2019

Mass Spectrometry: An Ideal Method For Rna Modification Analysis

Samuel Peter Wein
University of Pennsylvania, sam@samwein.com

Mass Spectrometry: An Ideal Method For Rna Modification Analysis

Abstract
Currently there is no good way to measure and find the exact location of multiple RNA modifications. Existing technology can effectively find single varieties of modifications, but cannot identify co-occurrence. As the field of proteomics has shown, mass spectrometry is a powerful and versatile technique assessing broad ranges of chemical modifications in the context of the cellular environment. In this project I used our expertise in proteomics to build a mass spectrometry based platform for identifying RNA modifications. I have since set up both software and analytical platforms querying RNA modifications, and used this platform to survey human tRNA samples and identify what modifications there are, and where they occur.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
Benjamin A. Garcia

Keywords
Mass spectrometry, OpenMS, RNA, RNA modification

Subject Categories
Biochemistry | Bioinformatics

MASS SPECTROMETRY: AN IDEAL METHOD FOR RNA MODIFICATION ANALYSIS

Samuel Wein

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Benjamin A. Garcia, Presidential Associate Professor of Biochemistry and Biophysics

Graduate Group Chairperson

Kim A. Sharp, Associate Professor of Biochemistry and Biophysics

Dissertation Committee

Kim A. Sharp, Associate Professor of Biochemistry and Biophysics

David W. Speicher, Professor & Co-Leader, Molecular & Cellular Oncogenesis Program

Jeremy E. Wilusz, Assistant Professor of Biochemistry and Biophysics

John Karanicolas, Associate Professor, Fox Chase Cancer Center

MASS SPECTROMETRY: AN IDEAL METHOD FOR RNA MODIFICATION ANALYSIS

Dedicated to Amanda Wein

ACKNOWLEDGEMENT

I would like to thank my wonderful wife Amanda for her help proofreading, and general support throughout the PhD process. Without her none of this would be possible. I would also like to thank both Hendrik and Byron for opening their homes to me when I was in Cambridge working on the NASE manuscript, as well as Timo, Hannes, and the rest of the OpenMS crew for answering all of my questions and for teaching me. I'd also like to thank Dave Jackson for introducing me to bioinformatics, and all of the wonderful encouragement he has provided. I'd like to thank all of the wonderful folks at Dock Street for serving me beer and listening to me grouse about the dissertation process, a fair portion of this document was written sitting at the bar in their Brewery.

ABSTRACT

MASS SPECTROMETRY: AN IDEAL METHOD FOR RNA MODIFICATION ANALYSIS

Samuel Wein

Benjamin A. Garcia

Currently there is no good way to measure and find the exact location of multiple RNA modifications. Existing technology can effectively find single varieties of modifications, but cannot identify co-occurrence. As the field of proteomics has shown, mass spectrometry is a powerful and versatile technique assessing broad ranges of chemical modifications in the context of the cellular environment. In this project I used our expertise in proteomics to build a mass spectrometry based platform for identifying RNA modifications. I have since set up both software and analytical platforms querying RNA modifications, and used this platform to survey human tRNA samples and identify what modifications there are, and where they occur.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

The process of research leading to this dissertation was not nearly as linear as it is presented in this work. There were lots of false starts, bits of tribal knowledge to be learned and a good deal of work that could be the subject of numerous other dissertations had I infinite time to pursue it. I hope that this work is helpful to any scholars who come after me, and I welcome any questions or requests for further insight into this research. This has been a labor of love and creativity and I am immensely grateful to have been given the opportunity to undertake it.

CHAPTER 1 : Background

1.1. Introduction

The wild world of RNA The study of RNA has been of interest to biologists for some time now. RNA fills many different roles in the cell; it is the intermediary between DNA and protein in the form of mRNA, it is the effector for translation in the form of tRNA, it constitutes a key component of the ribosome in the form of rRNA, and it is a regulatory molecule controlling translation levels in the form of miRNA. For almost as long as we have known that the sequence of RNA molecules (the nucleic acid sequence) affects its function, we have known that there are also other chemical modifications which shape the functionality of the molecule. Since these modifications mostly happen after the transcription from DNA to RNA, I will refer to them as post-transcriptional modifications (PTMs) throughout this work. It should be noted that there are also modifications that can occur to mRNA during translation (that is co-translationally)Afonin et al. (2012) for the purposes of this work I lump them in with PTMs. In this chapter I will explore broadly the existing understanding of RNA biology, and the effects of modifications on the RNA. I will then look at mass spectrometry (MS) as an analytical technique, its history, and how it can be used to solve critical problems that currently exist in the field of research surrounding modifications to RNA. To describe this discipline I use the term epitranscriptomics. I will also lead the reader through a close look at the field of proteomics, a much more well-developed field encompassing the study of proteins and their modifications by mass spectrometry.

What your traditional RNA sequencing techniques won't tell you The field of RNA sequencing is already relatively heavily populated. Scientists are able to sequence large swathes of RNA quickly and inexpensively with reverse transcription polymerase chain reaction (rtPCR) followed by "next-generation" DNA sequencing (see figure 1 for details) Bustin et al. (2005). Modifications which do not add or remove bases to the RNA molecule

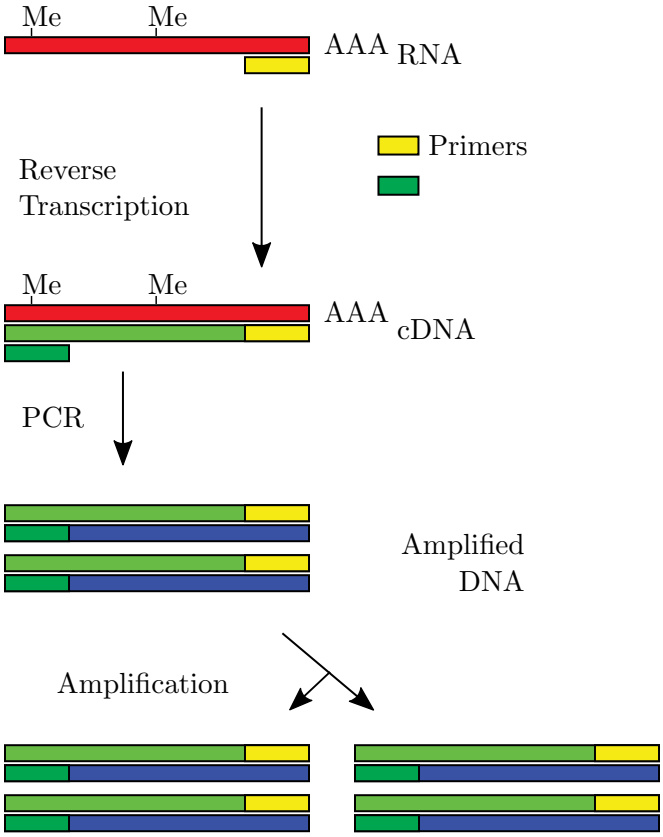

Figure 1: Reverse transcription polymerase chain reaction consists of two distinct steps, the initial reverse transcription followed by the amplification of reverse transcribed product through the polymerase chain reaction (PCR). In the reverse transcription step the enzyme reverse transcriptase creates a DNA copy of any RNA matching a short DNA-primer. This product is called cDNA. The cDNA is then amplified through successive rounds of PCR. During the reverse transcription phase any post-transcriptional modifications to the RNA are replaced in the cDNA by their canonical base paired nucleotide, making rtPCR incompatible with identifying RNA modifications. rtPCR is a prequisite to sequencing since all extant sequencing methods require substantially more nucleic acid than is available directly from biological sources. Wein (2019)

are not preserved during traditional methods of nucleotide sequencing. During the PCR amplification step prior to sequencing, modifications are not templated to new copies and are instead transcribed as whichever base they pair to. As we have become increasingly aware of the importance of these chemical modifications, it has become apparent that we need better options for sequencing modified RNA. Currently several techniques exist to detect single specific modifications. For example, bisulfite sequencing can be used to detect 5-methylcytosine Gilbert et al. (2016), and PSI-seq can detect pseudouridine Lovejoy et al. (2014), however these techniques only work for a single modification each. Leaving it impossible to determine co-occurrences of modifications. In order to identify and characterize these modifications, we have turned to the technique of mass spectrometry as it offers an unbiased and sensitive approach to determining analyte composition. MS allows us to detect the mass shift between unmodified and modified oligonucleotides and does not require amplification in order to provide enough sample for sequencing. Furthermore, by fragmenting nucleic acids and capturing tandem mass spectra, it is possible to localize which base is modified by comparing the mass shift of individual fragments. Other researchers have shown that it is eminently possible to observe miRNAs by mass spectrometry Kullolli et al. (2014), Yu et al. (2005) and other work on larger tRNA (Hossain and Limbach (2007)) and rRNAs (Taoka et al. (2016a)) has demonstrated that mass spectrometry provides a good mechanism for discovering and localizing modifications on these larger molecules Taoka et al. (2015). My work here extends this search and, crucially, presents a new software platform to make the analysis of complex modifications and complex mixtures tractable.

