA (mRNA), small nuclear RNA (snRNA) [3]. The sites of modification on a nucleoside are diverse and could include the ribose ring, the base, or both sites (Figure 1.2). Some of the modifications are as simple as methylations, such as 2'-O-methylcytidine, whereas others may involve multiple step additions to modify the sugar and base, such as 5-methoxycarbonylmethyl-2'-O-methyluridine.

![Figure 1.2 Representative structures of modified nucleosides](image)

A large number of modified nucleosides are found in tRNAs, which generally possess the greatest concentration of modified nucleosides, typically 8-12%. tRNA is an adaptor...
molecule of RNA serving as a link between genome information and protein synthesis, typically 73 to 94 nucleotides in length. The similar structure and size of tRNAs make the isolation of individual tRNAs very difficult. [4] Some modified nucleosides in tRNA are highly conserved in type and location, such as pseudouridine (Ψ), ribothymidine (T), and dihydrouridine (D). Such modifications have become part of nomenclature associated with tRNA structure. For example, the dihydrouridine stem and loop domain (DSL) in black and the ribothymidine, or TΨC, stem and loop (TSL) in blue of the cloverleaf secondary structure of tRNA are presented in Figure 1.3. Some conserved post-transcriptional modifications appear to be important in tRNA folding. [5]

Figure 1.3 The structure of domains of tRNA. The crystallographic structure of yeast tRNA\textsuperscript{Phe} is on the left [6]. The cloverleaf secondary structure is in the center. The sequence and secondary structure of the anticodon stem and loop domain from human tRNA\textsuperscript{Lys} is on the right. The figure was adapted from reference. [2]
RNA plays significant roles in regulation of numerous genes[7]. Post-transcriptionally modified nucleosides in RNA are a significant investment in genes, enzymes, substrate and energy. tRNA’s modifications play a critical role in gene expression. [8, 9] Because tRNA’s role is decoding the genome information to enable accurate and efficient protein synthesis, anticodon domain modification directly impacts the decoding process. [2] The modifications located within or adjacent to the anticodon of tRNA play important roles in the accuracy of codon binding, maintenance of the translational reading frame, and translocation of the tRNA from the A-site to the P-site of the ribosome, [9] Several mitochondrial diseases are well-characterized and are caused by a point mutation of mitochondrial tRNA gene. [10] For example, MELAS (mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes) is caused by the mt tRNA$^{\text{Leu(UUR)}}$ gene. These point mutations result in the deficiency of a modified nucleosides 5-taurinomethyluridine ($\tau$m$^5$U), which is the critical wobble modification for accurate translation.

The function of tRNA is performed by recognition between anticodons of tRNA and codons of messenger RNA. The redundancy of the genetic code results in 61 different anticodons available for tRNA decoding of mRNA, in addition to the three anticodons that recognize stop codons on the message. This redundancy leads to isoaccepting tRNAs, which are two or more tRNAs that carry the same amino acid but which have different tRNA sequences. Moreover, within an organism’s genome, one or more copies may exist for a particular tRNA gene. For example, *Escherichia coli* K12 MG1655 has 88 tRNA genes to express 53 different tRNAs which can decode 41 different codons.[11]

tRNA characterization is incomplete without an accurate characterization of the identities, sequence locations and amounts of modified nucleosides. At present, there are 45 different
modified nucleosides known in archaeal tRNAs, 46 different modified nucleosides known in bacterial tRNAs, and 50 different modified nucleosides known in eukaryotic tRNAs.[1] Clearly, the diversity of tRNAs, their modification status, and relative abundance in the cell make tRNA characterization a significant analytical challenge.

Despite the variety of approaches that have been developed to characterize tRNAs, the challenges associated with complete sequencing of tRNAs means that very few organisms (E. coli, Saccharomyces cerevisiae, Halobacterium volcanii, Bacillus subtilis and Mycoplasma capricolum) have been characterized extensively (but not always completely or accurately) at the tRNA level.[11, 12]

1.3 Methods for characterization of RNA

Several methods based on hybridization, chromatography, electrophoresis, and mass spectrometry have been developed to characterize tRNAs. DNA-RNA Hybridization forms double-stranded RNA and DNA that has a complementary nucleotide sequence. When DNA and RNA is heated to denaturation temperatures to form single strands and then cooled RNA can hybridize with DNA having a complementary sequence. Filter and solution hybridization assays were used for the characterization of yeast cytosolic tRNA\textsuperscript{Phe} and bovine mitochondrial tRNA\textsuperscript{Ser}.[13-15] This method for purification of tRNA is not quite efficient, requiring much more samples. For example, Yokogawa et al. purified 2-4 A260 units each of tRNA isoacceptors from 17.5 kg of bovine liver. [15]

Dong et al. used two-dimensional polyacrylamide gel electrophoresis to fractionate tRNA from E. coli, and isolated components were identified by hybridization to tRNA-specific
The design of DNA oligonucleotides sequence for the characterization of each tRNA isoacceptor is required. A variety of other hybridization-based approaches to tRNA detection have been reported,[17-19] with the most impressive and informative approach being the microarray-based detection of tRNAs developed by Pan and co-workers.[20, 21] The above methods are based on the hybridization between tRNA and the complementary DNA-oligonucleotides probe. The methods enable the sensitive and rapid characterization of tRNAs. For diverse modified nucleosides distributed in tRNAs, these methods are hindered due to the inability to recognize modified nucleosides.

Figure 1.4 Overview of mass spectrometric methods for characterization of RNA

1.3.1 ESI-MS and MALDI-MS of ribonucleic acids

Before the introduction of matrix assisted laser desorption/ ionization (MALDI) and electrospray ionization (ESI), it was difficult to characterize RNA oligomers using mass spectrometry. Compared to previous ionization sources, the common feature of these two
ionization methods is “soft” ionization, which can protect the majority of ionized precursors from dissociation after ionization. Nucleic acids can be characterized by ESI-MS[22] or MALDI-MS. [23, 24]The DNA or RNA oligomers can be ionized and form multiply charged ions via ESI or single charged ions via MALDI. It is difficult to couple separation techniques, such as High Performance Liquid Chromatography (HPLC), with the MALDI ionization source. Because of the easy coupling with HPLC, ESI is more popular for characterization of RNA. More molecules can be analyzed after HPLC separation. Three major analytical methods via mass spectrometry can be used for characterization of RNA in general (Figure 1.4), and can be classified into two types of methods, direct characterization and indirect characterization.

1.3.2 Direct characterization of intact RNA by mass spectrometry

Synthetic or natural purified individual RNA can be characterized by molecular mass measurement or top-down mass spectrometric approaches. [25-27] For instance, Limbach et al. used electrospray ionization quadrupole mass spectrometry for accurate determination of molecular masses of polynucleotides and small nucleic acids. [28] In the top-down approach, highly modified tRNA$^{\text{Val}}$ (76 nt) was characterized by collision-induced dissociation (CID) with 89% sequence coverage. [27] Although the individual RNA can be easily and rapidly analyzed by molecular mass measurement or top-down mass spectrometric approaches, a certain amount of individual RNA needs to be purified from a large amount of RNA mixtures, requiring multiple complex purification steps and more sample preparation. Typically, each organism has 25-50 tRNA isoacceptors, thus, the purification of large number of individual tRNA from total tRNA pools is still a challenge. Using the above two approaches, the high throughput analysis of tRNA is difficult.
1.3.3 Indirect characterization of intact RNA by mass spectrometry

Mass spectrometry-based characterization of RNAs is particularly useful for identifying modified nucleosides and placing those modifications into a specific sequence context for the RNA under analysis. The most common indirect mass spectrometry method is known as RNA mass mapping or RNase mapping.[29] RNA mass mapping is based on the use of specific RNases that generate smaller oligonucleotide digestion products amenable to separation and analysis, including sequencing via tandem mass spectrometry (MS/MS), by MALDI-MS [23, 24] or liquid chromatography mass spectrometry (LC/MS).[30, 31] Although RNA mass mapping is a powerful analytical method for RNA characterization in general, until recently the approach was hindered by the lack of suitable software for characterizing the resulting mass spectral data.[32-35]

In general, as Figure 1.5 shows, the strategy for the complete modification map of RNA includes two parts. Before LC-MS analysis, purification of individual RNAs is required. [36] In one part, the modified nucleosides profile can be obtained by enzymatic hydrolysis of RNA to nucleosides. In another part, the individual RNA needs to be digested to oligonucleotides using nucleotide-specific enzymes (e.g. RNase T1). The analysis of nucleosides or oligonucleotides can be performed by LC-MS/MS. The detailed structure information of nucleoside or oligonucleotides can be confirmed via tandem mass spectrometry. Currently, the most common method for purification of individual tRNAs is to use DNA oligonucleotides based on hybridization. [37] This method is quite sensitive, but labor-intensive and time-consuming.
Before nucleoside analysis by LC-MS/MS, enzymatic hydrolysis of tRNA to nucleosides is carried out using nuclease P1, alkaline phosphatase and venom phosphodiesterase, typically on a scale of 0.5 to 2 $A_{260}$ units (15 to 40 μg) of total tRNA. [39] Before oligonucleotide analysis by LC-MS/MS, sequence-specific digestion of tRNA to oligonucleotides is carried out by a ribonuclease (RNase T1), typically on a scale of 2 to 5 μg of total tRNAs. Compared to nucleosides analysis, less sample preparation of oligonucleotides analysis is required.
1.3.4  RNA fragmentation in tandem mass spectrometry

To further characterize the particular chemical nature of RNA, the structure information of RNA can be identified or confirmed by MS/MS. The complexity of MS/MS spectra of RNA makes them difficult to interpret. McLuckey [40] suggested the nomenclature for RNA fragmentation in MS/MS, which is shown in Figure 1.6. Differing from the dissociation of DNA, dissociation of RNA is independent of nucleobase loss and the cleavage of 5’ P-O bonding is predominant. [41] The dominant fragments generated from RNA in CID are c- and y-type product ions, irrespective of the type of ionization source [42] [43]. The complementary c- and y-type ions from 5’ P-O bond cleavage are dominant at low charge state and/or low excitation energies. [43] Higher charge state or higher excitation energies in CID increases fragmentation in MS/MS, resulting in more complex mass spectra. Thus, the MS/MS spectral data from the precursor at low charge state and/or low excitation energies can make mass spectral interpretation of oligonucleotides easier. Because there is a lack of software to de novo or automatically analyze MS/MS data of oligonucleotides, high throughput oligonucleotides analysis of MS/MS data is still a challenge.

Figure 1.6 Nomenclature for RNA fragmentation when performing MS/MS
1.3.5 RNA sequence isomer

The identification of isomeric digestion products by LC-MS is a challenge due to interference of the identical m/z value of other RNA sequence isomers. Typically, multiple methods based on chromatography coupled with mass spectrometry have been developed for the identification of RNA structural isomers, such as HPLC [44], Nano-HPLC[45], and ultra performance liquid chromatography (UPLC) [46]. Based on their characteristic dissociation pattern, tandem mass spectrometry has been applied for the discrimination of DNA sequence isomers [47] or qualitative and quantitative analysis of RNA sequence isomers generated from isomeric post-transcriptional modifications, such as pseudouridine vs. uridine [48, 49].

1.3.6 Comparative analysis of nucleic acids by mass spectrometry

The basic RNA mass mapping approach has been modified and adapted to enable comparative qualitative and quantitative analyses of RNA modifications.[37, 49, 50] For example, Songe-Møller et al. used RNA mass mapping of tRNAs from wild-type and enzyme-deficient mutants to identify differences in RNase T1 digestion product m/z values that revealed site-specific information on modified nucleosides.[37]. Their work for RNA modification mapping basically followed the flowchart of Figure 1.5. For example, to investigate tRNA^{Sec} modifications, individual tRNA^{Sec} isolated from total tRNAs pool is required. Two types of information - modified nucleoside profiling and sequence-specific enzymatic digestion products - were obtained by mass spectrometry (Figure 1.7). They directly analyzed modified nucleosides of purified tRNA^{Sec} from wild type and mutant mice by LC-MS (Fig. 1.7A) and RNase T1 digestion products of purified tRNA^{Sec} from wild type
and mutant mice by MALDI-TOF MS (Fig. 1.7B). In their work, each individual tRNA from the total RNA pool was isolated by specific DNA oligonucleotides beads using hybridization. Because the purification of individual tRNAs is very labor intensive, it means that more individual tRNA needs to be investigated, requiring more sample preparation.

Breuker and co-workers showed that differences in oligonucleotide fragmentation can be used to quantify pseudouridine in mixtures that contain sub-stoichiometric amounts of pseudouridine.[50] The majority of approaches for DNA analysis by mass spectrometry also involve comparative analyses rather than de novo sequencing.[51-53] A common experimental approach in genotyping is to identify the important sequence information within the sample by comparison against another component of the sample. For example, single-nucleotide polymorphism (SNP) can be identified using single-base extension assays with mass spectrometric detection, where the SNP is identified by characterizing the mass difference between the single base added to each allele rather than completely sequencing each allele.[53] In a different example of comparative analysis, van den Boom and co-workers used ribonuclease T1 (RNase T1) to generate digestion fingerprints that could be used to characterize a large number of PCR products for the presence of SNPs.[54]

Overall, the analytical strength of comparative analysis approaches is that only differences within the mass spectral data need to be analyzed, facilitating mass spectral interpretation. The above comparative analysis approaches were developed with MALDI-TOF MS. The advantage of MALDI-TOF mass spectral data interpretation is easy, because the data can easily be aligned manually. Comparative analysis has a potential application in LC-MS. However, LC-MS data is more complex than MALDI mass spectral data, because of the three-dimensional property of LC-MS data. Moreover, due to the lack of the software for
accurate alignment of oligonucleotides in LC-MS datasets, the comparative analysis by LC-MS datasets will be challenging.

Figure 1.7 A. Comparative analysis of modified nucleosides of tRNA\textsuperscript{Sec} by LC-MS; B. Comparative analysis of sequence-specific digestion products of tRNA\textsuperscript{Sec} by MALDI-MS (Figure from reference [37])
Figure 1.8 Flowchart for MS analysis of RNA. The types of software that are available for the processing of mass spectra of RNAs are shown in underlined bold type. (Figure was adapted from reference [34])

1.4 Bioinformatics tools for data analysis of mass spectral data from nucleic acids

The flowchart for the analysis of mass spectral data generated from RNA is illustrated in Figure 1.8. Due to the limitations of the top-down approach and technical limitations of separation of RNA (such as tRNAs), the sequence-specific cleavage by RNase (such as RNase T1, A, U2) or chemical degradation of (purified) RNA to oligonucleotides that can be separated and characterized must be performed before MALDI-TOF MS or LC-ESI-MS/MS analysis. RNase T1 specifically cleaves single stranded RNA after guanine. RNase
U2 specifically cleaves single stranded RNA after adenine. RNase A selectively cleaves single stranded RNA after uracil and cytosine. In practice, RNase A selectively cleaves RNAs after pyrimidines and has been found to cleave after 5-methylcytidine, 5-methyluridine, dihydrouridine, pseudouridine and 4-thiouridine in tRNAs. [57, 58]

Relevant software has been developed for RNA analysis and is listed in Table 1.1. According to its functions, software can be classified into four categories: (1) peak picking, (2) MS-based mass mapping, (3) MS/MS-based de novo sequencing, and (4) MS/MS-based database searching.