Both the analytical and bioinformatic methods for nucleotide mass spectrometry are less well developed in comparison to those for protein mass spectrometry. Here, I show my work to create methodology to perform nanoflow HPLC and tandem mass spectrometry on undigested miRNA-like oligonucleotides (˜20-25 nucleotide bases in length) and on longer RNA types (such as tRNA) when coupled with digestion by an enzyme, as well as adapt to and develop software tools for their analysis. I have expanded upon work by other groups to create an analytical method to separate and sequence RNAs, and have created new software

to analyze the tandem MS results by building off of OpenMS, an open-source set of libraries and programs developed to facilitate MS analysis Rost et al. (2016).

Mass spectrometry as an analytical technique Mass spectrometry is a technique with a long history of use for a variety of fields Thomson (1921). In mass spectrometry, molecules in the sample of interest are ionized (endowed with an electrical charge) and their mass to charge ratio is measured. This technique is broadly applicable, being useful for everything from metals (using ion coupled plasma mass spectrometry Balcaen et al. (2015)) to proteins and other biopolymers (using electrospray ionization mass spectrometry Fenn et al. (1989)). In this work I will discuss the history of mass spectrometry, which has led us to this point, and how the technique can be expanded to the new field of RNA modifications. Specifically I will look at lessons from the field of proteomics (the large-scale study of proteins) which has a relatively mature software and analytical ecosystem, and how we can extend those tools to RNA as another bio-polymer.

1.2. RNA

What is RNA? RNA, or RiboNucleic Acid, is a polymeric biomolecule composed of a chain of nucleotides. Much like the more widely discussed DNA, each nucleotide consists of a nitrogenous base, a 5-carbon sugar–Ribose, and a phosphate group. Crucially there are different bases with different chemical compositions. For RNA, these are canonically, Adenine (A), Guanine (G), Cytosine (C), and Uracil (U). These nucleotides form a linear sequence, with the base of one attached to the phosphate of the next. The identity of the base is important in a huge number of biological processes such as protein production, regulation of enzyme levels, and degradation of invading viral particles. Nomenclature for describing the different "sides" of the RNA talks about 5' and 3' ends. These correspond to which carbon on the ribose (numbered 1' to 5') is exposed. Unlike DNA, RNA is often single stranded, and forms a much wider variety of secondary and tertiary structures. Complementary bases (A pairs with U, C pairs with G) form hydrogen bonds between different portions of the same molecule, enabling the formation of a diverse group of key structures

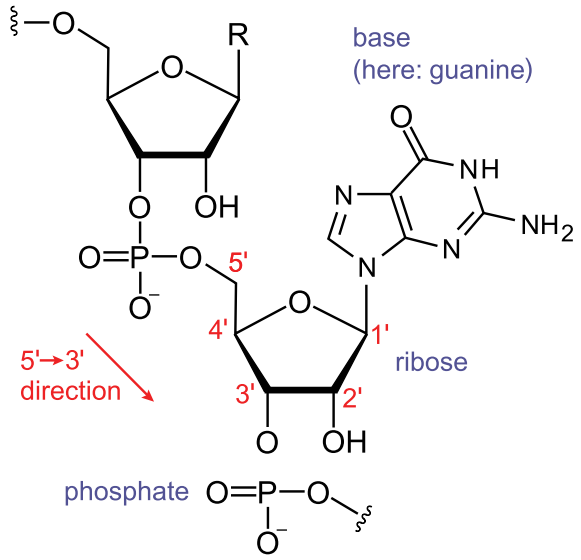

Figure 2: Diagram showing the components of an RNA nucleotide. Note that the carbons of the ribose are numbered 1'-5' (pronounced one-prime to five-prime) Sahib (2014)

such as stem-loops Svoboda and Cara (2006), which form the basis of both tRNA function and pre-miRNA recognition MacRae et al. (2007).

Messenger RNA Messenger RNAs (mRNA) act as an intermediary information carrier between genes, which are stored in DNA, and proteins, which are the main effector molecules in the cell. They are produced when the protein complex RNA polymerase transcribes the sequence of a gene into primary transcript mRNA which is then spliced and processed into mature mRNA. The sequence of bases in the mRNA prescribes the sequence of amino acids in the protein for which it corresponds. The translation from messenger RNA to proteins by the ribosome is dependent on specific recognition, by transfer RNA (tRNA), of a complementary three-mer of bases on the messenger RNA (the codon).

In eukaryotes, mRNA maturation consists of the removal of intronic sequences by the spliceosome complex. The remaining mRNA sections (called exons) are then joined together to form the final sequence. As well as splicing, eukaryotic mRNA undergoes the addition of a 5' terminal 7-methylguanosine called the 5'Cap. This nucleotide is linked to the first residue of the sequence by an unusual 5' to 5' triphosphate linkage, and is part of

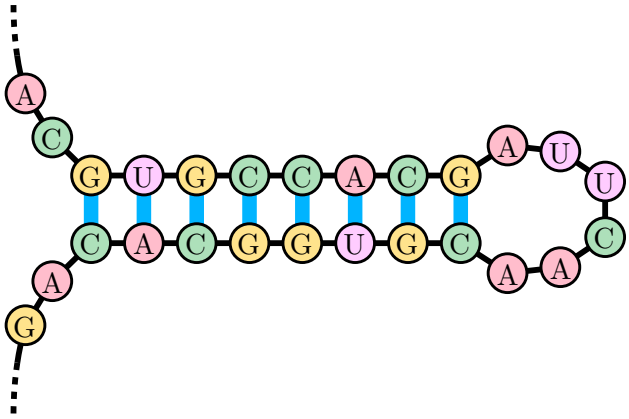

Figure 3: A stem-loop. The stem-loop is an important and simple secondary structure that RNA can form. It is a component of many larger RNA structures and is important in various RNA recognition Sakurambo (2006)

recognition by the ribosome for translation, as well as to prevent 5' RNases degrading the transcript. Maturation also includes polyadenylation of the 3' end of the mRNA. In this process a number of adenine nucleotides are added to the mRNA to prevent degradation, and to act as a signal to export the mRNA from the nucleus. Two cap-binding proteins (CBP20 and CBP80) interact with the transcription/export complex to move mature mRNAs through nuclear pores into the cytoplasm Kierzkowski et al. (2009).

Once in the cytoplasm, the mRNA is recruited to the ribosome by the binding of eukaryote initiation factors to the 5' cap. During the initiation phase, the ribosome is assembled around the mRNA, with the first transfer RNA binding to a 3 nucleic acid site called the start codon. The ribosome translocates the mRNA by three nucleotides (one codon), and the tRNA with an anticodon corresponding to the newly exposed mRNA codon binds to the newly unoccupied A site. Simultaneously, the amino acid attached to the previously resident tRNA is detached from the tRNA and added to the polypeptide chain that will become the translated protein. This process is called elongation. Elongation continues until the mRNA reaches a special sequence called the stop codon. The stop codon binds to a unique protein called a release factor and signals the end of translation, and the ribosome then releases the new protein.

Transfer RNA Transfer RNA (tRNA) is a type of small RNA which holds an amino acid and transfers it to the growing polypeptide during translation. tRNAs are typically between 76 and 90 nucleotides long and have a well conserved secondary structure Sharp et al. (1985). tRNAs begin life transcribed in the nucleus from DNA, much like other RNAs. Some then undergo splicing to remove introns, as well as extensive post-transcriptional modification necessary to perform their mature function, for example, Human tRNA molecules have an average of 13 modifications per cell Pan (2018a). The most abundant tRNA modifications are pseudouridine and 5-methylcytosine. Modifications are known to effect tRNA stability, localization, translation dynamics and ribosome binding. The tRNAs are then exported from the nucleus and covalently linked with their requisite amino acid by their appropriate aminoacyl-tRNA synthetase (an enzyme). Each type of tRNA has a three nucleotide long site called the anticodon, which contains the complementary three nucleotides to the mRNA codon for which it codes. At the beginning of the process of translating a mRNA into a protein, the initiation complex forms starting with the small ribosomal subunit first, which scans until a start sequence is found. The large ribosomal subunits is then recruited. The initiator (Methionine) tRNA completes the complex. The ribosome contains three tRNA sites. The initial Methionine starts in the middle (or P) site. Adjacent to the P site is the A site, which during every step of translating one mRNA triplet to an amino acid, receives a new tRNA matching the mRNA codon. The amino acid associated with this A site tRNA is then bonded, via a peptide bond, to the one or more amino acids attached to the P site tRNA, detaching the existing amino acids from the P site tRNA. Next, the ribosome progresses three bases down mRNA, shifting the now amino acid free P site tRNA to the E (or Exit) site, and shifting the A site tRNA with all of the attached amino acids into the P site. The elongation process then repeats until a "stop" codon is reached. Instead of a tRNA binding to the stop codon, a special protein called a release factor binds to the A site. Rather than add another amino acid, the release factor adds a water molecule to the last amino acid and then separates the amino acid chain from the ribosomal complex. The release factor then causes the disassembly of the ribosome-mRNA complex.

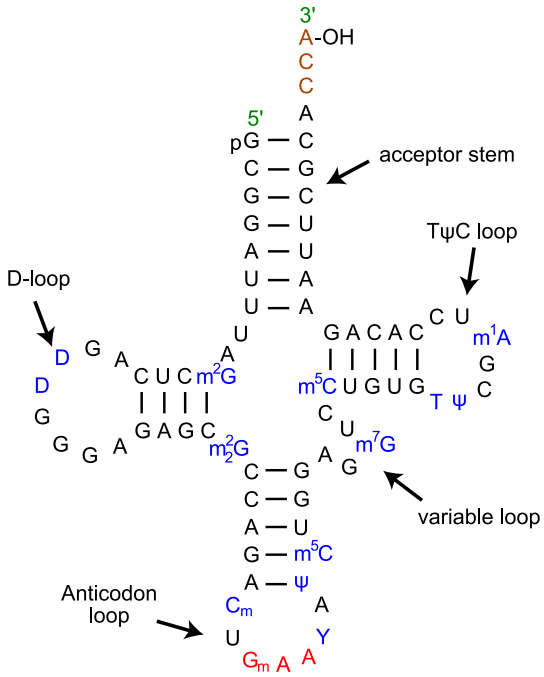

Figure 4: The secondary structure of a typical tRNA. Note that many residues are modified (in blue). tRNA is a heavily modified molecule, and the modifications are necessary for their proper functioning. The three nucleic acid sequence at the bottom in red is the anticodon. It contains nucleotides which are complementary to the codon for which it encodes. Yikrazuul (2010)

MicroRNA MicroRNAs are an important class of short non-coding RNA that down-regulate mRNA expression Lee et al. (1993). Mature miRNAs range in length from 19 to 25 nucleotides Bartel (2004), and are present in a wide variety of taxa across both the plant and animal kingdoms Llave et al. (2002). miRNA undergoes a complicated series of modifications between transcription and maturity . There are two major sources of miRNA. The first is independent genes that only transcribe a primary-miRNA (pri-miRNA), and miRNAs that are located in the introns of other genes. pri-miRNA biogenesis involves several steps. After transcription, the RNA is processed in the nucleus through binding of the Microprocessor complex Gregory et al. (2004). The Microprocessor complex contains Drosha, a RNase III enzyme, and Pasha/DGCR8, a double-stranded RNA-binding domain protein Han et al. (2004). Pasha recognizes the junction between the single-stranded flanking region and the stem of the pri-miRNA and also interacts with the stem and terminal

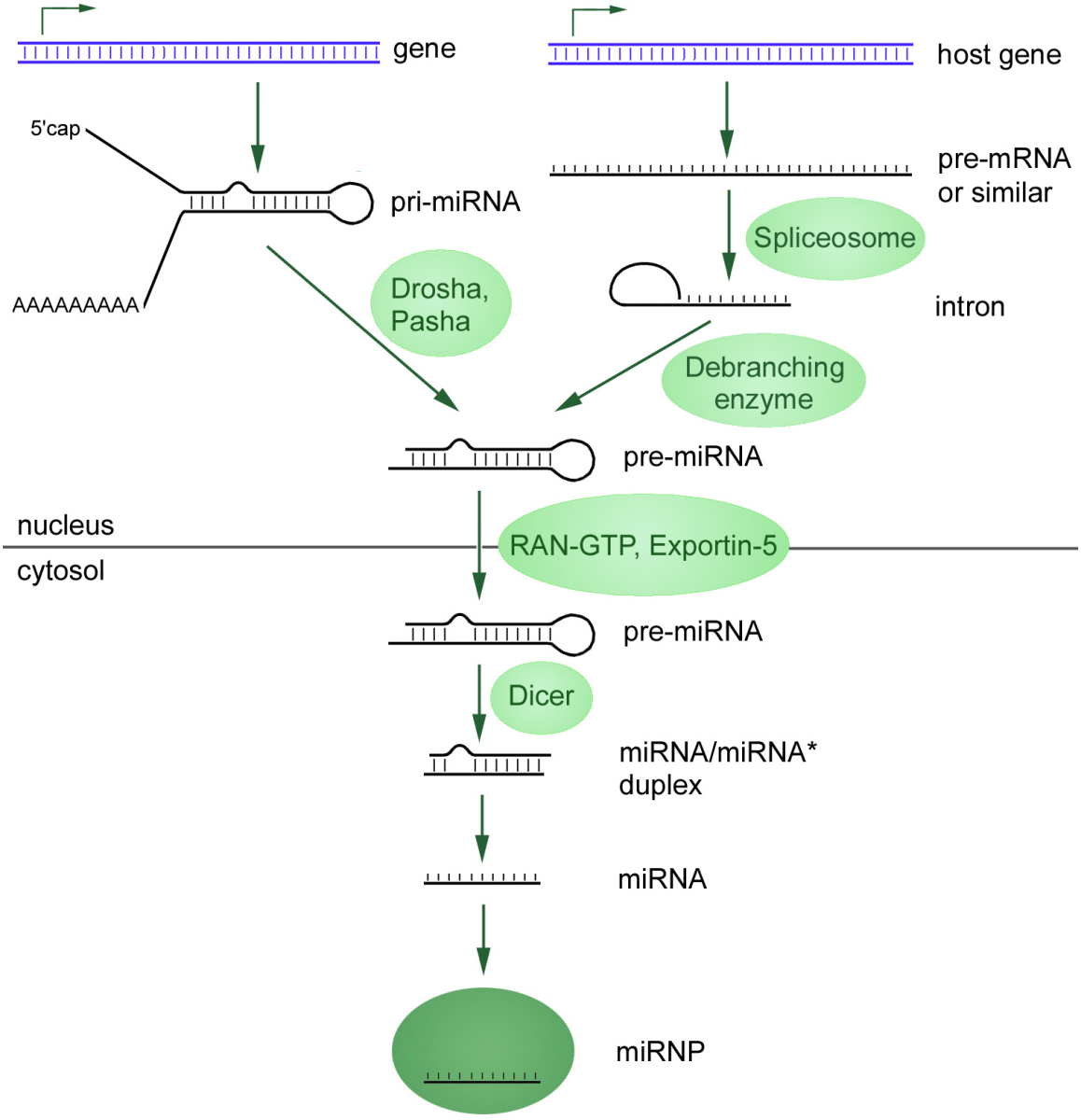

Figure 5: MicroRNA is transcribed from the genome either as an independent pri-miRNA or as part of an intron excised from another mRNA. If it is transcribed independently as pri-miRNA it is processed by the enzymes Drosha and Pasha into a hairpin loop of approximately 65 bases. If it is formed from an intron it is processed by a debranching enzyme. The result of both of these actions is called a pre-miRNA. The pre-miRNA is then exported from the nucleus by RAN-GTP and Exportin-5. The loop portion of the hairpin is cleaved by Dicer. The two remaining strands dissociate, one forming the mature miRNA and the other forming the miRNA*. The mature miRNA then is loaded into the RNA induced silencing complex where it prevents translation of mRNA complimentary (in whole or in part) to it. Narayanese (2012)

loop. Pasha positions Drosha at the proper location for cleavage. Drosha then cleaves the pri-miRNA producing two flanking fragments, and the ˜65nt stem loop region that is called the pre-miRNA. Pre-miRNAs, are then exported from the nucleus by a RAN-GTP dependent exportin called exportin-5. Once in the cytoplasm, pre-miRNAs are cleaved by Dicer, a RNase III type protein. This cleavage removes the loop from the pre-miRNA and produces a duplex of 22 nt miRNA. Of the two strands that were created by Dicer cleavage, the strand with the more stable base-pairing at its 5' end, called the miRNA*, is ejected and then degraded in most circumstances. The mature complex, containing TRBP, Dicer, the single strand of miRNA, and Argonaute, is called the RNA induced silencing complex (RISC). Once RISC is assembled, the miRNA localizes the complex to the target mRNA.

In healthy organisms miRNAs are responsible for regulation of cell fate, changes in organism life-cycle, and temporal regulation of development. The importance of miRNAs for normal development has been shown by the fact that knocking out Dicer is embryonic lethal at an early stage, causing the depletion of pluripotent stem cells. miRNAs also play a significant role in disease. In cancer they can act as either oncogenes or tumor suppressors, dependent upon what mRNA they target Shenouda and Alahari (2009).