### Table 1.1 Software for identification of nucleic acids based on mass spectrometry

<table>
<thead>
<tr>
<th>Software</th>
<th>Type</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass calculation &amp; peak picking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mongo Oligo Mass Calculator</td>
<td>Web-based</td>
<td><a href="http://mods.rna.albany.edu/masspecc/Mongo-Oligo">http://mods.rna.albany.edu/masspecc/Mongo-Oligo</a></td>
<td>-</td>
</tr>
<tr>
<td>SpiceCmd #</td>
<td>Desktop</td>
<td>-</td>
<td>Mitsui Knowledge Industory. Co. Ltd., Tokyo, Japan</td>
</tr>
<tr>
<td>Mass mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMM</td>
<td>Desktop</td>
<td><a href="http://yass.sdu.dk/RMM">http://yass.sdu.dk/RMM</a></td>
<td>[33]</td>
</tr>
<tr>
<td>MS/MS-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOS</td>
<td>Desktop</td>
<td>-*</td>
<td>[32]</td>
</tr>
<tr>
<td>COMPAS</td>
<td>Desktop</td>
<td>-*</td>
<td>[56]</td>
</tr>
</tbody>
</table>

# Commercial product
* Request copy from the author

1.4.1 Mass calculation and Peak picking

As mass spectrometry only can measure the \(m/z\) value of ions, the identification of RNA relies on its \(m/z\) value. The Mongo Oligo Mass Calculator is a common tool for calculation of mass values of DNA or RNA, or oligonucleotides, or oligonucleotides fragments by computer simulation based on the mechanism of dissociation in CID with the given
sequence(s) as input. Among the precursor ion peaks in mass spectral data, the monoisotopic peak is the only peak containing a single isotopic composition. The mass value of the precursor ion can be determined by the monoisotopic peak and charge state, which can be used for determination of the base composition of oligonucleotides. Therefore, the accurate identification, namely peak picking, of the monoisotopic peak plays a critical role in chemical analysis, quantitative analysis, and identification of RNA. Peak picking can be performed manually or computationally. For purified or synthetic oligonucleotides, peak picking can be easily identified by high resolution mass spectrometry. However, for mass spectral data obtained from biological samples, peak picking is much more difficult because the spectra are not so simple. Currently, there is no automated program available in the public domain specific for oligonucleotide peak picking in mass spectra. A commercial product SpiceCmd is available and specific for oligonucleotides analysis.

1.4.2 MS-based (Mass mapping)

Mass mapping software is designed for identification of the origin of a given RNA by comparing experimental mass spectral data of RNA fragments with the whole genome or RNA FASTA sequence databases. A software, named RRM, has been developed for RNA mass mapping.[33] The identification by RRM is based on the concept that masses of digestion products consist of specific fingerprints, which characterized the given RNA. Another alternative approach was developed to identify *E. coli* tRNAs by their “signature” digestion products (SDPs), whose masses are unique among all the digestion products. [55]

The mass-mapping database RNAccess, which contains a few organisms’ tRNA SDPs, can be used to identify tRNAs (Table 1.1). Due to the diversity of modified nucleosides and the difficulty of localization of post-transcriptional modification in tRNAs, there are a small
number of organisms’ tRNAs available in RNAccess. These two approaches have been proven successful for identification of RNA using specific enzymatic digestion products of RNA.

1.4.3 MS/MS-based de novo sequencing

MS/MS-based sequencing has been used for structure and sequence analysis of oligonucleotides. Generally, the determination of oligonucleotides sequences is interpreted manually. Due to the complexity of the MS/MS spectra of oligonucleotides, this manual interpretation is labor intensive, requiring extensive experience. An algorithm for automated interpretation of oligonucleotide sequence by MS/MS has been developed, [59] and was implemented in Simple Oligonucleotide Sequencer (SOS), a program for MS/MS-based de novo sequencing of DNA and RNA up to 20 nucleotides in length. [32] Another algorithm, COMPAS, was developed for automated sequencing of oligonucleotides by comparing the experimental MS/MS spectra of a given oligonucleotide with a predicted library of reference oligonucleotides fragmentation spectra. [60]

1.4.4 MS/MS-based database searching

Ariadne, a web-based database search engine, was designed and implemented for identification of RNA in biological samples by correlating experimental MS/MS data of RNA with a DNA/RNA sequence database. [35] First, it evaluates the matches between the masses of fragment ions in MS/MS spectrum of an RNase digestion product of RNA sample and that predicted from a candidate sequence in the DNA/RNA database. Second, the candidate sequences are mapped to RNA entries in the database and identified as particular RNAs by scoring the occurrences of the candidate sequences. A small number of
post-transcriptional modifications of RNA can be predicted by Ariadne, such as methylation of nucleotide base or sugar, by estimation of mass value shift from theoretical mass values.

To sum up, the above bioinformatics tools facilitate mass spectral data analysis and interpretation of unmodified or low level modified DNA or RNA or oligonucleotides. Because the post-transcriptional modifications of RNA are diverse, and some of them are hyper-modified, their mass spectral data are much more complex and difficult to interpret. The above bioinformatics tools do not have such powerful capability specific for post-transcriptional modifications in RNA

1.5 Background for dissertation research

![Figure 1.9 The endonuclease cleavage mechanism. The 3'-cyclic phosphate is an intermediate in enzymatic digestion process. After oxygen from water is incorporated into phosphate group, a stable isotope label from oxygen can be available in the 3'-phosphate (II).](image-url)
The goal of this work is to establish a mass spectrometry method for RNA analysis that benefits from the analytical strengths of comparative approaches, where only differences within the mass spectral data need to be analyzed. In contrast to prior RNA mass mapping approaches, the comparative analysis of RNA digest (CARD) is proposed. To simplify alignment of the digestion products in LC-MS data, the ribonuclease incorporation of isotopic labels is introduced.

Figure 1.0 General overview of $^{18}$O labeling and MS approach for quantitation of RNA (Figure from reference [62])
Enzymatic isotope labeling of RNA was developed by Berhane. [61] In the process of ribonuclease digestion, RNA is cleaved to form an intermediate, the cyclic phosphate product. After oxygen from water is incorporated into the phosphate group, a stable isotope label from oxygen (such as $^{18}$O) is available in the 3’ phosphate. (Figure 1.9) Enzymatic isotope labeling was implemented in the characterization of RNA by post-source decay (PSD) analysis. Due to 50:50 $^{16}$O and $^{18}$O labeling in ribonuclease digestion, the PSD product of RNA carrying the 3’-phosphate group will generate a doublet, simplifying the differentiation of c- and y-type ions in complex MS/MS spectra. [61]

Furthermore, enzymatic isotope labeling was applied to the quantification of RNA by MALDI-MS [62] and LC-ESI-MS [63]. Two RNA samples are digested with RNase T1 in $^{16}$O and $^{18}$O labeled water, separately. The two samples are then combined and analyzed by MALDI-MS (Figure 1.0) or LC-ESI-MS. The relative quantitation information of two RNA samples will be revealed from the relative ion abundance of $^{16}$O and $^{18}$O labeled digestion products in the mass spectra.

$$\frac{I_{18O}}{I_{16O}} = \frac{(I_{A+2} - b \cdot I_A)}{I_A}$$ (1.1)

Ion abundance ratios of $^{16}$O and $^{18}$O labeled digestion products can be calculated by equation (1.1), in which $I_A$ represents the monoisotopic peak abundance of the $^{16}$O labeled product, $I_{A+2}$ represents the combination of the monoisotopic peak abundance of $^{18}$O labeled products and $A+2$ isotopic peak abundance of $^{16}$O labeled products, $b$ is the percentage of $A+2$ isotopic peak abundance contributed from $^{16}$O labeled products compared to monoisotopic peak abundance of $^{18}$O labeled products. Because enzymatic isotope labeling was applied to relative quantification of RNA by LC-MS, the quantification information has
a potential application in automated analysis to discover the difference within the mass spectral data.

1.6 Conclusion

RNAs and their post-transcriptional modification play a significant role in cells, so the complete analysis of RNA including modification is required. Due to the powerful software and search engine for the mass spectral data of diverse post-transcriptional modified RNA, the goal of this dissertation is to establish a high throughput automated comparative analysis of RNA, combining ribonuclease incorporation of isotopic labeling with LC-MS analysis.

2.1 INTRODUCTION

The goal of this chapter is to establish a mass spectrometry method for RNA analysis that benefits from the analytical strengths of comparative approaches, such as those used in genotyping, where only differences within the mass spectral data need to be analyzed. In contrast to prior RNA mass mapping approaches, the proposed method for the Comparative Analysis of RNA Digests (CARD) combines ribonuclease incorporation of isotopic labels[62, 63] with mass spectrometry analysis. By incorporating different isotopic labels during digestion with RNase T1, an RNA of unknown sequence or modification status can be compared against an RNA of known sequence or modification status. Where the two RNA sequences are identical, a characteristic isotopic doublet will be detected; differences between the two RNAs are immediately identified by the absence of the isotopic doublet. Unlike RNase mapping or RNA sequencing methods, this approach does not require prior knowledge of unknown RNA sequence or modification status, rather it simply requires a suitable reference RNA for comparison. This work has been published in Analytical Chemistry. (2012), 84 pp.8607-13.

2.2 EXPERIMENTAL

2.2.1 Materials
Citrobacter koseri 14804 was purchased from American Type Culture Collection (ATCC, Manassas, VA). *Escherichia coli* MG1655 and JW4129-2 (miaA-Δ) strains were obtained from the *E. coli* Genetic Stock Center (Yale University, New Haven, CT). Bacto yeast extract, bacto tryptone, nutrient broth and nutrient agar were used as received from BD Bioscience (Sparks, MD). Triethylamine (TEA), 1,1,1,3,3,3-hexafluoroisopropanol (HFIP), magnesium acetate, ammonium chloride, Tri-Reagent, lysozyme chloride from chicken egg white, chloroform, 2-propanol, absolute ethanol and sodium chloride were from Sigma-Aldrich (St. Louis, MO). Ammonium acetate, potassium chloride, magnesium chloride and sodium dodecyl sulfate (SDS) were purchased from Fisher Scientific (Fairlawn, NJ). HPLC-grade methanol and acetonitrile were obtained from Honeywell Burdick & Jackson, Inc. (Muskegon, MI). Molecular biology grade tris-hydrochloride was purchased from Promega (Madison, WI). Sodium citrate was purchased from Mallinckrodt Baker, Inc. (Paris, KY). UltraPure agarose was purchased from Invitrogen Corporation (Carlsbad, CA). RNase T1 was purchased from Roche Molecular Biochemicals (Indianapolis, IN). Sep-Pak C18 cartridges were obtained from Waters (Milford, MA). H$_2$O$_{18}$ (95% purity) was used as purchased from Isotec (Miamisburg, OH). Nanopure water (18 MOhms) from a Barnstead (Dubuque, IA) nanopure system was used as a mobile phase solvent or autoclaved before use in enzymatic digestions. The 5’-biotin labeled oligodeoxynucleotide 5’Biotin-TGGAGGCAGGTCCGAGTGAAGTGAACC -3’ was obtained from Integrated DNA Technologies, Inc. (Coralville, IA).

2.2.2 Isolation of tRNAs

*C. koseri* was cultured with nutrient broth media, and both *E. coli* strains were cultured with Luria-Bertani media.[64] Cells were harvest at the end of mid-log phase (0.8-1.0 OD$_{600}$) and
frozen at -80 °C until further use. Transfer RNAs were isolated using Tri-Reagent as described.[65]

2.2.3 Purification of tRNA\textsubscript{Cys}

Transfer RNA-Cys(CGA) was purified from \textit{E. coli} and \textit{C. koseri} tRNAs using a biotinylated oligodeoxynucleotide with a sequence complementary to the 3’-sequence of \textit{E. coli} tRNA-Cys(CGA).[66] The biotinylated oligodeoxynucleotide was coupled to streptavidin agarose beads by 1 h incubation at room temperature. Purified total tRNA mixtures from the two organisms were used for all hybridization reactions. Transfer RNA-Cys was isolated by hybridization at 70 °C for 10 min followed by stirring for 2 h at room temperature. The bound tRNA-Cys was eluted from the oligonucleotide probe using 15 mM sodium chloride, 1.7 mM sodium citrate and a 0.1% SDS solution. Transfer RNA isolation was confirmed by gel electrophoresis and mass spectrometry.

2.2.4 Ribonuclease Digestions and Isotope Labeling

RNA was quantified using an Implen NanoPhotometer\textsuperscript{TM} Pearl (Implen GmbH, Munich, Germany). For all digestion and labeling experiments, the relative amounts of tRNA from the reference and candidate were equivalent. Experiments using purified tRNA\textsubscript{Cys} required 500 ng of each, while experiments using total tRNAs required 10 \(\mu\)g of total tRNA for both the reference and candidate. RNase T1 was precipitated from its original solution with acetone, resuspended and eluted in 1 ml of 75% aqueous acetonitrile from a Sep-Pak C18 cartridge. For RNase digestion, a ratio of 500 units of RNase T1 to 10 \(\mu\)g of tRNA was used. The reaction mixture was incubated in a 37 °C water bath for 2 h. For samples requiring
isotopic labeling with H$_2^{18}$O, the reaction mixture was first taken to dryness before being reconstituted in H$_2^{18}$O as previously described.[67]

2.2.5 Liquid Chromatography-Mass Spectrometry

All sample solutions were lyophilized and reconstituted in mobile phase A to a concentration of 0.5 μg μL$^{-1}$ and then analyzed from 10 μL (5 μg) injections. All LC/MS or LC/MS-MS analyses were performed using a MicroAS autosampler, Surveyor MS Pump Plus HPLC system and Thermo LTQ-XL (Thermo Scientific, Waltham, MA) mass spectrometer equipped with an electrospray ionization (ESI) source. Reversed phase chromatography was performed on an Xbridge®-MS C18 1.0 × 150 mm column, with 3.5 μm particle size and 50 Å pore size (Waters, Milford, MS) at a flow rate of 40 μL min$^{-1}$. Mobile phase A consisted of 16.3 mM TEA/400 mM HFIP at pH 7.0 in water; equal amounts of mobile phase A and methanol were combined to produce mobile phase B. The mobile phase gradient starting at 5 %B increased to 20 %B at 5 min, 30 %B at 7 min, and to 95 %B at 50 min. The mobile phase composition was held at 95 %B for 5 min prior to re-equilibrating the column for 15 min under initial mobile phase conditions before the next injection.

The ESI source was set to a capillary temperature of 275 °C and spray voltage of 4.50 kV. Data was recorded in negative polarity and profile data type using a zoom scan with the mass range from m/ξ 600 to 2000 in scan event 1, which acquires the m/ξ value of digestion products. In scan event 2, sequence information of the most abundant digestion product was obtained by data-dependent collision-induced dissociation (CID). The normalized collision energy was set to 35%, and the activation time was 30 ms. Data acquisition was through the Thermo Xcalibur software.
2.3 RESULTS AND DISCUSSION

2.3.1 Isotope Labeling for Comparative Analysis

The experimental outline for this comparative analysis approach is presented in Figure 2.1. This approach is enabled by previous advances in using RNase-mediated isotope incorporation to differentially label two RNA populations.[62, 63, 68] Within this approach, a reference RNA whose sequence is known at the post-transcriptional level is digested with an RNase and, during the digestion process, labeled with $^{16}$O at the resulting 3’-phosphate groups. The second, candidate, RNA is digested with the same RNase, except the 3’-phosphate groups are labeled with $^{18}$O. RNA sequences that are identical between the two samples will yield RNase digestion products that appear as doublets, separated by 2 Da ($^{16}$O vs. $^{18}$O) in the mass spectral data. Digestion products containing differences in the underlying RNA sequence or post-transcriptional modification status will yield singlets, which can then be characterized to identify the source of the singlet. To differentiate singlets arising from the reference or candidate RNA, an additional analysis using reverse-labeling can be employed wherein the reference RNA is labeled with $^{18}$O and the candidate RNA is labeled with $^{16}$O. Singlets initially arising from the reference will increase by 2 Da upon reverse labeling, while singlets initially arising from the candidate RNA will decrease by 2 Da upon reverse labeling.

2.3.2 Comparative Sequence of \textit{C. koseri} tRNA

As an initial illustration of this comparative analysis approach for RNA, tRNA-Cys(GCA) was isolated from \textit{E. coli} by hybridization to serve as the reference tRNA, as the complete sequence of this tRNA is known.[11] In a similar manner, tRNA-Cys was isolated from \textit{C. koseri}. \textit{C. koseri} and \textit{E. coli} both belong to the \textit{Enterobacteriaceae} family, yet the tRNA sequences
for *C. koseri*, at the post-transcriptional level, are presently unknown. The purified tRNA-Cys(CGA) from *E. coli* was digested in the presence of H$_2^{16}$O, and the purified tRNA-Cys from *C. koseri* was digested in the presence of H$_2^{18}$O. The RNase T1 digests were combined in equal amounts and analyzed by LC-MS/MS (Figure 2.2). Three singlets were detected. All other labeled digestion products (> monomer) were detected as doublets.