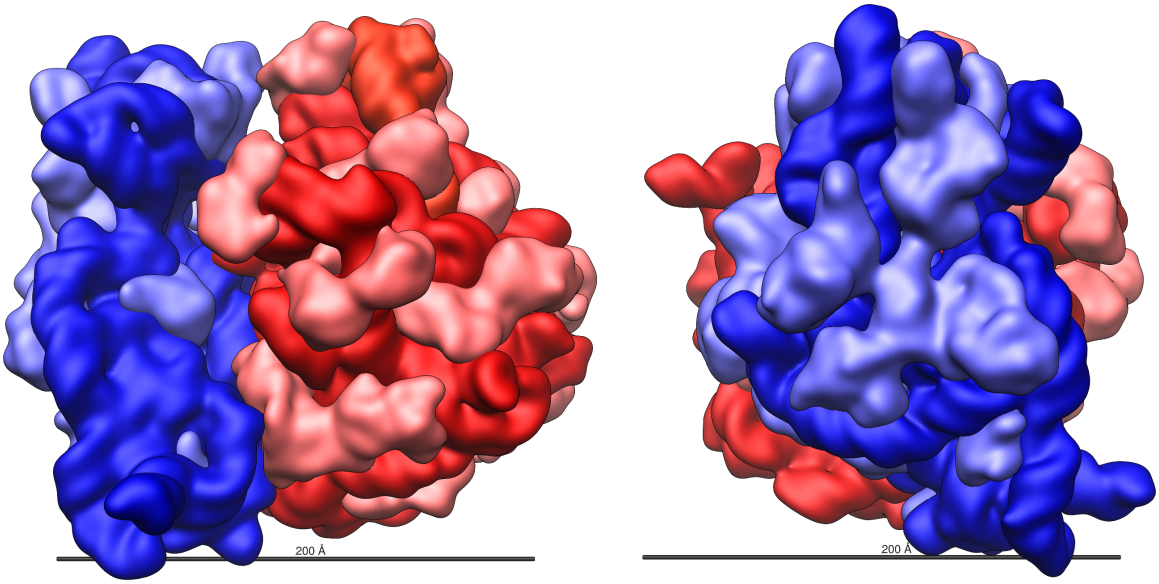

Figure 6: The structure of the E. coli ribosome. The rRNA components are in darker red and blue, with the protein components in lighter red and blue Vossman (2009)

Ribosomal RNA The ribosome is a key part of the protein translation process. It consists partially of proteins, as with most other enzymes; however, belying its ancient origin, it also consists of structural RNA components. These components, known as ribosomal RNA (or rRNA) have been known for the last 50 years to be highly modified Decatur and Fournier (2002). The ribosome consists of two subunits identified by their size (in Svedbergs). In eukaryotes the large subunit is 60S and the small subunit is 40S Noller (1984). In prokaryotes the large subunit is 50S and the small subunit is 30S.

The 50S subunit includes two RNAs, a 5S (˜120 nt in E. coli) and a 23S (˜2904 nt in E. coli) and 34 proteins Green and Noller (1996). The important nature of modifications to these rRNAs is apparent in the difficulty that the 23S rRNA has assembling into a functional 50S ribosome *in vitro*. In the absence of RNA modifying enzymes, 23S has great difficultly reconstituting a catalytically active 50S subunit. Indeed, it suffers a greater than five orders of magnitude decrease in this formation Green and Noller (1996). High resolution X-ray crystallography shows that modifications to rRNAs occur in the regions which are most closely linked to the proper functioning of the rRNA Chow et al. (2007).

1.3. Mass spectrometry

What is mass spectrometry Mass spectrometry is a complex of analytical techniques that measures the mass-to-charge ratio of compounds in a sample El-Aneed et al. (2009). By ionizing the analytes of interest, mass spectrometrists are able to use electromagnetic fields to separate, contain, and measure their mass-to-charge ratio. Since any sample of natural origin contains a small percentage of heavy isotopes of various atoms, the change in mass (as well as the percentage of atoms which are of a heavier isotope Yergey (1983)) is well known, and since the change in mass is necessarily an integer, the change in mass between the closest peaks in so-called isotopic clusters can then be used to calculate out the mass of any particles of interest. Mass spectrometry takes advantage of properties inherent to all forms of matter, because of this it is a widely applicable analytical technique, with uses spanning biology, chemistry and physics.

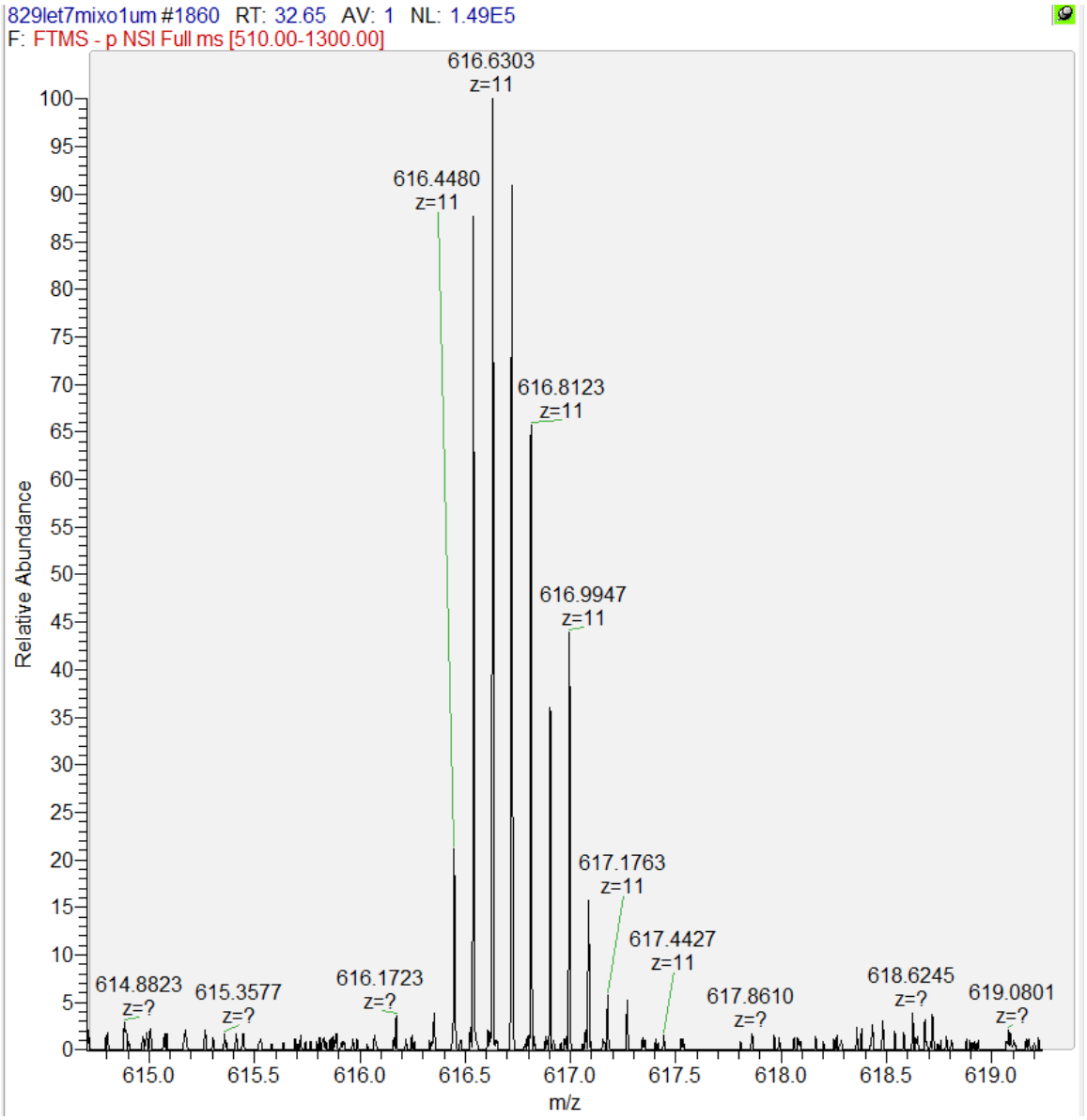

Figure 7: An example mass spectrum. The image is dominated by a set of isotopic peaks corresponding to a charge 11 microRNA. The x axis is the mass to charge ratio (M/Z). The y axis is the relative abundance of each mass peak, with the most abundant scaled to 100. Since the distance between the closest peaks in the isotopic set is 1/11 m/z we can calculate the actual mass of the analyte as 6791.9 Daltons.

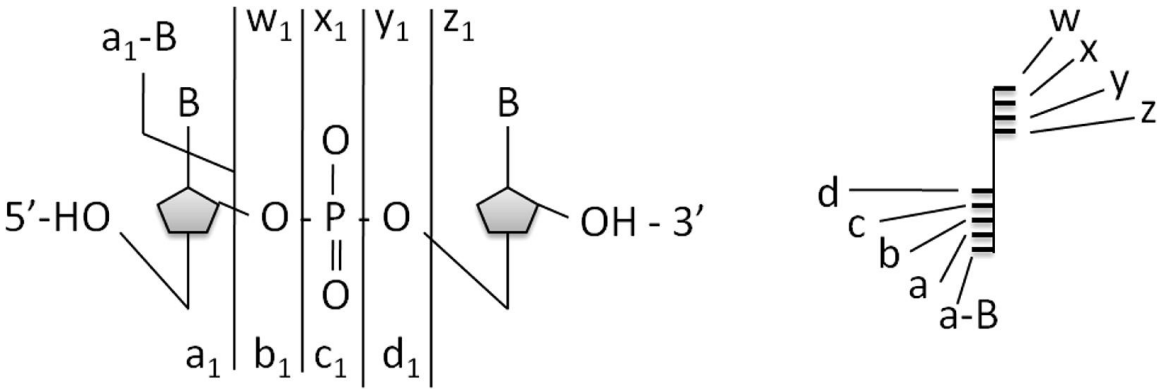

Figure 8: A schematic representation of the different locations at which RNA fragments under higher-energy collision dissociation. The vertical bars in the left figure show where the cleavage happens, and are annotated with a letter and a number in subscript denoting the convention for annotating fragments. The letter represents the fragment type while the number represents the number of nucleic acids remaining in the fragment. Additionally a-B ions are formed with the cleavage of the base from the ribose on the 5' side of the cleavage. The figure on the right shows different fragment ion types are represented in so-called "fork plots".

What is tandem mass spectrometry? In addition to being able to measure the mass of intact analytes, a form of mass spectrometry called tandem mass spectrometry (or MS/MS or MS^2) can be used to determine the mass of fragments of the analytes of interest. In tandem mass spectrometry, a range of ions are selected by their mass to charge ratio and are isolated using electromagnetic containment. This population of ions is then exposed to some mechanism for breaking it down. While there are a number of different mechanisms available on different instruments, the two types I used are higher-energy collision dissociation (HCD) and collision induced dissociation (CID). HCD and CID both function by increasing the kinetic energy of trapped ions using an electrical potential and then introducing a neutral gas. The fast-moving ions collide with the slow moving neutral gas, converting some portion of their kinetic energy into internal energy and breaking molecular bonds. By setting the correct amount of energy to use in fragmenting the sample, the experimenter is left with a population of fragments, ideally encompassing the total range of theoretical products. This technique is useful on biopolymers, which are composed of a sequence of monomers of known mass and have fragmentation patterns which are consistent and replicable. With

a complete population of the potential fragments, we can then reconstruct a "ladder" of peaks, each separated by the mass of the monomer present in one and absent in the other. Through this method we can establish the complete sequence of the polymer. An example of how this works for peptides is seen in the bottom right panel of figure 10. The arrows labelled A, G, and L show the mass shift between successive amino acids.

In this work I will focus on the ability of mass spectrometry to measure biopolymers, which has been traditionally applied to analyzing proteins, but which I will show is equally well suited to analyzing RNA.

Proteomics, where we came from Proteomics is the study of proteins at a large scale Bantscheff et al. (2012). It is a mature and well developed field with good software support for mass spectrometry. Our lab's expertise in proteomics is what has allowed for my rapid entry into the field of nucleic acid mass spectrometry. In a standard proteomics pipeline the workflow progresses from extracting proteins from the cell, to purifying them, and then separating them along a chromatographic gradient. The liquid is then pushed out of an emitter tip, while a voltage is injected into the liquid, creating charged droplets and ionizing the peptides. Ionized peptides enter the mass spectrometer where they are detected by the mass analyzer. Peaks of interest are then selected and fragmented to generate tandem mass spectra. The spectra are recorded, and sent to the attached computer. Software is then used to compare the experimental tandem mass spectra to theoretical mass spectra in a library of known peptide fragments. The existence of software to automate hypothesis testing is necessary to allow experiments which look at all known proteins for a species. Even with current software there are still limitations in the scale of experiments which can be done. Specifically, each addition of a potential modification increases the search space, i.e. the potential number of solutions that need to be iterated through, exponentially. These constraints make developing efficient, fast software both very necessary and very difficult to do. Thankfully at this time most major mass spectrometer manufacturers have developed solutions to handle at least modest sized proteomics data sets.

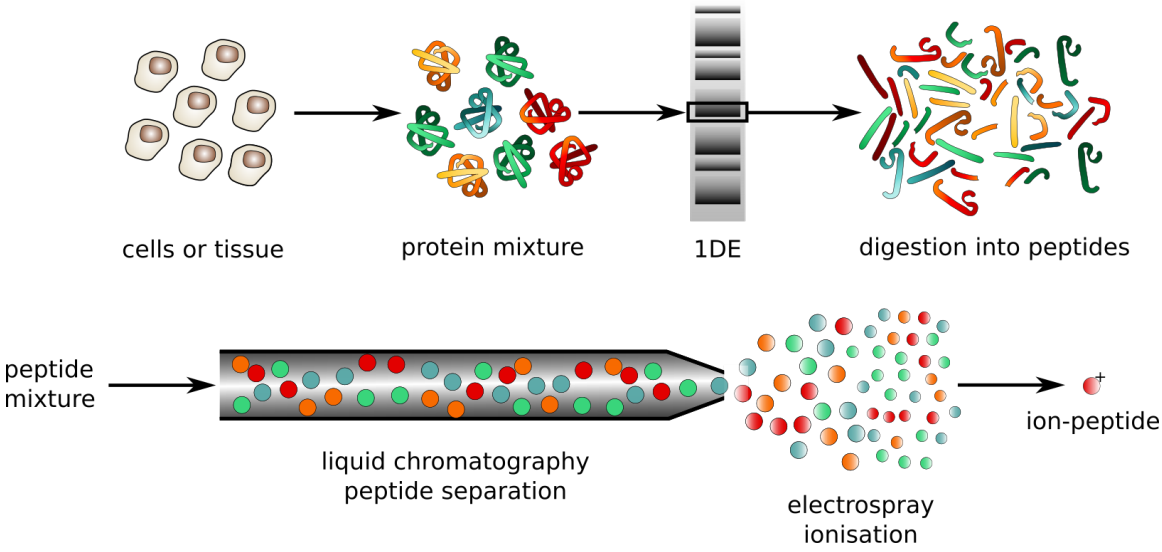

Figure 9: An traditional example of a proteomics protocol. The general workflow is the same as the one we use for analyzing RNA. The cells or tissues of interest are harvested, and the proteins are separated out from the rest of the cellular components. If the experimenter is curious about subcellular localization of proteins, further subdivision by cellular compartment follows. The protein mixture optionally is injected into a gel and undergoes gel electrophoresis to separate out proteins of different mass. Gel electrophoresis can be skipped if the experimenter in interested in multiple different proteins in the sample. The resulting selected proteins are digested by a protease enzyme into short oligomers called peptides. This digestion process makes both ionization and later identification of compounds easier. The peptide mixture is separated by chromatography, allowing peptides of different mass to reach the mass spectrometer at different times, and ultimately making identification simpler. At the end of the chromatography column the separated peptides are ionized via electrospray ionization. To accomplish this an electrical current is run through the liquid containing the peptides, at the same time that peptides are forced out of a small spray tip at high pressure. The combination of electrical repulsion from peptides of the same charge and heat from the inlet of the mass spectrometer causes the peptides to form small drops. Hupé (2012)

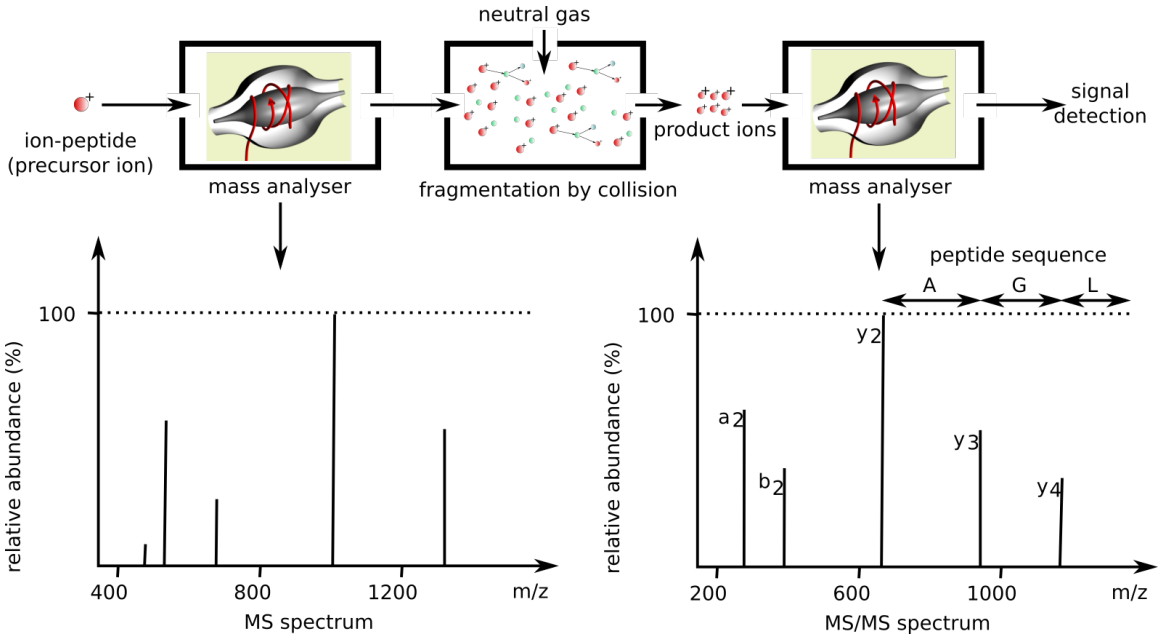

Figure 10: These ionized peptides enter the mass analyzer and their mass to charge is measured. Based on a set of predefined parameters masses of futher interest are selected and collected for fragmentation. The fragments, called product ions, are injected into the mass analyzer and their spectra are collected for analysis. The difference between successive fragment masses can be used to establish the sequence of the peptide.

Chromatography: gotta keep 'em separated Samples of biological origin inevitably contain a wide variety of different compounds. Separating these into manageable fractions is an important step in any workflow involving such samples. Figure 9 demonstrates this. Proteins are first separated through gel electrophoresis and then peptides are separated by high performance liquid chromatography (HPLC).