![Figure 2.1 Schematic outline of comparative sequencing by isotope-labeling and LC-MS where *E. coli* serves as the reference organism and *C. koseri* serves as the candidate (unknown) to be sequenced. tRNA endonuclease digestion products that are equivalent between organisms will appear as doublets (separated by 2 Da) in the mass spectral data; digestion products that are different between the two organisms will appear as a singlet.](image)

The first singlet (Fig. 2.2a) was identified as U[s$^{4}$U]AACAAGp from *E. coli* by MS/MS and a comparison with the known *E. coli* tRNA-Cys sequence. As noted above, the identity of this singlet can be confirmed by reverse-labeling and analysis (Fig. 2.2b). As expected if the singlet arises from the reference, after reverse-labeling the *m/z* value increased by 1 u.
(due to detection of the 2- charge state). Based on detecting this singlet, one can predict that the *C. koseri* tRNA-Cys sequence differs from its *E. coli* counterpart, either by a change in the identity of one (or more) bases or through a difference in the post-transcriptional modification status.

The next singlet (**Fig. 2.2c**) was found to be U[^4^]AGp arising from *C. koseri*, which was confirmed by reverse-labeling (**Fig. 2.2d**). Based on a comparison of the data analyzed at this point, tRNA-Cys from *C. koseri* contains G10 in place of A10 in the *E. coli* tRNA. The third and final singlet detected (**Fig. 2.2e,f**) corresponds to the RNase T1 digestion product CAAAGp from *C. koseri*, which is a result consistent with a single G to A sequence change noted above. As illustrated here, the combination of isotope labeling and LC-MS/MS can be used to obtain (or confirm) unknown RNA sequences at the post-transcriptional level in place of complete de novo sequencing of the isolated RNA. Moreover, with this comparative analysis approach, only three RNase T1 digestion products required further interrogation and analysis by MS/MS (**Figures 2.3-2.5**), where de novo approaches depend on complete sequence analysis of each and every RNase digestion product.

It must be noted that, as demonstrated, this approach relies primarily on LC/MS analysis. An analysis based on measured mass values (i.e., LC/MS) would not differentiate when sequence differences lead to sequence isomers in the RNase digestion products. LC-MS/MS can be used for sequence confirmation of detected singlets, however any isomeric differences between the reference and candidate sequences (e.g., pseudouridine vs. uridine or sequence isomers) would generate incorrect doublets that would be challenging to differentiate using the approach as described here. Such limitations should be kept in mind when utilizing this approach.
Figure 2.2 Singlets detected from comparative sequencing of tRNA$^{\text{Cy}}$ from C. koseri using tRNA$^{\text{Cy}}$ of E. coli as the reference sequence. (a) Singlet $\text{U}[^{34}\text{U}]\text{AACAAAGp}$ (m/z 1469.6, 2- charge) arising from purified E. coli tRNA$^{\text{Cy}}$ after labeling C. koseri with $^{18}$O. (b) The same singlet after labeling E. coli with $^{18}$O. The +1 increase in the m/z isotopic envelope confirms the singlet arises from E. coli. The low abundance peaks at < m/z 1470.6 arise from the use of < 100% $^{18}$O-labeled water. (c) Singlet $\text{U}[^{34}\text{U}]\text{AGp}$ (m/z 660.0, 2- charge) from purified C. koseri tRNA$^{\text{Cy}}$ after labeling C. koseri with $^{18}$O. (d) The same singlet after labeling E. coli with $^{18}$O. The -1 decrease in the m/z isotopic envelope confirms the singlet arises from C. koseri. (e) Singlet $\text{CAAAGp}$ (m/z 827.5, 2- charge) from purified C. koseri tRNA$^{\text{Cy}}$ after labeling C. koseri with $^{18}$O. (f) The same singlet after labeling E. coli with $^{18}$O. The -1 decrease in the m/z isotopic envelope confirms the singlet arises from C. koseri.
Figure 2.3 Collision-induced dissociation mass spectrum of the singlet U[s^{4}U]AACAAAAGp from E. coli.

Table 2.1 Listing of all m/z values used in oligonucleotide sequence assignment.

<table>
<thead>
<tr>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983.21</td>
<td>Y_6^-</td>
<td>1399.21</td>
<td>a_5^-B^-</td>
<td>1099.98</td>
<td>w_5^-</td>
</tr>
<tr>
<td>1919.12</td>
<td>c_8^-</td>
<td>1349.06</td>
<td>y_4^-</td>
<td>1020.02</td>
<td>y_5^-</td>
</tr>
<tr>
<td>1734.15</td>
<td>w_5^-</td>
<td>1316.57</td>
<td>y_2^-</td>
<td>955.95</td>
<td>c_5^-</td>
</tr>
<tr>
<td>1654.09</td>
<td>Y_6^-</td>
<td>1288.04</td>
<td>c_6^-2^-</td>
<td>770.97</td>
<td>w_2^-</td>
</tr>
<tr>
<td>1590.02</td>
<td>c_8^-</td>
<td>1284.96</td>
<td>c_4^-</td>
<td>740.89</td>
<td>a_3^-B^-</td>
</tr>
<tr>
<td>1429.04</td>
<td>W_4^-</td>
<td>1155.5</td>
<td>y_2^-</td>
<td>690.1</td>
<td>y_2^-</td>
</tr>
<tr>
<td>1402.08</td>
<td>[M-Ade]^2-</td>
<td>1123.46</td>
<td>c_7^-2^-</td>
<td>626.91</td>
<td>c_2^-</td>
</tr>
</tbody>
</table>
Figure 2.4 Collision-induced dissociation mass spectrum of the singlet U[s4U]AGp from *C. koseri*.

Table 2.2 Listing of all *m/z* values used in oligonucleotide sequence assignment.

<table>
<thead>
<tr>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1092.97</td>
<td>w3&lt;sup&gt;-&lt;/sup&gt;</td>
<td>690.98</td>
<td>y2&lt;sup&gt;-&lt;/sup&gt;</td>
<td>583.45</td>
<td>[M-Gua]&lt;sup&gt;2-&lt;/sup&gt;</td>
</tr>
<tr>
<td>1012.89</td>
<td>y3&lt;sup&gt;-&lt;/sup&gt;</td>
<td>626.9</td>
<td>c2&lt;sup&gt;-&lt;/sup&gt;</td>
<td>441.93</td>
<td>w&lt;sub&gt;1&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>955.92</td>
<td>c2&lt;sup&gt;-&lt;/sup&gt;</td>
<td>610.08</td>
<td>[M-H&lt;sub&gt;3&lt;/sub&gt;P&lt;sub&gt;4&lt;/sub&gt;O&lt;sub&gt;7&lt;/sub&gt;]&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>418.97</td>
<td>a&lt;sub&gt;2&lt;/sub&gt;B&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>770.93</td>
<td>w2&lt;sup&gt;-&lt;/sup&gt;</td>
<td>595.56</td>
<td>[M-s&lt;sup&gt;4&lt;/sup&gt;Ura]&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>362.00</td>
<td>y&lt;sub&gt;1&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>740.92</td>
<td>a&lt;sub&gt;2&lt;/sub&gt;B&lt;sup&gt;-&lt;/sup&gt;</td>
<td>592.05</td>
<td>[M-Ade]&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>304.95</td>
<td>c&lt;sub&gt;1&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Table 2.3 Listing of all *m/z* values used in oligonucleotide sequence assignment.

<table>
<thead>
<tr>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1349.03</td>
<td>y4</td>
<td>770.99</td>
<td>w2&lt;sup&gt;-&lt;/sup&gt;</td>
<td>633.02</td>
<td>c2&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>1291.00</td>
<td>c4&lt;sup&gt;-&lt;/sup&gt;</td>
<td>759.05</td>
<td>[M-Ade]&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>441.97</td>
<td>w&lt;sub&gt;1&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>1019.98</td>
<td>y3&lt;sup&gt;-&lt;/sup&gt;</td>
<td>714.02</td>
<td>w&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>362.03</td>
<td>y&lt;sub&gt;1&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
</tr>
<tr>
<td>961.95</td>
<td>c&lt;sub&gt;3&lt;/sub&gt;</td>
<td>691.07</td>
<td>y&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.3.3 Comparative Analysis of *E. coli* miaA-Δ mutant tRNAs

Another application of this comparative analysis approach for RNA characterization by mass spectrometry is for the identification of sequence-specific sites of post-transcriptional modification. While HPLC analysis of nucleoside digests is a common approach used in the identification of RNA modifying enzymes,[69] that approach cannot reveal the specific sequence location(s) for RNA modification. As differential modification of RNAs should generate singlets in a manner similar to that shown above where the underlying RNA sequence differs, I next sought to test this concept using a known mutant *E. coli* strain, which lacks the enzyme miaA necessary for the $N^\alpha$-isopentenyl modification to A37 in select *E. coli* tRNAs. Because $N^\alpha$-isopentenyladenosine ($i^\alpha$A) serves as the substrate for miaB catalyzed modification to 2-methylthio-$N^\alpha$-isopentenyladenosine (ms$^{2i\alpha}$A), this mutant should result in the absence of the modified nucleosides $i^\alpha$A and ms$^{2i\alpha}$A from select *E. coli*
tRNAs. The RNase T1 digestion products from tRNAs containing these modifications in the wild type strain (MG1655) as well as those tRNAs missing these modifications in the mutant strain (JW4129-2) should be detected as singlets using this comparative analysis approach.

Figure 2.6 Representative singlets detected from comparative analysis of total tRNAs from E. coli miaA-Δ mutant using total tRNAs of E. coli MG1655 as the reference. (a) Singlet ACU[Q]UΑ[m^2i6A]Α[Ψ]CUGp (m/z 1365.2, 3- charge) arising from E. coli MG1655 after labeling miaA-Δ mutant with ^18O. (b) The same singlet after labeling E. coli MG1655 with ^18O. The +0.66 increase in the m/z isotopic envelope, corresponding to 2 Da at the -3 charge state, confirms the singlet arises from E. coli MG1655. (c) Collision-induced dissociation mass spectrum of the singlet ACU[Q]UΑ[m^2i6A]Α[Ψ]CUGp from E. coli MG1655. The major peak corresponds to a loss of 114 Da from the molecular ion. This loss is consistent with loss of the 2-methylthio-6-isopentenyl group during CID. A listing of all m/z values used for sequence assignment can be found in Table 2.5.
Comparative analysis of tRNAs from the *E. coli* miaA-Δ mutant strain with those from *E. coli* MG1655, used as the reference, was performed using RNase T1 with analysis by LC/MS.

Sequence confirmation of singlet assignments was performed by examining data-dependent MS/MS spectra. A total of seven singlets were identified as arising from *E. coli* MG1655 and
three singlets were identified as arising from the miaA-Δ mutant strain (Table 2.4). A representative example includes the RNase T1 digestion product ACU[Q]UA[ms2BG]A[Ψ]CUGp from *E. coli* MG1655 detected as a singlet (Fig. 2.6a) and confirmed to arise from the reference (MG1655 strain) by reverse labeling (Fig. 2.6b). The corresponding RNase T1 digestion product ACU[Q]UAAA[Ψ]CUGp from the *E. coli* miaA-Δ mutant (JW4129-2) was detected as a singlet (Fig. 2.7a), and also confirmed by reverse labeling (Fig. 2.7b). Representative MS/MS data for confirmation of the assigned sequences are shown in Figures 2.6c and 2.7c. Based on the known reference tRNA sequences from *E. coli*, this specific RNase T1 digestion product arises from tRNA-Tyr (Table 2.4).

<table>
<thead>
<tr>
<th><em>E. coli</em> MG1655 tRNA</th>
<th>RNase T1 Singlet Digestion Product</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sec (UCA)</td>
<td>ACUUCA[15A]AUCCAGp</td>
<td>4214.6</td>
</tr>
<tr>
<td>Trp (CCA)</td>
<td>U[Cm]UCCA[ms2BG]AACCGp</td>
<td>3944.6</td>
</tr>
<tr>
<td>Phe (GAA)</td>
<td>AA[ms2BG]A[Ψ]CCCCGp</td>
<td>3319.5</td>
</tr>
<tr>
<td>Cys (GCA)</td>
<td>CA[ms2BG]A[Ψ]CCCGp</td>
<td>2685.4</td>
</tr>
<tr>
<td>Ser (UGA)</td>
<td>A[ms2BG]AACCGp</td>
<td>2403.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><em>E. coli</em> miaA-Δ Mutant tRNA</th>
<th>RNase T1 Singlet Digestion Product</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyr (QUA)</td>
<td>ACU[Q]UAAA[Ψ]CUGp</td>
<td>3984.6</td>
</tr>
<tr>
<td>Trp (CCA)</td>
<td>U[Cm]UCCAAAACCGp</td>
<td>3830.5</td>
</tr>
<tr>
<td>Phe (GAA)</td>
<td>AAAA[Ψ]CCCCGp</td>
<td>3205.5</td>
</tr>
</tbody>
</table>

These data demonstrate that the comparative analysis approach can be used to identify differences in RNA modification at a sequence specific level. Not only were these
differences identified without requiring complete de novo sequencing of the entire complement of *E. coli* tRNAs, no fractionation or purification of the entire tRNA digest was done prior to comparative analysis. The 10 singlets listed in Table 2.4 could be identified from among the hundreds of doublets present from the total tRNA digests, and these 10 singlets simplify further characterization of modification differences among the two strains. Some of the anticipated singlets from the miaA-Δ mutant strain were not detected (i.e., tRNAs Leu, Sec, Cys, Ser). However, singlets for each of these tRNAs that include either ms2i6A or i6A were detected in the reference sample. One difficulty encountered in this application of the comparative analysis approach is that the absence of modifications can lead to RNase digestion products whose *m/z* values overlap with other, common RNase digestion products from other tRNA components of the mixture. Still, it is possible to suspect that these modifications are lost from those tRNAs in the miaA-Δ mutant strain as the tRNAs from the reference strain for each were detected as singlets. It remains to be examined how complex the initial digestion mixture may be before this approach becomes limiting due to limited dynamic range or overlapping *m/z* values.

Table 2.5 Listing of all *m/z* values used in oligonucleotide sequence assignment of data shown in Figure 2.6c.

<table>
<thead>
<tr>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1883.86</td>
<td>y_{11} ^2^-</td>
<td>1465.26</td>
<td>a_{9-B} ^2^-</td>
<td>1229.74</td>
<td>w_{7} ^2^-</td>
</tr>
<tr>
<td>1867.40</td>
<td>c_{11} ^2^-</td>
<td>1410.28</td>
<td>c_{4} ^-</td>
<td>1189.72</td>
<td>y_{7} ^2^-</td>
</tr>
<tr>
<td>1830.30</td>
<td>a_{9-B} ^-</td>
<td>1408.26</td>
<td>c_{8} ^2^-</td>
<td>1065.20</td>
<td>w_{9} ^2^-</td>
</tr>
<tr>
<td>1731.30</td>
<td>y_{10} ^2^-</td>
<td>1382.74</td>
<td>w_{8} ^2^-</td>
<td>1053.20</td>
<td>a_{4-B} ^-</td>
</tr>
<tr>
<td>1716.32</td>
<td>c_{5} ^-</td>
<td>1342.78</td>
<td>y_{8} ^2^-</td>
<td>1053.16</td>
<td>w_{5} ^-</td>
</tr>
<tr>
<td>1713.70</td>
<td>c_{10} ^-</td>
<td>1328.64</td>
<td>[M-Cyt] ^3^-</td>
<td>1025.20</td>
<td>w_{6} ^-</td>
</tr>
<tr>
<td>1688.06</td>
<td>w_{5} ^-</td>
<td>1328.26</td>
<td>[M-Ura] ^3^-</td>
<td>1022.14</td>
<td>c_{8} ^-</td>
</tr>
<tr>
<td>1618.26</td>
<td>w_{9} ^2^-</td>
<td>1320.30</td>
<td>[M-Ade] ^3^-</td>
<td>973.14</td>
<td>y_{4} ^-</td>
</tr>
<tr>
<td>1618.26</td>
<td>a_{10-B} ^2^-</td>
<td>1315.26</td>
<td>[M-Gua] ^3^-</td>
<td>939.14</td>
<td>c_{3} ^-</td>
</tr>
<tr>
<td>1608.26</td>
<td>y_{5} ^-</td>
<td>1282.22</td>
<td>[M-ms2i6A] ^3^-</td>
<td>748.16</td>
<td>w_{2} ^-</td>
</tr>
<tr>
<td>1578.34</td>
<td>y_{9} ^-</td>
<td>1279.22</td>
<td>y_{4} ^-</td>
<td>668.10</td>
<td>y_{2} ^-</td>
</tr>
<tr>
<td>1561.30</td>
<td>c_{9} ^2^-</td>
<td>1243.60</td>
<td>c_{7} ^2^-</td>
<td>633.16</td>
<td>c_{2} ^-</td>
</tr>
<tr>
<td>1524.34</td>
<td>a_{9-B} ^-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.6 Listing of all $m/z$ values used in oligonucleotide sequence assignment of data shown in Figure 2.7c.