HPLC turns out to be an incredibly useful tool for separating all sorts of different biological molecules. It consists of a column containing an adsorptive solid stationary phase and a liquid mobile phase that is a mixture of solvents which changes during the experiment. The sample is loaded onto the column in a mixture called the loading buffer. The composition of the loading buffer is such that the sample adsorbs to the particles of the stationary phase, and will stick there no matter how much loading buffer is passed over them.

What happens next is dependent on what form of HPLC is being conducted. For the

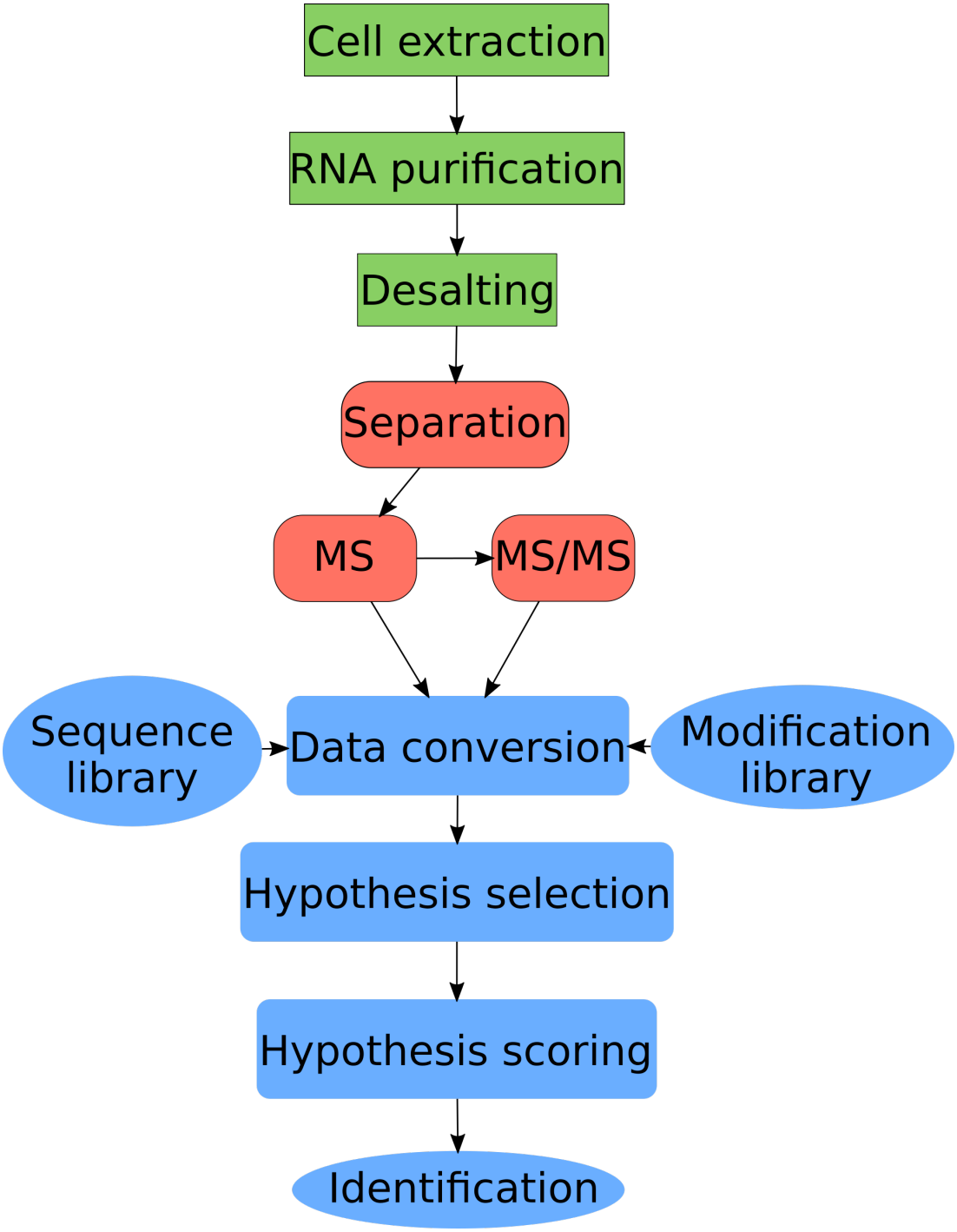

Figure 11: A flowchart of the components of this project. Nodes are colored based on a general grouping of processes. Green are *ex vivo* (out of cells), red are analytical, and blue are computational.

purposes of this work, we will discuss what is called reverse-phase HPLC. In the standard reverse-phase HPLC used in proteomics, the sample is loaded in an aqueous buffer onto a stationary phase which is composed of a hydrophobic material. The hydrophobic peptides bind strongly to the hydrophobic stationary phase. A mixture of solvents is then run over the column at high pressure (>50 bar). The mobile phase starts out being predominantly aqueous, and through the experiment the mixture transitions to containing more and more organic components (acetonitrile is a common example hydrophobic solvent used in proteomics). Depending on the chemical properties of the peptides bound to the stationary phase, they will have a different affinity for the aqueous and organic phases. Because of this they will be pushed off the stationary phase at different times during the experiment i.e. at different proportions of aqueous versus organic phases. This separation by chemical property is the core purpose of chromatography and allows the scientist to capture mass spectra of different compounds at different times.

The chemistry of nucleic acids makes separating them by HPLC more challenging than for peptides. Chemically, nucleic acids contain both hydrophilic and hydrophobic moieties (i.e. parts), making the process of forcing them to adduct to a hydrophobic stationary phase difficult. Traditional experiments that do not involve mass spectrometry typically use cation-exchange chromatography Junowicz and H. Spencer (1969). Unfortunately, the salts necessary to perform cation exchange are not compatible with mass spectrometry. The good news is that there is a solution. The experimenter can add ion-pair reagents to act as "adapters" and allow nucleic acids to bind to standard reverse-phase chromatography columns Lin et al. (2007). Ion-pair reagents contain both chemical elements which are hydrophobic and which are hydrophilic. The hydrophobic portion interacts with the stationary phase leaving the hydrophilic portion to interact with the hydrophobic phosphate groups on the nucleic acid. In this manner it is possible to bind and separate different nucleic acids by reverse-phase chromatography.

Unfortunately, adding ion-pair reagents has certain downsides. First, they are very difficult

to remove from equipment (the HPLC, MS, and any tubing connecting them) once introduced. For any porous or semi-porous surface they are functionally impossible to entirely remove. Second, they suppress positive ionization in the mass spectrometer for as long as they remain there, making sharing an instrument with a colleague who does ion-pair chromatography a trying ordeal at best. Lastly, even in the negative charge ionization mode that we use for nucleic acid mass spectrometry, the ion-pair reagents partially inhibit the electrospray ionization of the nucleic acids Gustavsson et al. (2001). This results in issues keeping a stable spray from the electrospray tip as well as difficulties generating enough signal in the mass spectrometer. Nonetheless (despite quite substantial testing) ion-pair reverse phase HPLC remains the best technology for separating nucleic acids in a way that is compatible with mass spectrometry, and it remains the method that I use throughout this work.

Why do we care about RNA modifications? RNA is an extensively modified biological macromolecule. Over 150 chemically distinct modifications have been reported, ranging from simple methylation of the ribose or nucleobase to large additions such as the conversion of guanosine to wybutosine. The presence of methylated adenine, cytosine, and guanine in RNA was uncovered in the 1960s Borek and Srinivasan (1966), and pseudouridine has been referred to as the fifth base for decades Davis and Allen (1957). However, widespread interest in these epigenetic marks (primarily N6-methylated adenosine, m6A) has been raised by recent reports that underscore their importance in a wide variety of developmental signalling. For example in stem cells, the intracellular effector proteins SMAD2 and SMAD3 promote binding of the m6A writer complex to a subset of mRNAs associated with early cell fate decisions Bertero et al. (2018). Likewise, a number of modifications are associated with disease. It has been demonstrated that the loss of taurine modification in the anticodon of mitochondrial tRNA-Leu is responsible for mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS) Kirino et al. (2004). m6A is implicated in obesity, as a target of FTO (obesity-associated protein, an m6A demethlylase) Jia et al. (2013). The presence of m6A is also associated with increased injury-induced protein translation in

adult mouse dorsal root ganglia. Loss of either m6A writer or reader results in defects in functional axon regeneration Weng et al. (2018). Aberrant methylation of cytosine-5 (m5C) in tRNAs has been linked to neuro-developmental disorders Abedini et al. (2018). The proportion of the RNA which has been modified is also of interest. Frequently in tRNA's for example modifications may be present at a basal level but be expressed at an increased level to increase their ribosomal affinity Pan (2018a). Being able to measure both the presence and stoichiometry is thus important.

The recent interest in RNA epigenetics has been spurred by technical advances in next-generation sequencing technology, which has allowed modifications in mRNA to be profiled individually. All of the approaches based on next generation sequencing, such as, Solexa/Illumina sequencing, use antibodies to immunoprecipitate modified RNA and/or apply chemical treatments to alter it and read out modifications as mutations or truncations in the preparation of cDNA Li et al. (2016) Helm and Motorin (2017). The primary caveat of these methods is that only a single type of modification can be profiled in each experiment, and specific chemical and/or antibody reagents do not exist for every modification. Further complications can be caused by lack of specificity of the existing antibodies, in particular m6A and m6Am Linder et al. (2015). Antibody binding is, at best, a rather inexact science. Cross reactivity abounds, and due to the stochastic nature of antibody generation, there are many antibodies that work well for a given technique (e.g. western blotting) but fail for other techniques (e.g. immunoprecipitation).

Tandem mass spectrometry is currently the only technique that can directly and comprehensively characterise chemical modifications in RNA sequences, by comparison of mass spectra with a sequence spectral database Kullolli et al. (2014). Prior to the advent of massively parallel (or next-generation) sequencing, there was substantial interest in using mass spectrometry to sequence nucleic acids Apffel et al. (1997). However, these efforts largely ceased as it became apparent that there were other sequencing methods available that had greater throughput. As more and more of the biological relevance of modifications

to nucleic acids has emerged over the last few years, there has been renewed interest in using mass spectrometry for characterising modification. The majority of this work has focused on reducing the RNA to mono-nucleosides and applying workflows analogous to metabolite analysis Su et al. (2014). While these techniques are effective in determining what modifications are present in a sample, all information about the location and co-occurrence of modifications is lost. This information is critical in complex samples to allow attributing modifications to specific RNAs. Even in simpler cases, modification location and co-occurrence may be important for a phenotypic effect; for example, in microRNA, 2'-O-methylation of the 3'-most nucleic acid sterically inhibits 3' exonuclease digestion (i.e. prevents enzymatic breakdown of miRNA by 3' exonuclease enzymes)Abe et al. (2014). For this reason there is interest in analysing samples in as close to their native states as possible. However, intact oligonucleotides are challenging to separate via chromatography that is compatible with mass spectrometry. The current approach of choice is reversed-phase ion-pair liquid chromatography Huber and Oberacher (2001).

In addition to the experimental challenges, difficulties emerge in interpreting the acquired data. Impressive steps towards automating data analysis have been made by several tools, including SOS in 2002 Rozenski and McCloskey (2002), Ariadne in 2009 Nakayama et al. (2009), Oma and Opa in 2012 Nyakas et al. (2012), and RNAModMapper in 2017 Yu et al. (2017), all of which are database-matching scripts or programs that decode the complicated patterns of oligonucleotide fragmentation. However, none of these existing software solutions currently offers key features necessary to analyze data from large-scale experiments. First, no software can efficiently handle RNA oligonucleotide spectral searches – especially of more complex samples or involving many different modifications – in batch-compatible fashion. Second, statistical validation strategies such as false-discovery rate estimation are not implemented. This leads to unreliable sequence assignments and subjective manual assessment of spectra for validation. Third, existing solutions do not tie into any larger analytical framework, making integration with other (e.g. quantitative) data difficult. In contrast, shotgun proteomics has been sequencing peptides reliably for many years, and the

inference, identification and quantification of proteins from constituent peptides has been automated to such a degree that the field has matured into answering biological questions at a more fundamental level Gillet et al. (2016).

1.4. A roadmap

For this work, I used the standard proteomics experiment as a template, and adapted the techniques for use with RNA. The initial target of my research was specifically focused in looking at miRNA modifications. miRNAs are an ideal length for this work. They are long enough to be unique, but short enough to be able to be effectively ionized in their entirety. As the project has progressed, it has become apparent that both the analytical method and the software are applicable to longer types of RNA as well. The development of this project consisted of implementing a protocol to extract RNA from cells, as well as a chromatography and analytical solution, and developing software to analyze the data. I developed two generations of software during the period of this research. The results in Chapter 2 were produced with the first pass software, and the results in chapter 3 come from the final product. I describe the analytical portion of this project in chapter 2 and describe the software portion of the project in chapter 3. Since creating meaningful data from a sample of any complexity requires both the analytical and computational platform to be worked out, the reader is advised that there is some crosstalk between the two chapters. Chapter 2 also contains a brief discussion of the creation of a sheath spray assist device for stabilizing the electrospray for nucleic acids. While this device did not end up being used in the final iteration of the analytical method, it still is an important development for which broader applicability is definitely possible. I continue to work on improving it in the hopes of it being used in future experiments elsewhere. Chapter 3 will also discuss some of the experiments that have been conducted to show both the proper functioning of my method and to show its broad applicability to a variety of samples. It also contains a brief description of work to use OpenMS' label-free quantification techniques to assess the stoichiometry of modifications. Chapter 4 concludes with a summary of where the field is

after all of this work, and future directions in which the field can progress.

CHAPTER 2 : Analytical Methods and Analytical Challenges

2.1. Introduction

The first step toward being able to analyze mass spectrometry data is to generate mass spectrometry data. Prior to developing any software, I needed to develop an analytical system capable of producing data to analyze. In this chapter I discuss the trials and tribulations involved in creating the analytical system, why nucleic acids are particularly difficult to consistently analyze by mass spectrometry, and some of the ways to make them easier to manage.

The analytical system that I developed consists of several parts. The first, extracting total RNA from cells, was followed by separating RNA into fractions by length. Purification was needed to further clean up the sample and reduce salt content (I discuss the challenges associated with this step later in this chapter). The complex mixture of RNA then needed to be separated in such a way as to limit the number of RNA species in single MS1 (i.e. parent, or full) scans, and the subsequent potential for missing triggering MS2 (i.e. daughter or dependent) scans in a timely fashion. After the separation, the RNA needed to be ionized in order to be visible in the mass spectrometer. I will start by discussing the chromatographic separation, the ionization, and the mass spectrometry. I will then move on to looking at my early experiments to determine the sensitivity of my system, and its ability to separate mixtures containing multiple different RNAs.

The analytical system that I used to separate different RNA went through several phases. Initial attempts to directly inject samples containing only a single RNA using the Triversa Nanomate device were unsuccessful. I then moved to using ion pair reverse phase chromatography, both to separate different species of RNA in the same sample and to separate RNA from non-RNA contaminants in the same sample. A substantial amount of time was spent adjusting the buffer composition to produce consistent results. In order to offset the disadvantages inherent in needing to use ion-pair reagents, I developed an axial sheath

spray assist device to improve spray stability. While the addition of this device substantially improved my results, it ended up being complicated to run. Thus I ultimately moved to a system based off of Thermo Fisher EASY-Spray columns.

The reasoning for choosing miRNA (analogues) for the initial analytical work was twofold. First, the length (and therefore mass) of miRNA is constrained to a length which is within the mass range of the instruments that I use. That is, I do not need to do any further digestion to create fragments of an analyzable length. The second reason is more complex; recent work has shown that miRNAs can contain a variety of modifications that affect their function. These include base additions like uridylation, as well as modifications to existing bases and sugars. A particularly well characterized modification is the addition of a methyl group to the 2' oxygen on the ribose of the 3' nucleotide of the miRNA Abe et al. (2014). In both Drosophila and Arabidopsis it has been shown that this addition changes the shape of the 3' end of the miRNA, slowing the activity of 3' exonucleases, therefore increasing the lifespan the molecule Li et al. (2005). Oxidation by reactive oxygen species is another modification that has been linked to increasing the inhibitory activity of certain miRNA species in cardiac ischemia Wang et al. (2015).

2.2. Experiments

2.2.1. Chemicals

Chemicals used for HPLC were: Triethylamine (TEA), Hexafluoro-2-propanol (HFIP), methanol, HPLC grade water, Ethylenediaminetetraacetic acid (EDTA), HPLC grade acetonitrile. Synthetic miRNAs were ordered from Integrated DNA Technologies (IDT). These included dme-miR-34 (UGGCAGUGUGGUUAGCUGGUUGUG), dme-let7 (UGAGGUAGUAGGUUGUAUAGU), and dme-BANTAM (UGAGAUCAUUUUGAAAGCUGAUU). Variants of miR-34 and let7 were also ordered with a 2'-O-methylation on the 3' terminal nucleic acid. MiR-34 was additionally ordered with a 2'-O-methylation on the 5' terminal nucleic acid. All synthetic miRNA samples were HPLC purified by IDT. I chose these three RNA

analogues as they correspond to miRNA that have been well studied. A schematic of the liquid flow in my instrumentation is provided in figure 12.

2.2.2. Direct infusion with the NanoMate

The Triversa NanoMate is a chip-based electrospray ionization device which allows direct infusion of very small quantities of sample through a chip-based nozzle and into the mass spectrometer. It is an ideal piece of equipment for analyzing relatively pure simple samples, and is faster and requires less setup than chromatography. It also compares favorably to traditional direct infusion from a syringe in that it allows the use of substantially less sample and substantially lower flow rates (which correspond to higher ionization efficency and minimizing reagent usage). Since nucleic acid mass spectrometry requires ionizing in negative mode, I had to deal with the associated challenges of producing a stable spray without arcing over the very small distance between the chip and the ion capillary. In my initial experiments I varied the composition of the liquid in which I suspended the nucleic acid oligomers from 100% water to 100% acetonitrile. Unfortunately, what I found was that as the acetonitrile concentration was increased the stability of the spray improved, but the suspension of the nucleic acid became more tenuous. This resulted in a series of experiments which showed that I could either get spray for a very short period of time followed by the nucleic acid precipitating and clogging the spray nozzle, or I could have a higher percentage of water, preventing precipitation but also preventing a spray that was stable enough to collect data from. Given these competing constraints, I determined that the NanoMate was not an ideal solution for oligonucleotide mass spectrometry.