<table>
<thead>
<tr>
<th>$m/z$</th>
<th>Fragment</th>
<th>$m/z$</th>
<th>Fragment</th>
<th>$m/z$</th>
<th>Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1937.48</td>
<td>$y_6^-$</td>
<td>1524.30</td>
<td>$a_5^-$</td>
<td>1244.22</td>
<td>$a_8^-$</td>
</tr>
<tr>
<td>1831.28</td>
<td>$a_6^-$</td>
<td>1521.18</td>
<td>$y_9^2^-$</td>
<td>1186.76</td>
<td>$c_7^2^-$</td>
</tr>
<tr>
<td>1826.82</td>
<td>$y_{11}^2^-$</td>
<td>1504.22</td>
<td>$c_9^2^-$</td>
<td>1172.78</td>
<td>$w^2_7$</td>
</tr>
<tr>
<td>1810.38</td>
<td>$c_{11}^2^-$</td>
<td>1410.28</td>
<td>$c_4^2^-$</td>
<td>1132.74</td>
<td>$y_{7}^2^-$</td>
</tr>
<tr>
<td>1716.30</td>
<td>$c_5^-$</td>
<td>1408.16</td>
<td>$a_9^-$</td>
<td>1079.54</td>
<td>$a_{7}^2^-$</td>
</tr>
<tr>
<td>1714.22</td>
<td>$w_{10}^2^-$</td>
<td>1359.12</td>
<td>$w_4^-$</td>
<td>1053.22</td>
<td>$w_{5}^2$</td>
</tr>
<tr>
<td>1713.78</td>
<td>$a_{11}^-$</td>
<td>1351.28</td>
<td>$c_6^2^-$</td>
<td>1022.24</td>
<td>$c_{6}^2^-$</td>
</tr>
<tr>
<td>1688.18</td>
<td>$w_5^-$</td>
<td>1289.52</td>
<td>$[M-Cyt]^3^-$</td>
<td>1008.12</td>
<td>$w_{6}^2^-$</td>
</tr>
<tr>
<td>1674.22</td>
<td>$y_{10}^2^-$</td>
<td>1289.52</td>
<td>$[M-Ura]^3^-$</td>
<td>973.16</td>
<td>$y_3^-$</td>
</tr>
<tr>
<td>1656.84</td>
<td>$c_{10}^2^-$</td>
<td>1285.56</td>
<td>$y_{9}^2^-$</td>
<td>939.12</td>
<td>$c_3^2$</td>
</tr>
<tr>
<td>1608.24</td>
<td>$y_5^-$</td>
<td>1282.52</td>
<td>$[M-Ade]^3^-$</td>
<td>748.08</td>
<td>$w_2^-$</td>
</tr>
<tr>
<td>1561.20</td>
<td>$a_{10}^-$</td>
<td>1279.14</td>
<td>$y_4^-$</td>
<td>668.12</td>
<td>$y_{2}^-$</td>
</tr>
<tr>
<td>1561.20</td>
<td>$w_9^-$</td>
<td>1277.20</td>
<td>$[M-Gua]^3^-$</td>
<td>633.14</td>
<td>$c_2^2$</td>
</tr>
</tbody>
</table>

2.4 CONCLUSIONS

An approach based on the comparative analysis of RNAs, designated the CARD (Comparative Analysis of RNA Digests) approach, has been developed for efficiently analyzing one or more RNA sequences without resorting to complete RNA sequencing and which is suitable for the characterization of unknown RNAs. The CARD approach provides a relatively straightforward method for obtaining or confirming RNA sequence and/or modification information at a far higher scale than is currently possible, and the approach is enabled by encoding information relating to differences between the reference and unknown RNA(s) through isotopic labeling. Although RNase-mediated isotopic labeling was used here, the approach should be directly compatible with other stable-isotope labeling approaches used in RNA mass spectrometry.[70] Moreover, because this approach is directly compatible with previously reported quantitation strategies for RNA analysis.[62, 63]
future iterations could provide qualitative and quantitative characterization of unknown RNAs within a single analysis. Although tRNA sequencing was illustrated using a single isolated tRNA in this work, the comparative analysis of the *E. coli* wild type and miaA-Δ mutant strains for modification analysis using unfractionated tRNAs suggests CARD can readily be extended to tRNA sequencing at the organism level for bacterial, archaeal or eukaryal tRNAs, providing suitable reference sequences are available. Application to extend this approach into this area is presented in the next chapter.
Chapter 3. Mass Spectrometry Sequencing of Transfer Ribonucleic Acids by the Comparative Analysis of RNA Digests (CARD) Approach

3.1 INTRODUCTION

In Chapter 2, the Comparative Analysis of RNA Digests (CARD) approach[71] was developed. The CARD approach was illustrated by identifying post-transcriptional modifications site-specifically within \( E. coli \) tRNAs, and a proof-of-concept example of sequencing a single isoaccepting tRNA was presented. In this chapter, I expand the utility of CARD to encompass a more global approach for tRNA sequencing. \( E. coli \) tRNAs were chosen as the set of reference (known) sequences that enabled the comparative sequencing of tRNAs whose complete sequences, including post-transcriptional modifications, are currently unknown. Differences between the reference tRNA sequence and that of the unknown are identified by the presence of singlets in the mass spectrum, whereas identical sequences between the reference and unknown appear as doublets due to the use of \(^{16}\text{O}\) and \(^{18}\text{O}\) isotope labeling. With this method, I demonstrate total tRNA sequencing for an organism sharing high homology to the reference organism without requiring purification of individual tRNAs and without need for de novo approaches. The advantages of the CARD approach are also illustrated by the characterization of an incorrect tRNA sequence in the Sprinzl tRNA database[11] for one of the \( E. coli \) tRNAs previously characterized over 20 years ago.[72] This work has been published in *Analyst*, (2013) **138**, pp 1386-1394.
3.2 EXPERIMENTAL

3.2.1 Materials

*E. coli* MRE 600 and *Citrobacter koseri* 14804 were purchased from American Type Culture Collection (ATCC, Manassas, VA). All other materials used are in Chapter 2.

3.2.2 Isolation of tRNAs

Bacterial cultures and isolation of tRNAs were performed as described in Chapter 2.

3.2.3 Ribonuclease Digestions and Isotope Labeling

The procedure for enzymatic digestion and isotope labeling are described in Chapter 2.

3.2.4 Liquid Chromatography-Mass Spectrometry

LC-MS methods, the mobile phase used and data analysis steps are described in Chapter 2.

3.2.5 Sequence Analysis

All tRNA gene sequences (tDNA sequences) were obtained from the Genomic tRNA Database (http://gtrnadb.ucsc.edu/)[73] or from BLAST searches against the *C. koseri* genome (NC_009792.1).[74] *E. coli* tRNA sequences with annotated post-transcriptional modifications were obtained from the tRNAdb 2009 database (http://trnadb.bioinf.unileipzig.de/)[11] or Modomics (http://modomics.genesilico.pl/).[12] Theoretical sequences of RNase digestion products were obtained using Mongo Oligo Calculator (http://rnamdb.cas.albany.edu/RNAmods/masspec/mongo.htm) or in-house software.
3.3 RESULTS AND DISCUSSION

The experimental outline for the CARD approach is presented in Figure 3.1. This approach is enabled by previous advances in using RNase-mediated isotope incorporation to differentially label two tRNA populations.[62, 63, 68] Details surrounding the development of this approach are described in Chapter 2.

![Figure 3.1](image)

Figure 3.1. Schematic outline of comparative sequencing by isotope-labeling and LC-MS where _E. coli_ serves as the reference organism and _C. koseri_ serves as the candidate (unknown) to be sequenced. tRNA endonuclease digestion products that are equivalent between organisms will appear as doublets (separated by 2 Da) in the mass spectral data; digestion products that are different between the two organisms will appear as a singlet.

3.3.1 Comparative Sequencing of _C. koseri_ total tRNAs

To illustrate this comparative sequencing approach at the level of total tRNAs, _C. koseri_ was chosen as a suitable candidate organism for several reasons. _C. koseri_ and _E. coli_ both belong to the _Enterobacteriaceae_ family. A BLAST search of 14 randomly chosen _E. coli_ tRNA genes yielded hits to 14 tRNA genes from _C. koseri_, and of those genes identified in the BLAST
search, a significant majority (11/14) shared 100% sequence identity with their \textit{E. coli} counterparts (Table 3.1). Thus, \textit{C. koseri} was presumed to be a suitable model to test the approach as only a few singlets would be predicted based on the BLAST results from a subset of tRNA genes. Moreover, because \textit{C. koseri} can be safely cultured under standard laboratory conditions, it provided a rather straightforward experimental system for us to use.

Table 3.1 Basic Local Alignment Search Tool (BLAST) results for searching 14 randomly chosen \textit{E. coli} tRNA genes against the \textit{C. koseri} genome. The value reported is the %sequence homology between the \textit{E. coli} gene and that found in \textit{C. koseri}.

<table>
<thead>
<tr>
<th>\textit{E. coli} query sequence:</th>
<th>tRNA ID</th>
<th>\textit{E. coli}</th>
<th>\textit{C. koseri}</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA: Alanine 1 (Ala1) [A]</td>
<td>DA1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Cysteine (Cys) [C]</td>
<td>DC1660</td>
<td>100%</td>
<td>98%</td>
</tr>
<tr>
<td>tRNA: Glutamic Acid 1 (Glu1) [E]</td>
<td>DE1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Phenylalanine (Phe) [F]</td>
<td>DF1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Glycine 1 (Gly1) [G]</td>
<td>DG1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Histidine (His) [H]</td>
<td>DH1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Isoleucine 1 (Ile1) [I]</td>
<td>DI1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Leucine 1 (Leu1) [L]</td>
<td>DL1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Methionine (Met) [M]</td>
<td>DM1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Glutamine 1 (Gln1) [Q]</td>
<td>DQ1660</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Arginine 1 (Arg1) [R]</td>
<td>DR1660</td>
<td>100%</td>
<td>98%</td>
</tr>
<tr>
<td>tRNA: Serine 1 (Ser1) [S]</td>
<td>DS1664</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>tRNA: Threonine 1 (Thr1) [T]</td>
<td>DT1660</td>
<td>100%</td>
<td>97%</td>
</tr>
<tr>
<td>tRNA: Valine 1 (Val1) [V]</td>
<td>DV1662</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

3.3.2 Creating Anticipated tRNA Sequences

Before undertaking the experimental procedure outlined in Fig. 3.1, I first calculated anticipated singlets and doublets for \textit{C. koseri} comparative analysis. The process used here for calculating anticipated singlets and doublets for all tRNAs is illustrated in Figure 3.2 for a single tRNA. First, a reference tRNA sequence is obtained, and the gene sequence of the candidate tRNA with highest homology to the reference is aligned to the reference tRNA.
Next, the assumption was made that post-transcriptional modifications will be conserved between closely related organisms, therefore any post-transcriptional modifications present on the sequence of the *E. coli* (reference) tRNA were added at the same sequence location of the *C. koseri* (candidate) tRNA (Fig. 3.2b). If the site of post-transcriptional modification differs in nucleotide identity between the reference and candidate sequences, then such post-transcriptional modifications are not added to the candidate sequence. One may confirm that the candidate organism contains the same post-transcriptional modifications as the reference through LC-MS analysis of nucleoside digests.[69] The modified nucleosides from *C. koseri* total tRNAs were shown in Figure 3.3 and these detected modified nucleosides are identical to *E. coli*.

**Figure 3.2.** Procedure for generating anticipated tRNA sequences from the candidate organism. (a) The reference (RNA) sequence and the candidate gene (DNA) sequence are aligned. (b) Where base identity is conserved between the two sequences, known post-transcriptional modifications present on the reference tRNA are placed onto the candidate tRNA sequence to generate the final, predicted candidate tRNA sequence. These sequences can then be digested in silico with the RNase of choice to predict the minimum number of singlets that should appear in the mass spectrum.
Table 3.2. Predicted singlets arising during comparative sequencing of *E. coli* and *C. koseri* total tRNAs when tRNAs digested using the denoted RNases. Sequence difference identified by bold, underlined text. RNase T1 interference-free singlets are denoted by an asterisk, *.

<table>
<thead>
<tr>
<th>tRNA</th>
<th>RNase/Expected Singlet Digestion Product</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> Leu (CAG)</td>
<td>None</td>
<td>2872.3</td>
</tr>
<tr>
<td><em>C. koseri</em> Leu (CAG)</td>
<td>U2/ GUGUUCUAp</td>
<td>2872.3</td>
</tr>
<tr>
<td><em>E. coli</em> Asp (GUC)</td>
<td>U2/ G[D]CGG[D][D]Ap</td>
<td>2611.4</td>
</tr>
<tr>
<td><em>C. koseri</em> Asp (GUC)</td>
<td>T1/ [D]UGp</td>
<td>977.1</td>
</tr>
<tr>
<td></td>
<td>U2/ G[D]UGG[D][D]Ap</td>
<td>2612.4</td>
</tr>
<tr>
<td><em>C. koseri</em> Leu (CAA)</td>
<td>Cannot be identified by RNase T, A or U2</td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em> Cys (GCA)</td>
<td>T1/ U*[sU]<em>AACAAAGp</em></td>
<td>2941.4</td>
</tr>
<tr>
<td><em>C. koseri</em> Cys (GCA)</td>
<td>T1/ CAAAGp</td>
<td>1655.3</td>
</tr>
<tr>
<td><em>E. coli</em> Leu (UAA)</td>
<td>U2/ [T][Y]CCCUCGGCGUUCGCGUUGGGGGG[mU][*p]UCAp</td>
<td>9310.1</td>
</tr>
<tr>
<td><em>C. koseri</em> Leu (UAA)</td>
<td>T1/ UCCCAUCCGGp</td>
<td>1655.3</td>
</tr>
<tr>
<td></td>
<td>U2/ [T][Y]CCCUCGGCGUUCGCGUUGGGGGG[mU][*p]UCAp</td>
<td>9311.1</td>
</tr>
<tr>
<td><em>E. coli</em> Arg (UCU)</td>
<td>T1/ AC[s<em>C]U[mnm</em>U]CU[‘<em>A’]AGp</em></td>
<td>3387.5</td>
</tr>
<tr>
<td></td>
<td>A/ [‘*A’]AG[p]</td>
<td>1472.2</td>
</tr>
<tr>
<td></td>
<td>U2/ C[s<em>C]U[mnm</em>U]CU[‘*A’]Ap</td>
<td>2713.4</td>
</tr>
<tr>
<td><em>C. koseri</em> Arg (UCU)</td>
<td>T1/ C[s<em>C]U[mnm</em>U]CU[‘<em>A’]AGp</em></td>
<td>3068.4</td>
</tr>
<tr>
<td></td>
<td>A/ [‘*A’]AG[p]</td>
<td>1471.2</td>
</tr>
<tr>
<td><em>E. coli</em> Tyr (GUA)</td>
<td>None</td>
<td>2312.3</td>
</tr>
<tr>
<td><em>C. koseri</em> Tyr (GUA)</td>
<td>U2/ GU[Gm]GCCAp</td>
<td>2312.3</td>
</tr>
<tr>
<td><em>E. coli</em> fMet (CAU)</td>
<td>A/ GAAGAUp</td>
<td>2001.3</td>
</tr>
<tr>
<td></td>
<td>U2/ UCGUCG[mU][‘*Y’]CAp</td>
<td>3535.4</td>
</tr>
<tr>
<td><em>C. koseri</em> fMet (CAU)</td>
<td>T1/ [mU][‘*Y’]CAAUCUGp</td>
<td>3198.4</td>
</tr>
<tr>
<td></td>
<td>U2/ UCUGCCCCCGGCAp</td>
<td>4129.5</td>
</tr>
<tr>
<td><em>C. koseri</em> Arg (CCG)</td>
<td>Cannot be identified by RNase T, A or U2</td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em> Val (GAC)</td>
<td>U2/ pGCGUCCG[s*U]Ap</td>
<td>3005.3</td>
</tr>
<tr>
<td><em>C. koseri</em> Val (GAC)</td>
<td>U2/ pGCGUCCG[s*U]Ap</td>
<td>3006.3</td>
</tr>
<tr>
<td><em>E. coli</em> Arg (UCU)</td>
<td>T1/ CCC[s<em>C]U[mnm</em>U]CU[‘*A’]AGp</td>
<td>3668.5</td>
</tr>
<tr>
<td><em>C. koseri</em> Arg (CCU)</td>
<td>T1/ CCC[s*C]UCCU[‘*A’]AGp</td>
<td>3624.5</td>
</tr>
<tr>
<td></td>
<td>U2/ GCCC[s*C]UCCU[‘*A’]Ap</td>
<td>3624.5</td>
</tr>
<tr>
<td><em>E. coli</em> Sec (UCA)</td>
<td>U2/ GUUGGGGCGCGGCGGUGCCCGGCAp</td>
<td>7846.0</td>
</tr>
<tr>
<td><em>C. koseri</em> Sec (UCA)</td>
<td>U2/ GCGGUCGCCCCGGCAp</td>
<td>4248.6</td>
</tr>
<tr>
<td><em>E. coli</em> Thr (GGU)</td>
<td>T1/ [m*G]UCCCAGp</td>
<td>2272.3</td>
</tr>
<tr>
<td></td>
<td>U2/ G[m*G]UCCCAp</td>
<td>2272.3</td>
</tr>
<tr>
<td><em>C. koseri</em> Thr (GGU)</td>
<td>T1/ [m<em>G]UCCCCAp</em></td>
<td>2577.4</td>
</tr>
<tr>
<td></td>
<td>U2/ G[m*G]UCCCCAp</td>
<td>2577.4</td>
</tr>
</tbody>
</table>
Figure 3.3 Modified nucleosides in *Citrobacter koseri* total tRNAs. 1, dihydrouridine (D). 2, pseudouridine (Ψ). 3, 5-carboxymethoxyuridine (cmo^5^U). 4, 3-(3-amino-3-carboxypropyl)uridine (acp^3^U). 5, 2-thiocytidine (s^2^C). 6, 5-methyluridine (m^5^U). 7, (unknown). 8, 2′-O-methylguanosine (Gm). 9, 1-methylguanosine (m^1^G). 10, N^6^-threonylcarbamoyl adenosine (t^6^A). 11, 2-methyladenosine (m^2^A). 12, N^6^-methyladenosine (m^6^A). *, 2-methylthio-N^6^-isopentenyladenosine (ms^2^i^6^A)(data not shown in UV chromatogram).

Following the procedure outlined in Figure 3.2, the 53 known, sequenced tRNAs from *E. coli* that are present in the tRNA[11, 12] and Modomics[11, 12] databases were compared with the 82 tRNA genes identified in *C. koseri* using tRNAScan-SE.[73] This process yields a total of 47 predicted *C. koseri* tRNAs, with 34 of those tRNAs predicted to have identical sequences – assuming post-transcriptional modifications are conserved – to their *E. coli* counterparts (Table 3.3). An advantage of the comparative sequencing approach is that only 13 *C. koseri* tRNAs are predicted to differ from their *E. coli* counterparts; it is this smaller group of tRNAs whose sequence differences from the reference need to be identified by comparative sequence analysis. The majority of the sequence differences among the 13 tRNAs arise from C to U (cytidine to uridine or vice versa) changes between...
*E. coli* and *C. koseri*. Because the nominal mass difference between cytidine and uridine is only 1 Da, such changes cannot easily be identified using RNases A, T1 or U2 with this comparative sequencing approach. However, other changes, such as G to A (guanosine to adenosine or vice versa), insertions/deletions, or changes in detected post-transcriptional modifications should be amenable to identification, if those changes result in predicted RNase digestion products that do not overlap with other, existing digestion products of the same or similar *m/z* values.

### 3.3.3 Predicting Anticipated Mass Spectral Singlets

After constructing all the anticipated *C. koseri* tRNA sequences, these sequences as well as those from *E. coli* were in-silico digested with RNases (*Table 3.2*). Of the 13 predicted *C. koseri* tRNAs that would have sequence differences from their *E. coli* counterparts, 11 of these tRNAs theoretically could be identified via detection of singlets after RNase digestion. When limiting the analysis to RNase T1, at least 11 singlets are anticipated – four arising from the reference organism and seven from the candidate organism. Of those 11 singlets, four arise from C to U changes, five arise from A to G changes and two arise from insertions/deletions. Not included in *Table 3.2* are singlets that would arise from *E. coli* tRNAs that do not appear to have a *C. koseri* counterpart (e.g., tRNA Ala (UGC)).

Similar calculations can be done using other RNases, such as RNase A, which would generate at least three singlets, and RNase U2 (limited to specificity for adenosines), which would generate at least 18 singlets. Because these predictions are based only on differences at the primary sequence level, additional singlets could arise when differences are manifested in the post-transcriptional modifications between the two organisms. Regardless of the
RNase chosen, the number of predicted singlets remains low consistent with the fact that there is high tRNA gene sequence conservation between these two organisms.

**3.3.4 Comparative Sequencing by LC-MS**

Comparative sequencing of tRNAs from *C. koseri* with those from *E. coli* was performed using RNase T1 with analysis by LC-MS. Sequence confirmation of singlet assignments was done by examining data-dependent MS/MS spectra. As noted in Table 3.2, at least 11 singlets arising from six tRNAs are anticipated. Comparative sequencing of total tRNAs from *C. koseri* generates a large number of doublets whose \( m/z \) values may interfere with detection of the anticipated singlets. In particular, any RNase T1 digestion product from any tRNA that is within 2 Da of an expected singlet can lead to mass spectral interferences. Thus, anticipated singlets were classified into two categories – those that are interference-free and those that are interference-prone (Table 3.2). In this initial work, interference-prone singlets were presumed to co-elute with any interfering digestion products during LC/MS analysis. The anticipated singlets in each of these categories were analyzed experimentally.

**3.3.5 Interference-Free Singlets**

The interference-free predicted singlets after RNase T1 digestion, identified by an asterisk, *, in Table 3.2, are *E. coli*-Cys, *E. coli*-Arg (UCU), *C. koseri*-Arg (UCU) and *C. koseri*-Thr (GGU). The anticipated singlet for tRNA-Cys was detected during analysis of total tRNAs (Figure 3.4). The singlet (U[s]\text{*U}\text{AAACAGp}) arising from *E. coli* (Fig. 3.4a) was confirmed by reverse-labeling and analysis (Fig. 3.4b). As expected, if the singlet arises
from the reference, after reverse-labeling the nominal mass will increase by 2 Da (1 \( m/z \) value due to detection of the 2- charge state).

![Mass spectra](image)

Figure 3.4. Mass spectra corresponding to a detected singlet when (a) \( C. \) koseri is labeled with \( ^{18}O \) and (b) \( E. \) coli is labeled with \( ^{18}O \). The singlet \( \text{U}[s^4\text{U}]\text{AACAAAGp} \) (\( m/z \) 1469.6, 2- charge) arises from the \( E. \) coli tRNA-Cys(GCA) as confirmed by the +1 increase in the \( m/z \) isotopic envelope after \( ^{18}O \)-labeling of \( E. \) coli.

The arginine UCU tRNAs of \( E. \) coli and \( C. \) koseri are predicted to differ by an A to G sequence change at position 30. No evidence was found in the mass spectral data for the two anticipated RNase T1 singlets arising from tRNA-Arg(UCU), even when the amount of total tRNA digest was increased from 5 \( \mu \)g to 20 \( \mu \)g (Figure 3.5). As discussed further below,
because tRNA-Arg(UCU) is a low abundance tRNA in E. coli,[16] singlets from this tRNA were below the limits of detection for this method.

Figure 3.5 Extracted ion chromatograms corresponding to the anticipated doubly- and triply-charged ions of AC[\text{s}'\text{\textsc{C}}}U[m\text{m}5\text{U}]CU[t\text{\textsc{A}}]AGp, a singlet from tRNA-Arg(UCU), at 5 μg and 20 μg of E. coli total tRNA after RNase T1 digestion loaded on column. The XIC response at 31.5 min corresponds to a doubly-charged ion consistent with a base composition of C₃U₂A₂Gp + methyl, which is not the anticipated singlet.
Figure 3.6. Mass spectra corresponding to the anticipated singlet of $[\text{m}^7\text{G}]\text{UCCCCAGp}$ from *C. koseri* tRNA-Thr(GGU) when (a) *C. koseri* is labeled with $^{18}\text{O}$ and (b) *E. coli* is labeled with $^{18}\text{O}$. In this case, the doubly charged ion is characterized by a doublet rather than the expected singlet.

The only other interference-free singlet was expected for *C. koseri*-Thr (GGU). Surprisingly, $[\text{m}^7\text{G}]\text{UCCCCAGp}$ was detected as a doublet, rather than the predicted singlet (Figure 3.6a,b). The detection of this predicted singlet as a doublet, in the absence of any expected interferences, is evidence that one of the sequences used for singlet prediction is in error. Further, because a doublet was found for a mass value consistent with $[\text{m}^7\text{G}]\text{UCCCCAGp}$, the error would reside in the *E. coli* sequence present in the tRNAdb.[11] The tRNA sequence for *E. coli* Thr1 present in the tRNAdb yields a difference from the anticipated *C. koseri* Thr1 that corresponds to an internal C in the RNase T1 digestion product of interest.
(Figure 3.7a). Unfortunately, the original publication[72] reporting the sequence for Thr1 does not include any specific sequencing data to determine whether the long run of C-residues would lead to an error between what is experimentally determined here versus that reported in the tRNAdb. However, miscounting of cytidines near the variable loop is possible when using the electrophoresis and thin-layer chromatography method for tRNA sequencing.[75]

a) **tRNA Database Sequence for *E. coli* tRNA-Thr1 (GGU)**

5’−GCU GAU AUG GCU CAG DDG GDA GAG CGC ACC CUU GGU (m^6t^5A)AG GGU GAG (m^7G)UC CCC G(m^5U)Ψ CGA AUC UGG GUA UCA GCA CCA−3’

b) **Aligned gene sequences for tRNA-Thr1 (GGU)**

5’−GCT GAT ATG GCT CAG TTG GTA GAG CGC  E. coli Thr1 gene
5’−GCT GAT ATG GCT CAG TTG GTA GAG CGC  C. koseri Thr1 gene

ACC CTT GGT AAG GGT GAG GTC CCC AGT TCG  E. coli Thr1 gene
ACC CTT GGT AAG GGT GAG GTC CCC AGT TCG  C. koseri Thr1 gene

ACT CTG GGT ATC AGC ACC A−3’  E. coli Thr1 gene
ACT CTG GGT ATC AGC ACC A−3’  C. koseri Thr1 gene


c) **Identified Sequence for C. koseri tRNA-Thr1 (GGU)**

5’−GCU GAU AUG GCU CAG DDG GDA GAG CGC ACC CUU GGU (m^6t^5A)AG GGU GAG (m^7G)UC CCC AG(m^5U)Ψ CGA AAU CUG GGU AUC AGC ACC A−3’

**Corrected Sequence for *E. coli* tRNA-Thr1 (GGU)**

5’−GCU GAU AUG GCU CAG DDG GDA GAG CGC ACC CUU GGU (m^6t^5A)AG GGU GAG (m^7G)UC CCC AG(m^5U)Ψ CGA AAU CUG GGU AUC AGC ACC A−3’

Figure 3.7. (a) Sequence of tRNA-Thr1 (GGU) obtained from the tRNAdb 2009 database (http://trnadb.bioinf.uni-leipzig.de/).[11] (b) Sequence alignment of tRNA-Thr1 (GGU) genes from *E. coli* K12 W3110 strain and *C. koseri*. The two genes are identical. (c) The corrected sequence for *E. coli* tRNA-Thr1 (GGU).
To further corroborate these experimental findings, the gene encoding tRNA-Thr1 from *E. coli* K12 W3110 (gi|85674274:3423435-3423510) was examined and compared to the sequence for Thr1 from *C. koseri*. Indeed, the gene sequence for *E. coli* Thr1 is identical to that for *C. koseri* Thr1 (*Figure 3.7b*). Thus, the error lies within the reference organism tRNA sequence that, when corrected, would not generate any predicted singlets for this tRNA. Based on this finding, the *E. coli* tRNA-Thr1 sequence in the tRNAdb is incorrect, and the corrected sequence is shown in *Figure 3.7c*. Moreover, it should be noted that the Modomics database contains the correct sequence for *E. coli* Thr1,[12] although there is no primary literature that supports that particular sequence entry within Modomics. These results illustrate the importance of having accurate tRNA reference sequences.

### 3.3.6 Interference-Prone Singlets

A total of seven predicted singlets are anticipated to be interference-prone, due to possible mass spectral overlaps with other, known RNase T1 digestion products. Moreover, four of those seven would arise from C to U sequence differences between *E. coli* and *C. koseri*. A typical example of an interference-prone predicted singlet is seen in the analysis of the *C. koseri*-Asp (GUC) predicted sequence [D]UGp (*Figure 3.8*). A number of interfering ions, likely originating from the ubiquitous digestion products CCGp (m/z 972.1), (CU)Gp (m/z 973.1), and UUGp (m/z 974.1), are present. Positive identification of this, and other C to U sequence differences, would require use of a uridine- or cytidine-specific ribonuclease, which is unfortunately not commercially available.

One interference-prone predicted singlet that arises from a G to A sequence change between *E. coli* and *C. koseri* is the RNase T1 digestion product UCCCACUCCGp from *C. koseri*-Leu.
(UAA). This singlet has a mass value is 3134.4. While there are no other \textit{C. koseri} RNase T1 digestion products that would have the same predicted mass value, an RNase T1 digestion product from tRNA-Gly(CCC), AUUCCCUUCGp, has a mass value of 3136.4. The mass difference between these two RNase T1 digestion products is \(\sim 2\) Da, leading to potential interferences during data analysis. \textbf{Figure 3.9a} is the mass spectral results when the candidate tRNAs were labeled with \(^{18}\text{O}\), and \textbf{Figure 3.9b} contains the mass spectral results after reverse-labeling (the reference, \textit{E. coli}, tRNAs are labeled with \(^{18}\text{O}\)).

![Figure 3.8](image)

\textbf{Figure 3.8} Mass spectrum corresponding to the anticipated singlet [DUGp] from \textit{C. koseri} tRNA-Asp(GUC). A number of interfering ions, likely originating from the ubiquitous digestion products CCGp \((m/z\ 972.1)\), (CU)Gp \((m/z\ 973.1)\), and UUGp \((m/z\ 974.1)\), are present and can interfere with detection of the DUGp singlet.
Figure 3.9. Mass spectra corresponding to anticipated singlet from *C. koseri* tRNA-Leu(CAG) when (a) *C. koseri* is labeled with $^{18}$O and (b) *E. coli* is labeled with $^{18}$O. The tRNA-Leu(CAG) singlet is predicted at $m/z$ 1566.2 (2- charge) when *C. koseri* is labeled with $^{16}$O (*E. coli* with $^{18}$O). Interfering doublets arising from tRNA-Gly(CCC) at $m/z$ 1567.2 (2- charge) and 5S rRNA at $m/z$ 1566.6 (2- charge) hinder identification of the tRNA-Leu(CAG) singlet because of isotopic overlap when *C. koseri* is labeled with $^{18}$O.

When analyzing the 2- charge states for these RNase T1 digestion products, the all-light isotope for UCCCACUCCGp ($^{16}$O-labeled) should be found at $m/z$ 1566.1, while the interfering oligonucleotide, AUUCCCUUCGp, should be detected at $m/z$ 1567.1. In addition to the expected ions, another interfering oligonucleotide present at $m/z$ 1566.6 was detected. Such a mass value could easily arise by replacing one uridine with one cytidine in AUUCCCUUCGp to yield a base composition of AC$_3$U$_3$Gp. Although no tRNA sequence is predicted to have that base composition, *E. coli* 5S rRNA does contain the sequence (87)UCUCCCCAUG(96), and thus is the likely source of this interference. While these
results support the expectation that the expected singlet from tRNA-Leu (UAA) is present, it is clear from these two examples that interfering RNase T1 digestion products can hinder accurate sequence identification in certain circumstances.

Another interference-prone predicted singlet that arises from a G to A sequence change between E. coli and C. koseri is the RNase T1 digestion product from C. koseri-Cys, CAAAAGp (Figure 3.10). Although this singlet could not be unequivocally identified due to interferences with other RNase T1 digestion products present in the sample, there is a clear change in the isotopic distribution that is consistent with this RNase T1 digestion product arising from C. koseri. Moreover, it was previously shown that this digestion product is present and detectable as a singlet when tRNA-Cys is purified from bacterial cell cultures.[71]
Thus, in some cases additional confirmation of anticipated singlets that are interference-prone will require improved chromatographic conditions, characterization by additional RNase digestions or application of the CARD method to less complex mixtures.

Of the 82 \textit{C. koseri} tRNA genes examined against \textit{E. coli} tRNAs, a total of 36 tRNA sequences (Table 3.3) were identified based on the RNase T1 data generated here when taking into account both the singlets analyzed as well as the lack of detecting any unanticipated singlets. Because RNase U2 is not commercially available, predicted differences in tRNAs Arg(CCU), Asp(GUC), Tyr(GUA) and Val(UAC) cannot be experimentally verified. Even limiting the experimental studies to a single RNase, the ability to rapidly obtain sequence information at the post-transcriptional level from \textasciitilde80\% of an organism’s tRNAs by this comparative sequence approach provides a uniquely powerful analytical method for tRNA sequencing studies.

3.3.7 Limits of a Comparative Sequencing Approach

As demonstrated above, this comparative sequencing approach for total tRNAs is easily implemented using LC-MS/MS with \textit{E. coli} as a reference organism and a candidate organism that also arises from the \textit{Enterobacteriaceae} family. It seems very likely that other organisms within the \textit{Enterobacteriaceae} family could similarly be sequenced, as BLAST analysis of other bacteria in this family using the same 14 randomly chosen tRNA genes examined earlier (Table 3.1) yields results similar to those obtained for \textit{C. koseri}. Those results were expanded by conducting a theoretical examination of singlets that would be anticipated for the entire complement of tRNAs when \textit{Salmonella enterica} is the candidate organism (Table 3.4). Here, four RNase T1 singlets should be detected if there are no differences in post-transcriptional modifications. Moreover, this initial bioinformatic
approach can be conducted relatively quickly to ascertain the appropriateness of a given reference/candidate set of tRNAs for the CARD approach.

Figure 3.11. (a) Collision-induced dissociation mass spectrum of the singlet \(U[s^4U]AACAAAGp\) from \(E.\ coli\) total tRNAs. (b) Collision-induced dissociation mass spectrum of the doublet \([m^7G]UCCCCAGp\).
<table>
<thead>
<tr>
<th>Table 3.3</th>
<th>tRNA Sequences of Chirobatr Koser</th>
</tr>
</thead>
</table>

The table presents the tRNA sequences for various character codes, including:

- Val (GAC)
- Tyr (GAC)
- Sec
- Leu (GAG)
- Lys
- Arg
- Pro
- Phe
- Met m
- Ser5
- Ser2
- Phe1
- Val2B
- Tyr1
- Gly3
- Arg2
- Met m

The percentages for each character code are also provided.
Table 3.4: tRNA Sequences of Salmonella enterica
At a different extreme, this approach cannot be used when the reference and candidate organisms are phylogenetically quite distinct. For example, using tRNA genes from *Thermus thermophilus* HB8 as the sequencing candidate while retaining *E. coli* as the reference, nearly 40 mass unique expected RNase T1 digestion products would be detected as singlets. Moreover, Carell and co-workers have shown that tRNA modification status varies according to phylogenetic distance among organisms.[76] Thus, a significant fraction of the peaks in the mass spectral data would be singlets arising from both organisms, limiting the utility of this comparative analysis approach.

The approach described here relied primarily on LC-MS analysis. LC-MS/MS was only used when necessary to confirm sequences of detected singlets (Figure 3.11). An analysis based on measured mass values would not differentiate when sequence differences lead to sequence isomers in the RNase digestion products. Further, if two singlets were sequence isomers, they would be challenging to differentiate using the approach as described here unless chromatographic conditions could be optimized to separate sequence isomers.

Because this method for comparative sequencing utilizes differential labeling of samples, other RNA-labeling approaches could also be used. For example, Waghmare and Dickman demonstrated the use of stable-isotope labeling of RNA for improved LC-MS and LC-MS/MS analysis of RNase digests.[70] The particular advantage noted by those authors is that labeling with $^{15}$N-enriched media allows one to more accurately determine oligonucleotide composition, because the number of nitrogens can be used as an additional constraint in oligonucleotide assignments. Another advantage noted in that publication arises when multiple oligonucleotides co-elute. With differential stable-isotope labeling, compositional isomers can be identified based on the detected isotope distribution patterns. In a similar manner, for the comparative sequencing method described herein, the use of
stable-isotope labeling, rather than enzyme-mediated labeling, may improve singlet identification for those singlets that are interference-prone when only $^{16}\text{O}$ and $^{18}\text{O}$ labeling is used.

Figure 3.12. Calculated codon usage frequencies for E. coli, C. koseri and T. thermophilus. Comparative sequencing is more effective when the codon usage frequencies of the reference and candidate organisms are similar.
Another limitation revealed in these initial studies involves the comparative sequencing of low abundance tRNAs. As found for the scarce tRNA-Arg(UCU), tRNA sequences that are predicted based on gene information may not always be present in sufficient abundance within the sample for detection. While the absence of singlets from both *E. coli* and *C. koseri* tRNA-Arg(UCU) confirmed in this case that this tRNA is of low abundance in both organisms, there may be comparisons where singlets that were not originally predicted are present because the other corresponding tRNA (from either the reference or the candidate organism) is at levels below the limits of detection.

Related to the issues above, it also seems likely that the comparative sequencing approach is best utilized when the tRNA distributions and codon usage frequencies are similar between the reference and candidate organisms (Figure 3.12). It is well documented that organisms possess a codon usage bias, based on the distribution of synonymous codons within any particular organism’s genome.[73] Because tRNA abundance is related to an organism’s codon usage bias,[16, 77, 78] singlets could arise in the experimental data that were not originally predicted if the codon usage bias, and therefore the resulting tRNA abundance, of the reference is significantly different from that of the candidate organism. Singlets arising from the reference would reflect the absence of the corresponding tRNA within the candidate, and likewise singlets arising from the candidate would reflect the absence of the corresponding tRNA within the reference.

### 3.4 CONCLUSIONS

An improved understanding of the functional significance of post-transcriptionally modified nucleosides in ribonucleic acids (RNAs), especially transfer RNAs (tRNAs), requires analytical methods that can efficiently obtain sequence information at the RNA level.
Comparative sequencing of tRNAs provides a relatively straightforward method for obtaining or confirming tRNA sequence information at a far higher scale than is currently possible. While the CARD approach has been applied here to tRNA sequencing, the concept should be equally applicable to other classes of RNAs, although sample complexity may limit its utility when applied to extremely large ribosomal RNAs. In addition, although only bacterial tRNAs have been investigated to date, this approach can be used for comparative sequencing of archaeal or eukaryotic tRNAs, providing suitable reference sequences are available. In the next Chapter, I will present how to process the data analysis rapidly and discover new information using CARD and informatics.
Chapter 4. A Novel Bioinformatics Tool for Targeted and Untargeted RNA in LC-MS

4.1 INTRODUCTION

In Chapters 2 and 3, the expected singlet digestion products were predicted based on gene sequence information and post-transcriptional modification trends. Searching singlets in LC-MS datasets can be performed using the expected singlet digestion products based on tRNA-gene prediction. The data analysis method in LC-MS datasets using predicted singlet digestion products is not only inefficient and time consuming, but it is difficult to search new information with unknown or new post-transcriptional modifications in large LC-MS datasets from the CARD approach as well. Here I describe a novel bioinformatics approach based on Visual Basic scripts for rapid analysis and data mining of singlets in LC-MS datasets. I tested this script by using two datasets from individual tRNA and total tRNAs, respectively, and show that this script executes fast and generates robust and reliable results.

4.2 EXPERIMENTAL

4.2.1 Datasets for development of VBA script

All experimental datasets for testing the script were obtained from experiments in Chapters 2 and 3. These LC-MS datasets are from purified individual tRNA\(^{\text{Cys}}\) and total tRNAs RNase T1 digestion products.
4.2.2 Methods
Before LC-MS datasets from F (forward labeling) and R (reverse labeling) were divided into N segments as ASCII by Xcalibur in Excel spreadsheet, the LC-MS data was smoothed in the Qual Browser in Xcalibur. Each segment of mass spectral data consists of mass spectra in 2 minutes retention time range. The script uses the algorithm in Figure 4.1 and Visual Basic for Application (VBA) codes to process all data. The script includes algorithms for peak picking and comparative analysis processing. The peak picking contains algorithms for isotopic peaks detection, which can detect mass-to-charge ($m/z$) values and ion abundance of A and A+2 peaks and charge state assignment. Comparative analysis processing compares the isotopic distribution patterns or $m/z$ value shift between F and R files. The algorithm of comparative analysis processing identifies singlets in LC-MS datasets based on either $m/z$ values that shift between F and R files as well as the change in A and A+2 isotope peak contributions, namely the relative ratio of ion intensity of A and A+2 isotopic peaks ($I_{\lambda}/I_{\lambda+2}$). In addition, the direction of the $m/z$ shift or the change of isotope peak contributions can be used to identify the source of any singlet.

4.3 RESULTS AND DISCUSSION

4.3.1 Algorithms of VBA script
The VBA script algorithms were designed for automatic and rapid identification of the singlets that were predicted and identified manually in Chapter 2, as well as for discovery of new unpredicted singlets. The criterion in comparative analysis processing for the difference in the $I_{\lambda}/I_{\lambda+2}$ ratio was set from 170% to 600%.
To enable the script to better differentiate singlets from doublets and interferences, another criterion, the b value, was introduced. The b value represents the percentage of A+2 isotopic peak abundance compared to A isotopic peaks abundance of oligonucleotides. Linear regression analysis of the correlation of b values and mass values of oligonucleotides (Figure 4.2) shows that a linear relationship with an $R^2$ of 0.9841. Oligonucleotides plotted were generated from signature digestion products of E. coli total tRNAs[55]. The b values of each digestion product were calculated using Xcalibur, based on their molecular compositions.
In Chapters 2 and 3, the mass values of all the RNase T1 digestion products detected by LC-MS are below 5000 Da. As Figure 4.2 shows, the b value is less than 1.3 for all oligonucleotides below 5000 Da. Thus, in comparative analysis processing, the criteria was set that if the b values (the ratio of $I_A/I_{A+2}$) from the co-eluted peaks between the two experiments are both $\geq 1.3$, the peaks are not considered to be singlets. For the VBA script application in tRNA$^{Cys}$ datasets, after using the b value criterion, seven interference peaks are reduced to one interference peak in the output of script. Thus, with the b value criterion, the interference from doublets or noise signals can be significantly reduced.

![Figure 4.2 The linear regression line of b value and mass values of oligonucleotides](image)

**Figure 4.2** The linear regression line of b value and mass values of oligonucleotides

### 4.3.2 Identification of targeted RNA by VBA script

The VBA script was firstly tested on two LC-MS datasets acquired from individual tRNA$^{Cys}$ of *E. coli* and *C. koseri*. In the data processing, the peak picking detected 206 and 140 peaks in LC-MS datasets from forward labeling and reverse labeling experiments, respectively. After comparative analysis processing, six unique singlet candidates were exported from the hundreds of detected peaks (Table 4.1). The output of results includes the singlet
information: mass value, \(m/z\) value, charge state, the ratio of \(I_A/I_{A+2}\), retention time (RT), and ion abundance (intensity). The three singlets, predicted from gene difference and identified manually in Chapter 2, were identified by four unique singlet candidates from the script output (Table 4.2). The three singlets identified by the script are \(\text{U}[\text{s}^{4}\text{U}]\text{AACAAAGp}\) from \(E.\ coli\), and \(\text{CAAGp}\) and \(\text{U}[\text{s}^{4}\text{U}]\text{AGp}\) from \(C.\ koseri\). The singlet candidate at mass 1647.0 Da is considered as an interference due to the low ion abundance. Therefore, using the VBA script, the identification of predicted singlets is consistent with the ones predicted from tRNA-gene sequences as described in Chapter 2. Moreover, the entire data processing step is rapid, requiring less than two minutes to process the datasets from two experiments.

Table 4.1  The complete output from the script after processing of LC-MS data from tRNA-Cys of \(E.\ coli\) and \(C.\ koseri\). This script identifies \(m/z\) values that shift between labeling experiments as well as differences in A and A+2 isotope peak contributions. The direction of the \(m/z\) shift can be used to identify the source of the singlet.

<table>
<thead>
<tr>
<th>Source</th>
<th>Forward Labeling (tRNA-Cys) (E.\ coli) ((^{16}\text{O}) / C.\ koseri) ((^{18}\text{O})</th>
<th>Reverse Labeling (tRNA-Cys) (C.\ koseri) ((^{16}\text{O}) / E.\ coli) ((^{18}\text{O})</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass (Da)</td>
<td>(m/z)</td>
<td>charge</td>
<td>ratio</td>
</tr>
<tr>
<td>1322.1</td>
<td>660.0</td>
<td>-2</td>
<td>0.22</td>
</tr>
<tr>
<td>1647.1</td>
<td>822.5</td>
<td>-2</td>
<td>2.39</td>
</tr>
<tr>
<td>1679.1</td>
<td>838.3</td>
<td>-2</td>
<td>0.38</td>
</tr>
<tr>
<td>2941.2</td>
<td>1469.6</td>
<td>-2</td>
<td>0.79</td>
</tr>
<tr>
<td>2963.2</td>
<td>1480.6</td>
<td>-2</td>
<td>0.73</td>
</tr>
<tr>
<td>2639.3</td>
<td>1318.6</td>
<td>-2</td>
<td>4.95</td>
</tr>
<tr>
<td>2639.3</td>
<td>1318.6</td>
<td>-2</td>
<td>4.95</td>
</tr>
<tr>
<td>2641.3</td>
<td>1319.6</td>
<td>-2</td>
<td>0.86</td>
</tr>
<tr>
<td>2641.3</td>
<td>1319.6</td>
<td>-2</td>
<td>0.86</td>
</tr>
</tbody>
</table>

4.3.3 Identification of untargeted RNA by VBA script

Surprisingly, a new singlet \((m/z\ 1318.6, -2\ charge state)\) was discovered by the script in the original LC-MS datasets from tRNA\(^{\text{Cys}}\). Based on this singlet shift, the script identified the peak \((m/z\ 1318.6)\) as the singlet generated from \(C.\ koseri\) tRNA\(^{\text{Cys}}\) (Table 4.1, Figure 4.3b).
This new singlet could result from differential post-transcriptional modification. CA[ms\textsuperscript{2i6A}][\Psi][CGp] (m/z 1341.6 u) is the only RNase T1 digestion product of tRNA\textsubscript{Cys} whose mass value is closest to this newly identified singlet. It was found that the mass value for CA[i\textsuperscript{6A}][\Psi][CGp] matches this new singlet. This new singlet sequence (m/z 1318.6, -2 charge state) was identified and confirmed as CA[i\textsuperscript{6A}][\Psi][CGp] based on collision-induced dissociation tandem mass spectrum (Figure 4.4).

<table>
<thead>
<tr>
<th>Interference-Free Singlets Digestion Sequence</th>
<th>Mass (Da)</th>
<th>Sodium adduct mass (Da)</th>
<th>Experimental mass (Da)</th>
<th>Mass difference (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U[s\textsuperscript{4U}]AACAAAGp (E. coli)</td>
<td>2963.4</td>
<td>2963.2</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>2941.4</td>
<td>2941.2</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>CAAAGp (C. koseri)</td>
<td>1677.3</td>
<td>1677.0</td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>1655.3</td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>U[s\textsuperscript{4U}]AGp (C. koseri)</td>
<td>1320.1</td>
<td>1320.0</td>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>

The doublet CA[ms\textsuperscript{2i6A}][\Psi][CGp] and the singlet CA[i\textsuperscript{6A}][\Psi][CGp] arise from sequence C(33)-G(40) in tRNA\textsubscript{Cys} of C. koseri. A(35), when fully modified, becomes ms\textsuperscript{2i6A}. The enzyme miaA is specific for adding the N\textsuperscript{6}-isopentenyl modification to A35 in tRNA\textsubscript{Cys}. N\textsuperscript{6}-isopentenyladenosine (i\textsuperscript{6}A) serves as the substrate for miaB catalyzed modification to 2-methylthio-N\textsuperscript{6}-isopentenyladenosine (ms\textsuperscript{2i6}A). Thus, if A(35) is not fully modified to ms\textsuperscript{2i6}A, it could be modified to be i\textsuperscript{6}A. According to the relative abundance of \textsuperscript{16}O and \textsuperscript{18}O labeled peaks in Figure 4.3a, the level of CA[ms\textsuperscript{2i6A}][\Psi][CGp] of tRNA\textsubscript{Cys} in C. koseri is relatively lower than the level in E. coli. The difference arises from hypomodification of ms\textsuperscript{2i6}A to i\textsuperscript{6}A.
4.3.4 Application of script for LC-MS data analysis of total tRNAs

After confirming the VBA script could identify singlets in tRNA\textsuperscript{Cys}, the script was tested on the LC-MS data from total tRNA. The datasets from total tRNAs digestion products are more complex and contain more information and higher noise signals. Using peak picking, the script detected 349 and 249 peak clusters in LC-MS datasets from forward labeling and reverse labeling experiments of total tRNAs, respectively. Because the datasets from total tRNAs contain higher noise intensity than the ones from tRNA\textsuperscript{Cys}, the threshold value setting in peak picking is much higher than previously. Thus, the number of peak detected in
data from total tRNA is only two times of the detected in tRNA\(^{\text{Gys}}\). By comparative analysis processing, 16 peaks were considered as singlet candidates (Table 4.3).

Figure 4.4 Collision-induced dissociation tandem mass spectrum of CA[i^3A]A[^Ψ]CCGp

Given that there could be retention time shift when using LC separation, the retention time tolerance was set at 2 min. Thus, seven singlet candidates were found within this retention time tolerance. After manual identification of these candidates, seven candidates, which are false singlets, resulted from the deficiency of peak picking algorithms. Five peak candidates, which were manually identified as false singlets, resulted from the different expression of tRNA genes. For example, the peak at \(m/z\) 976.00 in forward labeling experiment is a false
singlet. The peak picking detected the monoisotopic peak ($m/z$, 976.00) as monoisotopic peak in the peaks cluster (Figure 4.5a). Moreover, to mine and discover singlets as many as possible, the criteria for monitoring the significant change of singlets in two experiments was low. For example, the peaks at $m/z$, 971.44 u (Figure 4.5b) in forward labeling is the common digestion product UU[m\(^7\)G]UCGp (mass 1944.9 Da) from tRNA\(^{\text{His}}\) of \(E.\ coli\) and \(C. koseri\). Due to the relative low level of UU[m\(^7\)G]UCGp in \(C. koseri\) compared to \(E. coli\) and the tolerant comparative criteria, this doublet UU[m\(^7\)G]UCGp was exported as a candidate singlet.

Table 4.3 The complete output from the script after processing of LC-MS data from total tRNAs of \(E. coli\) and \(C. koseri\). This script identifies $m/z$ values that shift between labeling experiments as well as differences in A and A+2 isotope peak contributions. The direction of the $m/z$ shift can be used to identify the source of the singlet.
In the comparative analysis of total tRNAs from *E. coli* and *C. koseri*, U[<sup>s</sup>U]<sup>4</sup>AACAAAGp (mass 2941.1 Da) is the only interference-prone predicted singlet, which was manually identified in Chapter 2 and identified by the VBA script in Table 4.3. The new unpredicted singlet CA[i<sup>6</sup>A]A[p60]<sup>g</sup>CCGp (mass 2639.1 Da) was identified by the script in LC-MS datasets of total tRNAs as well, and confirmed in mass spectrum data manually (Figure 4.6b). The mass values of 2685.2 Da and 2707.1 Da (sodium adduct) were exported by the script due to the relative low level of CA[<sup>32</sup>i<sup>6</sup>A]A[p]<sup>Ψ</sup>CCGp in *C. koseri* tRNA<sub>Cys</sub> (Figure 4.6a).

It may be difficult to differentiate this doublet from other singlets. The digestion product CA[<sup>32</sup>i<sup>6</sup>A]A[p]<sup>Ψ</sup>CCGp was identified manually in mass spectral data from RNase T1 digestion products of total tRNAs from single *C. koseri* (Figure 4.7). It turns out that the
RNase T1 digestion product CA[ms2i6A][Ψ]CCGp (m/z 1341.66, -2 charge state) is present in *C. koseri*.

**Figure 4.6** Comparative analysis of total tRNAs of *E. coli* and *C. koseri*. (a) Mass spectra of CA[ms2i6A][Ψ]CCGp (b) Mass spectra of CA[i6A][Ψ]CCGp.

**Figure 4.7** Mass spectrum of CA[i6A][Ψ]CCGp from single *C. koseri* total tRNAs
4.3.5 *C. koseri* tRNA<sub>Cys</sub> sequence

According to the tRNA<sub>Cys</sub> gene sequence between *E. coli* and *C. koseri*, three interference-prone singlet digestion products can be predicted (Table 4.4). The sequence of tRNA<sub>Cys</sub> containing ms<sub>2i6A</sub> can be confirmed by identification of three singlets using the previous CARD approach. Using script data analysis of LC-MS datasets, mapping of variable post-transcriptional modifications is significantly improved. In Table 4.4, four singlet digestion products can be rapidly identified. Compared to the prediction-based approach of Chapters 2 and 3, the script can rapidly identify differences from both gene and post-transcriptional modifications. In sample data from total tRNAs, the difference resulting from post-transcriptional modifications was identified by this script. Two sequences of *C. koseri* tRNA<sub>Cys</sub> were identified and confirmed by MS/MS data. The sequence including ms<sub>2i6A</sub> and the one including i<sup>6</sup>A are listed in Figure 4.8.

**Table 4.4** Comparison of the singlets identified using the script outlined in Figure 4 to those that are identified by comparing tRNA gene sequences between two organisms. The bottom singlets were identified from within the LC-MS datasets from RNase T1 digestion of total tRNAs.

<table>
<thead>
<tr>
<th>Interference-Free Singlets Digestion Product</th>
<th>Mass (Da)</th>
<th>Sample Datasets</th>
<th>Predicted by gene difference?</th>
<th>Identified by Script?</th>
</tr>
</thead>
<tbody>
<tr>
<td>U[s&lt;sup&gt;4&lt;/sup&gt;U]AACAAAGp</td>
<td>2941.4</td>
<td><em>E. coli</em> tRNA&lt;sub&gt;Cys&lt;/sub&gt;</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CA[i&lt;sup&gt;6&lt;/sup&gt;A][Ψ]CCGp</td>
<td>2639.4</td>
<td><em>C. koseri</em> tRNA&lt;sub&gt;Cys&lt;/sub&gt;</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>CAAAGp</td>
<td>1655.3</td>
<td><em>C. koseri</em> tRNA&lt;sub&gt;Cys&lt;/sub&gt;</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>U[s&lt;sup&gt;4&lt;/sup&gt;U]AGp</td>
<td>1320.1</td>
<td><em>C. koseri</em> tRNA&lt;sub&gt;Cys&lt;/sub&gt;</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>U[s&lt;sup&gt;4&lt;/sup&gt;U]AACAAAGp</td>
<td>2941.4</td>
<td><em>E. coli</em> Total tRNAs</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CA[i&lt;sup&gt;6&lt;/sup&gt;A][Ψ]CCGp</td>
<td>2639.4</td>
<td><em>C. koseri</em> Total tRNAs</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
4.4 CONCLUSIONS

This novel bioinformatics tool based on VBA script for identifying singlets generated in the comparative analysis approach for tRNA characterization was found to be equal to or better than the significantly more time-consuming approach of manual identification from tRNA gene-based predictions. For any RNase T1 digestion products arising from tRNA hypomodification, their presence is not readily predicted from tRNA gene sequence information. For instance, an unexpected singlet, CA[^6A][Ψ]CCp, from *C. koseri* was identified using this script. By using this script for data analysis, not only the targeted RNA can be rapidly identified, but untargeted differences can discovered from large mass spectral datasets.
Chapter 5. Identification of RNA Sequence Isomers by Isotope Labeling and LC-MS/MS

5.1 INTRODUCTION

As noted in Chapters 2 and 3, some digestion products generated during the CARD approach share identical $m/z$ values but possess different sequences. The identification of isomeric digestion products by LC-MS is not possible, unless these isomers can be separated at the HPLC step. To address this limitation, here I developed a tandem mass spectrometry (MS/MS) approach to identify tRNA sequence isomers by tracking isotope labeled fragment ions within the tandem mass spectrum. Isomeric sequences will generate product ions during collision induced dissociation (CID) that can be used to differentiate the sequences. The y-type ions dissociated from identical digestion product sequences will form doublets separated by 2 Da in the tandem mass spectrum due to $^{16}$O and $^{18}$O labeling (Figure 5.1). In contrast, some y-type ions from RNA sequence isomers will appear as singlets. After reverse labeling, these singlets will shift by 2 Da. Therefore, the ambiguities of identification of RNA sequence isomers are reduced by the presence of unique fragments and by recognition of peaks shift as discussed in Chapter 2. This MS/MS step can be performed with the CARD approach to improve the identification of singlets.
5.2 Experimental

5.2.1 Materials

*Escherichia coli* tRNA^{Tyr} and tRNA^{Glu} were purchased from Sigma Aldrich. All other materials used are described in Chapter 2.

5.2.2 Isolation of tRNAs

Bacterial cultures and isolation of tRNAs were performed as described in Chapter 2.

5.2.3 Ribonuclease Digestions and Isotope Labeling

The procedure for enzymatic digestion and isotope labeling are described in Chapter 2.

5.2.4 Liquid Chromatography-Tandem Mass Spectrometry

All LC/MS and LC/MS-MS analyses were performed using a MicroAS autosampler, Surveyor MS Pump Plus HPLC system and Thermo LTQ-XL (Thermo Scientific, Waltham, MA) mass spectrometer equipped with an electrospray ionization (ESI) source. The ESI source was set to a capillary temperature of 275 °C and spray voltage of 4.50 kV. Data was recorded in negative polarity and profile data type using a zoom scan with the mass range from \( m/\zeta \) 600 to 2000 in scan event 1, which acquires the \( m/\zeta \) value of digestion products. In scan event 2, sequence information of the most abundant digestion product was obtained by data-dependent collision-induced dissociation (CID) with the optimized isolation width 6.
5.3 Results and Discussion

Within the CARD approach, RNA sequences that are identical between the two samples will yield RNase digestion products that appear as doublets, separated by 2 Da (\(^{16}\text{O}\) vs \(^{18}\text{O}\)) in the
mass spectral data, but there may be some unidentified RNA sequence isomers. For example, the RNase digestion product CAG was labeled with $^{16}$O and $^{18}$O. Its sequence isomer, ACG, labeled with either $^{16}$O or $^{18}$O will co-elute with digestion product CAG and appear as a doublet in mass spectral data. To address this limitation, the basis of this approach for RNA sequence isomer identification is presented in Figure 5.1. Within this comparative MS/MS approach, the identical RNase digestion products will yield y-type fragment ions that appear as doublets, separated by 2 Da ($^{16}$O vs $^{18}$O) in the tandem mass spectral data. RNA sequence isomers can be identified by the characteristic y-type fragment ions as singlets.

5.3.2 Optimization of MS/MS Conditions

As an initial illustration of this comparative MS/MS approach for $^{16}$O and $^{18}$O labeled RNA, the MS/MS condition were optimized to simultaneously isolate $^{16}$O and $^{18}$O labeled digestion products for CID. The isolation width parameter determines the total width of the gate applied around a selected precursor ion. The isolation widths (4 u to 6 u) were used to select 3-mer and 10-mer oligonucleotide digestion products labeled by $^{16}$O and $^{18}$O. The 10-mer oligonucleotide digestion product AAUCCCCUAGp from tRNA-Tyr1 generated doublet y-type product ions with an isolation width of 4 u (Figure 5.2.d). When the isolation width was increased to 6 u, the 3-mer oligonucleotide digestion product CAGp from tRNA-Tyr1 yielded doublet y-type product ions (Figure 5.2.c).
Figure 5.2 Optimization of isolation width during MS/MS to enable detection of the isotopic doublet present on y-type fragment ions.
tRNA-Glu2 Sequence
GUCCCUUCGUCΨAGAGGCCCAGGACACCGGCCU(mnm^{5}s^{2}U)UC(m^{2}A)CGGCUGUAACAGGGGTTΨCGAAUCCCUAGGGGACGCCA

tRNA-Tyr1 Sequence:
GGUGGGG(s^{4}U)UCCCGAGC(Gm)GCCAAAGGGAGCAGACUQUA
(ms^{2}i^{6}A)AΨCGCCGUACUCGACUUCGAGGTΨCGAAUCCUUCGCCACCACCA

Figure 5.3 The sequences of tRNAs-Glu2 and Tyr1. The bold are the sequences of the digestion product sequence isomers ACGp and CAGp generated by RNase T1.

Figure 5.4 Mass spectra of two oligonucleotides ACGp and CAGp
Figure 5.5 Tandem mass spectra illustrating the identification of ACGp in the presence of the interfering sequence isomer, CAGp, using isotope-labeling and tandem mass spectrometry. (a) Collision induced dissociation mass spectrum of oligonucleotides from tRNA-TyrI,GluII with $^{16}$O and tRNA-GluI with $^{18}$O. (b) Collision induced dissociation mass spectrum of multiple oligonucleotides from tRNA-TyrI,GluII with $^{18}$O and tRNA-TyrI with $^{16}$O.
5.3.3 Comparative MS/MS Analysis of tRNA-Glu2/Tyr1 Model

Two standard tRNAs-Glu2 and Tyr1 whose sequences are known (Figure 5.3) were used to build a test model for this MS/MS method. A mixture of tRNA-Glu2 and tRNA-Tyr1 was digested initially with $^{16}$O water, while tRNA-Tyr1 alone was digested in $^{18}$O water. The RNase T1 digests were combined in equal amounts and analyzed by LC-MS/MS (Figure 5.3). In LC-MS analysis, the RNase T1 $^{16}$O-labeled digestion product ACGp from tRNA-Glu2 cannot be identified due to the presence of $^{16}$O- and $^{18}$O-labeled sequence isomer, CAGp from tRNA-Tyr1 (Figure 5.4a). Reverse labeling mass spectral data cannot be used to identify or confirm the presence of digestion product ACGp (Figure 5.4b). Only minor change in relative abundance of $^{16}$O and $^{18}$O labeled products was shown after reverse labeling. Within the optimized CID MS/MS method, the digest ACGp was easily identified by its fragment CGp ion, namely $y_2$ ion ($m/z$ 667.1) (Figure 5.5a). Because isotope labeling is limited to the 3'-phosphate of the RNase T1 digestion products, this label will be present in all $y$-type fragment ions after MS/MS. As noted above, the identity of sequence isomer ACGp can be confirmed by reverse labeling (Figure 5.5b). As expected, if the $y$-type ion fragment arises from ACGp, after reverse labeling, the $m/z$ value increased by 2 u (due to detection of 1- charge state).

Table 5.1 RNA sequence isomers predicted based on RNase T1 digestion of E. coli wt and miaA-Δ mutant tRNAs.

<table>
<thead>
<tr>
<th>tRNA</th>
<th>Sequence</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wt Asp(QUC)</td>
<td>AAUACCUGp</td>
<td>2572.3</td>
</tr>
<tr>
<td>Mutant Asp(QUC)</td>
<td>AAUACCUGp</td>
<td>2572.3</td>
</tr>
<tr>
<td>wt Leu (GAG)</td>
<td>CCCAAUAGp</td>
<td>2571.4</td>
</tr>
<tr>
<td>Mutant Leu (GAG)</td>
<td>CCCAAUAGp</td>
<td>2571.4</td>
</tr>
<tr>
<td>Mutant Cys(GCA)</td>
<td>CAAA[ψ]CCGp</td>
<td>2571.4</td>
</tr>
</tbody>
</table>
Table 5.2 Expected y-type fragment ion $m/z$ values for CAA[$\Psi$]CCGp, CCAAUAGp and AAUACCUGp

<table>
<thead>
<tr>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
<th>Fragment</th>
<th>m/z</th>
</tr>
</thead>
<tbody>
<tr>
<td>y₁⁻</td>
<td>362.0</td>
<td>y₁⁻</td>
<td>362.0</td>
<td>y₁⁻</td>
<td>362.0</td>
</tr>
<tr>
<td>y₂⁻</td>
<td>667.1</td>
<td>y₂⁻</td>
<td>691.1</td>
<td>y₂⁻</td>
<td>668.1</td>
</tr>
<tr>
<td>y₃⁻</td>
<td>972.1</td>
<td>y₃⁻</td>
<td>997.1</td>
<td>y₃⁻</td>
<td>973.1</td>
</tr>
<tr>
<td>y₄⁻</td>
<td>1278.2</td>
<td>y₄⁻</td>
<td>1326.2</td>
<td>y₄⁻</td>
<td>1278.2</td>
</tr>
<tr>
<td>y₅⁻</td>
<td>1607.2</td>
<td>y₅⁻</td>
<td>1655.2</td>
<td>y₅⁻</td>
<td>1607.2</td>
</tr>
<tr>
<td>y₅⁻²⁻</td>
<td>803.1</td>
<td>y₅⁻²⁻</td>
<td>827.1</td>
<td>y₅⁻²⁻</td>
<td>803.1</td>
</tr>
<tr>
<td>y₆⁻</td>
<td>1936.3</td>
<td>y₆⁻</td>
<td>1960.3</td>
<td>y₆⁻</td>
<td>1913.2</td>
</tr>
<tr>
<td>y₆⁻²⁻</td>
<td>967.6</td>
<td>y₆⁻²⁻</td>
<td>979.6</td>
<td>y₆⁻²⁻</td>
<td>956.1</td>
</tr>
<tr>
<td>y₇⁻²⁻</td>
<td>1132.2</td>
<td>y₇⁻²⁻</td>
<td>1132.2</td>
<td>y₇⁻²⁻</td>
<td>1120.6</td>
</tr>
<tr>
<td>c₁⁻</td>
<td>304.0</td>
<td>c₁⁻</td>
<td>304.0</td>
<td>c₁⁻</td>
<td>328.0</td>
</tr>
<tr>
<td>c₂⁻</td>
<td>633.1</td>
<td>c₂⁻</td>
<td>609.1</td>
<td>c₂⁻</td>
<td>657.1</td>
</tr>
<tr>
<td>c₃⁻</td>
<td>962.1</td>
<td>c₃⁻</td>
<td>914.1</td>
<td>c₃⁻</td>
<td>963.1</td>
</tr>
<tr>
<td>c₄⁻</td>
<td>1291.2</td>
<td>c₄⁻</td>
<td>1243.2</td>
<td>c₄⁻</td>
<td>1292.2</td>
</tr>
<tr>
<td>c₅⁻</td>
<td>1597.2</td>
<td>c₅⁻</td>
<td>1572.2</td>
<td>c₅⁻</td>
<td>1597.2</td>
</tr>
<tr>
<td>c₆⁻</td>
<td>1902.3</td>
<td>c₆⁻</td>
<td>1878.2</td>
<td>c₆⁻</td>
<td>1902.3</td>
</tr>
<tr>
<td>c₇⁻²⁻</td>
<td>1103.1</td>
<td>c₇⁻²⁻</td>
<td>1103.1</td>
<td>c₇⁻²⁻</td>
<td>1103.6</td>
</tr>
</tbody>
</table>

5.3.4 Application of Comparative MS/MS Analysis to Real Biological Samples

The application of this comparative MS/MS analysis approach for RNA sequence isomer characterization by tandem mass spectrometry is used for comparative analysis of *E. coli* miaA-Δ mutant tRNAs. In Chapter 2, the singlet CAAA[$\Psi$]CCGp from miaA-Δ mutant cannot be identified as singlet in mass spectral data, due to the interference of other two RNA sequence isomers CCAAUAGp and AAUACCUGp. The mass difference among these RNase digestion products is within 1 Da (Table 5.1), which cause the difficulty in separation by chromatography and identification in mass spectrometry. Within the comparative MS/MS approach, the singlet digestion product CAAA[$\Psi$]CCGp could be identified by some unique y-type product ions in CID MS/MS. In Table 5.2, the y-type product ion in bold from CAAA[$\Psi$]CCGp, whose masses were unique compared to the y-
type products of other RNA sequence isomers, can be used as unique ions for identification of CAAA\[Ψ\]CCGp in CID MS/MS.

Figure 5.6 Identification of tRNA-Cys CCAAUAGp from *E. coli* miaA-D mutant in presence of interfering sequence isomers CCAAUAGp and interfering digestion product AAUACCUGp (a)(c) Collision induced dissociation mass spectrum of multiple oligonucleotides from *E. coli* wt with \(^{16}\text{O}\) and miaA-Δ mutant with \(^{18}\text{O}\). (b)(d) Collision induced dissociation mass spectrum of multiple oligonucleotides from *E. coli* wt with \(^{18}\text{O}\) and miaA-Δ mutant with \(^{16}\text{O}\).
The mixture digestion products in the MS/MS spectra are $^{18}$O-labeled CAAA[Ψ]CCGp, $^{16}$O- and $^{18}$O-labeled CCCAUAGp and $^{16}$O- and $^{18}$O-labeled AAUACCUGp. Due to the complexity of multiple oligonucleotides, mass spectral interpretation is very complex. Using the approach used for ACGp model, the $y$-type fragment peak at $m/z$ value 973.1 ($y_3$ ion) from $^{18}$O-labeled CAAA[Ψ]CCGp of miaA-Δ mutant can be identified (Fig. 5.6a, 5.6c). After reversing labeling, the identity of CAAA[Ψ]CCGp can be confirmed and analysis based on the increasing intensity of $^{16}$O-labeled $y_3$ ion ($m/z$ 972.1) in Fig 5.6b, 5.6d. The other potential singlets cannot be identified because they are interfered by CID fragments of $^{16}$O and $^{18}$O labeled isomers or they are a low abundance fragment due to the terminal fragments (such as $y_2$, $y_6$ ions in 7-mer oligonucleotides).

5.4. Conclusions

A MS/MS method based on CARD approach has been developed for efficiently analyzing RNA sequence isomers. In addition, because this comparative MS/MS analysis approach is directly compatible with previous CARD approach [71, 79], comparative analysis of RNA sequence isomer by LC-MS/MS and comparative analysis of RNA by LC-MS can be performed within a single analysis. Therefore, this approach enables the simultaneous identification of interference-free singlets and interference-prone singlets, which represents sequence difference or differential modification in tRNA level. This approach can be used with the CARD approach to significantly improve the identification of tRNA differences. Moreover, the ambiguities of identification of interference-prone singlets (RNA sequence isomers) can be reduced by the presence of unique fragments and by recognition of peaks shift of these fragments after reverse labeling. When the number of RNA sequence isomers
is less than or equal to two, the identification of the singlet digestion product in the presence of RNA sequence isomers is possible. If the number of RNA sequence isomer is more than two, the tandem mass spectral data can be quite complex for interpretation. For example, the singlet AAAACCGp from tRNA-Ser can not be identified by its unique product ion in CID MS/MS in the presence of three RNA sequence isomers. With the presence of more RNA sequence isomers, the less the number of unique product ions from the singlet digestion product that can be identified in MS/MS.
Chapter 6 Conclusions and Future Perspectives

6.1 Summary

The overall goal of this dissertation is to develop mass spectrometric approaches for high-throughput automated comparative analysis of RNAs. To accomplish this goal, two major aims were developed. First, the mass spectrometric method for high-throughput comparative analysis of RNA must be developed without requiring purification of individual RNAs. Second, an approach for the rapid processing of data from large mass spectral datasets must be developed.

In Chapter 2, an approach based on the comparative analysis of RNAs, designated the CARD (Comparative Analysis of RNA Digests) approach, was developed for efficiently analyzing one or more RNA sequences without resorting to complete RNA sequencing and which is suitable for the characterization of unknown RNAs. The CARD approach provides a relatively straightforward method for obtaining or confirming RNA sequence and/or modification information at a far higher scale than is currently possible, and the approach is enabled by encoding information relating to differences between the reference and unknown RNA(s) through isotopic labeling. Although RNase-mediated isotopic labeling was used here, the approach should be directly compatible with other stable-isotope labeling approaches used in RNA mass spectrometry.[70] Moreover, because this approach is directly compatible with previously reported quantitation strategies for RNA analysis,[62, 63] future iterations could provide qualitative and quantitative characterization of unknown RNAs within a single analysis. Although tRNA sequencing was illustrated using a single isolated tRNA in this work, the comparative analysis of the *E. coli* wild type and miaA-Δ mutant strains for
modification analysis using unfractionated tRNAs suggests CARD can readily be extended to tRNA sequencing at the organism level for bacterial, archaeal or eukaryal tRNAs, providing suitable reference sequences are available.

Next, the application of this CARD approach for sequencing of total transfer RNAs from *Citrobacter koseri* is presented in Chapter 3. Using CARD, about 80% of the tRNAs from the bacterium *Citrobacter koseri* can be sequenced using ribonuclease T1 with *E. coli* tRNAs as the reference. During these studies, I also discovered a sequence error for *E. coli* tRNA-Thr1, and I used this method to confirm the correct sequence for that tRNA. While the CARD approach has been applied here to tRNA sequencing, the concept should be equally applicable to other classes of RNAs, although sample complexity may limit its utility when applied to extremely large ribosomal RNAs. In addition, although only bacterial tRNAs have been investigated to date, this approach can be used for comparative sequencing of archaeal or eukaryotic tRNAs, providing suitable reference sequences are available.

In Chapter 4, a novel bioinformatics tool for identification of both targeted and untargeted RNA in LC-MS datasets is presented. This novel bioinformatics tool, based on VBA scripts for identifying singlets generated in the comparative analysis approach for tRNA characterization, was found to be equal to or better than the significantly more time-consuming approach of manual identification from tRNA gene-based predictions. For any RNase T1 digestion products arising from tRNA hypermodification, their presence is not readily predicted from tRNA gene sequence information. For instance, an unexpected singlet, CA[\text{\text{\textsuperscript{6}A}}][\text{\text{\textsuperscript{P}}}]{\text{CCG}}_p$, from *C. koseri* was identified using this script. By using this script for data analysis, not only can the targeted RNA be rapidly identified, but new information can be discovered.
In Chapter 5, a method was developed for identification of RNA sequence isomers by isotope labeling and tandem mass spectrometry. This MS/MS method based on the CARD approach was developed for efficiently analyzing RNA sequence isomers. In addition, because this comparative MS/MS analysis approach is directly compatible with my previous CARD approach [71, 79], comparative analysis of RNA sequence isomers by LC-MS/MS and comparative analysis of RNA by LC-MS can be performed within a single analysis. Therefore, this approach enables the simultaneous identification of interference-free singlets and interference-prone singlets, which represent sequence differences or differential modifications in tRNAs. Moreover, the ambiguities of identification of interference-prone singlets (RNA sequence isomers) can be reduced by the presence of unique fragments and by recognition of the peak shifts of unique fragments after reverse labeling.

6.2 Future experiments

6.2.1 Simplifying data analysis of CARD approach

In Chapters 2 and 3, RNA isolated from conventional culture media contains a natural isotopic distribution. When the digestion products from these RNAs are combined with $^{18}$O labeled RNAs, the mass spectral data is quite complex. To reduce the difficulty of mass spectral interpretation, isotopically-enriched media can be used to simplify the mass spectral data. For example, the doublet of $^{16}$O and $^{18}$O labeled $[\text{m}^7\text{G}]UCCCCAGp$ from tRNA-Thr(GGU), which has been identified as doublet in Figure 3.5 of Chapter 3, cultured in conventional media is presented as a complex doublet mass spectral data in Figure 6.1a. In contrast, the doublet from Carbon 12 ($^{12}$C) enriched isotopic media is much simpler and easier to recognize as a doublet (Figure 6.1b). Other applications of enriched isotopic
labeling can be used for the identification of interference-prone singlet digestion products, and for the identification of RNA sequence isomers.

![Figure 6.1 Comparison of the doublet [m’G]UCCCCAGp of tRNA-Thr(GGU) cultured in natural culture media and isotope enriched media](image)

6.2.2 Software for comparative analysis of RNA

In Chapter 4, the algorithms based on the VBA script for comparative analysis of RNA were developed. The input of LC-MS datasets is not completely automated, so future improvements in data input are necessary. The raw files generated from the LTQ-XL mass spectrometer can be converted to the mzXML format. The challenge will be revise the code to readily read this format. This improvement would allow the high-throughput screen of tRNAs from multiple mutant organisms (or disease tissues) with the ones from wild type (or normal cells). The CARD approach combined with the high-throughput data analysis method could have potential applications in clinical studies and disease diagnostic in the future.
The development of data analysis software is laborious and time-consuming. Thus, besides the development of the algorithm based on the VBA script, existing software should be explored for comparative analysis of RNA. The evaluation of four non-commercial tools (MzMine, MSight, msinspect, and MetAlign) for computational processing of LC-MS datasets has been reported. [80] The above four tools are mainly used for label-free proteomics. Because the software has the functions for label-free quantitative proteomics and LC-MS profiling, the concept of comparative analysis of label-free RNA is proposed and would be enabled using the relevant software for data processing of LC-MS datasets.

6.2.3 The concept of label-free comparative analysis of RNA

Quantitative proteomics has been developed using either stable isotope labeling or label-free approaches. LC-MS quantitation of peptides is performed by comparing peak intensities between multiple runs obtained by continuous detection in MS mode. To compare data from multiple LC-MS runs, the software must detect, match and align MS peaks. The most recent software packages that are available for comparative LC-MS have been reported. [81] Based on this concept, the direct comparative analysis of RNA without isotope labeling would be possible using the alignment function in LC-MS datasets of the relevant software, such as PLGS, MzMine, or OpenMS. Once the automated and accurate alignment of peaks in LC-MS datasets can be performed, a label-free comparative approach for RNA will be enabled. Within CARD approach, reverse labeling provides a straightforward way to confirm the differences according to the singlets shift, which is so far a best way to assure the accuracy for singlets identifications, especially for the confirmation of new information from the differences. Compared to the reverse labeling confirmation in CARD, a label-free
approach will be a big challenge to accurately confirm the differences in MS modes due to the quite complex LC-MS datasets. The development of a label-free comparative approach for RNA will be a long term and tough task.

6.2.4 Potential application of comparative analysis of RNA

The CARD approach provides a relatively straightforward method for obtaining or confirming RNA sequence and/or modification information at a far higher scale than is currently possible, and the approach is enabled by encoding information relating to differences between the reference and unknown RNA(s) through isotopic labeling. Although the CARD approach was illustrated using a known reference, CARD could be readily extended to directly compare unknown mutants with unknown reference, where only differences would be presented as singlets that can be further characterized to identify how the candidate differs from the reference, even when the reference is unknown. Using the algorithms for the identification of untargeted singlets in Chapter 4, comparative analysis of two unknown RNA samples to discover their differences is possible. This concept of comparative analysis using an unknown RNA reference would have a potential impact on the RNA biomarker or RNA molecular diagnostics in human diseases, and it also provides a novel way to investigate the relationship between post-transcriptional modification of RNA and human diseases, such as cancer.

In addition, the CARD approach has been extended to comparative sequencing of tRNAs from homologous organisms, such as Citrobacter koseri. To extend this method to more applications in the field of RNA, the largest phylogenetic distance between the reference and
the candidate's tRNA sequencing using CARD approach could be preliminarily investigated by comparative genomics. We could evaluate the success of the CARD approach based on the difference between two organisms’ tDNA sequences. The greater the difference between these tDNA sequences reduces the number of reference sequences required.

In all, to improve and apply the CARD approach into more applications, multiple aspects could be improved, applied and developed: software for data analysis, homologous sequencing applications, and the concept of label-free comparative approach for RNA. Using high-throughput automated comparative analysis of RNA, more and more discoveries regarding RNA and their post-transcriptional modification will be readily completed.
Bibliography