2.2.3. nanoLC-MS/MS

The HPLC system that I used consisted of an Eksigent AS2 autosampler and an Eksigent NanoLC-Ultra 2D+. The preliminary experiments were performed on a Thermo LTQ, with the majority collected on a Thermo Velos Orbitrap-LTQ. On the Orbitrap instrument, both MS1 and MS2 spectra were collected using the Orbitrap, a high resolution mass detector,

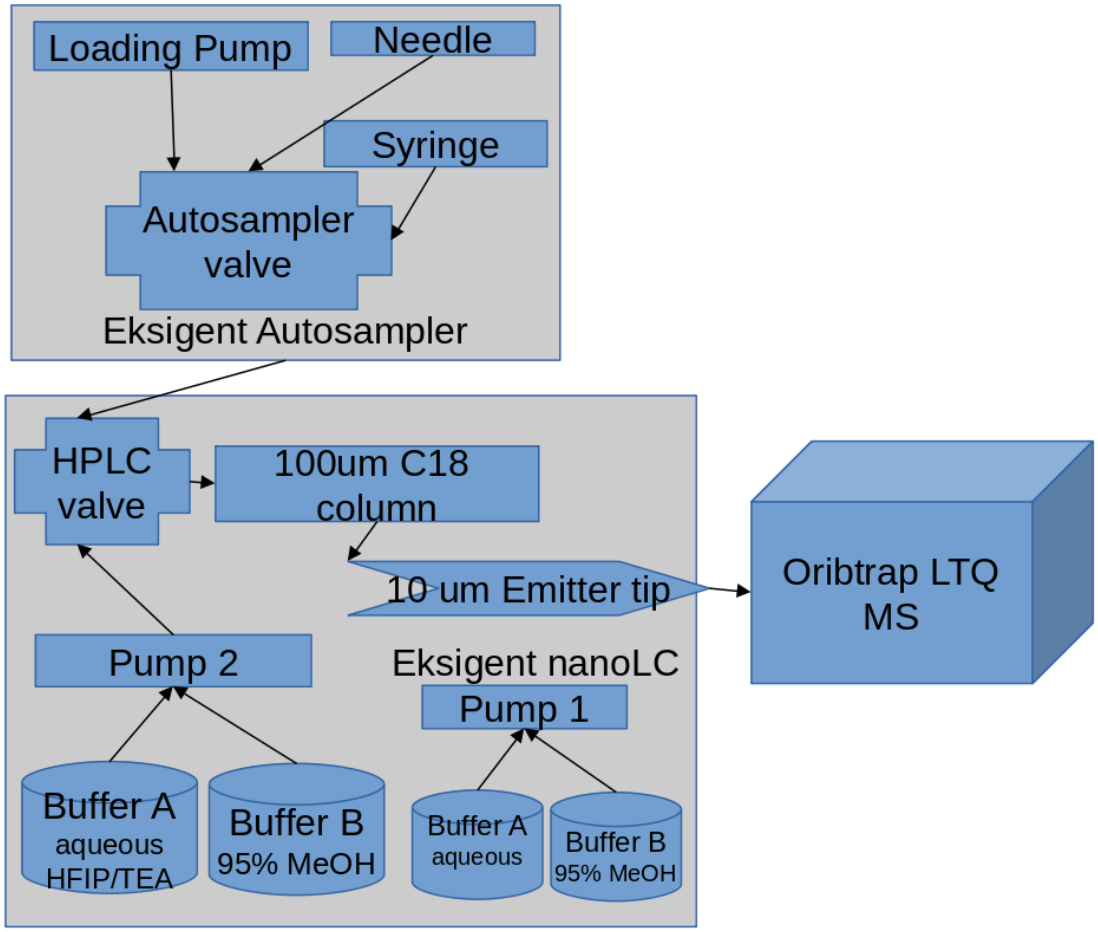

Figure 12: A schematic of the instrumentation used in the nano-LC system. The sample is loaded from vials in the autosampler by the syringe. The autosampler valve then switches, putting the loading pump in line with the loaded sample. The loading buffer is pushed from the loading pump through the HPLC valve and into the HPLC column. This flow pushes the sample onto the column where it sticks. Once loading is complete, the loading pump shuts off and the HPLC valve rotates, putting pump 2 in line with the column. Simultaneously the Orbitrap energizes the emitter tip, starting electrospray and starting to acquire spectra. Pump 2 begins running mostly buffer A (the aqueous buffer), and as the experiment progresses the amount of buffer B being pumped increases, causing the nucleic acids on the column to elute off when the percentage of organic solvent reverses their adduction to the column. Note that this diagram describes a one column setup. For some experiments I also added a trap column before the main column allowing sample to be loaded to the trap column with a much higher flow rate than the main column could tolerate.

with resolution of 30,000 or 60,000. Fragmentation was performed by collision induced dissociation (CID) with a normalized collision energy of 35, an isolation window of 1 m/z (mass to charge) and isolation time of 10 ms. Synthetic miRNAs were diluted to 1uM, 100nM, 10nM, and 1nM by the addition of HPLC grade water containing 1mM ethylenediaminetetraacetic acid (EDTA). The addition of low concentration EDTA was observed to both increase column longevity as well as decrease salt adduct formation. Samples were loaded from the autosampler using an aqueous loading buffer containing 100 mM HFIP, and 1.7mM TEA adjusted to pH 7.5. Samples were loaded onto a 5cm X 100 μm trap column containing 3μm C18-aq resin. Loading to the trap was accomplished using the same 100 mM HFIP and 1.7mM TEA buffer that was used in the autosampler, at a rate of 3uL/min for a duration of 10 minutes. The longer loading period allowed for better removal of the EDTA and any attached salts. Following loading, the sample was eluted into the MS on a 60 minute gradient with buffer A being the same as the loading buffer, and buffer B consisting of 90% methanol and 10% acetonitrile. The analytical column used was a 75μm ID x 15cm C18-aq column packed in-house, attached to a 10μm electrospray tip (also packed with C18-aq, to minimize bubble formation from pressure drop after the column). Analysis was performed over a 60 minute gradient at 250nL/minute with the following profile: 0-12 minutes: 95.2% A, 12-15 minutes: 95.2% to 90% A, 15-42 minutes: 90% to 64.8% A, 42 to 43.5 64.8% to 30% A, 43.5-52.5 minutes 30% A, 52.5-54 minutes 30% to 95.2% A, 54-60 minutes 95.2% A. ESI voltage was set at -1.8kV and the capillary temperature was set at 180C°. I performed these experiments in negative mode MS, as the phosphate backbone is not amenable to ionization in positive mode. To separate different oligonucleotides, I used ion-pair reverse-phase liquid chromatography. In order to achieve a level of sensitivity that will allow detection of miRNAs from cell samples, I use nanoflow conditions (250nL/min). Performing chromatography at a lower flow rate increases the proportion of ions which are analyzed by the MS, and improves sensitivity compared to a higher flow rate. The system was flushed out using a wash-cycle between sessions and after every 6 runs. In most runs, MS1 profile data was recorded in the range of 520-1200 m/z (effectively filtering out

the lower m/z background produced by HFIP). I also acquired data spanning 520-850 m/z and 520-1800 m/z in my preliminary experiments. Spectra were recorded using Thermo's acquisition software Xcalibur.

2.3. Results and Discussion

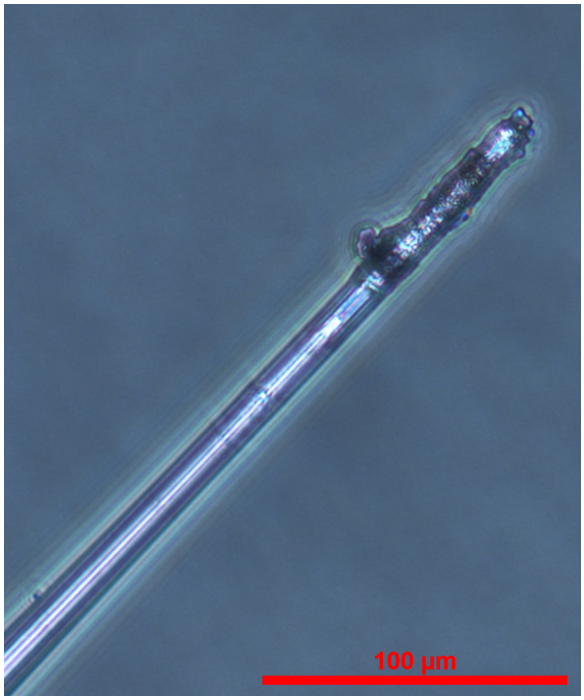

Figure 13: A micrograph of an electrospray tip which has been coated in salt.

Handling salt Salt is a major issue for mass spectrometry. It inhibits ionization and has a nasty tendency to build up at the spray tip. This issue is much more pronounced with nano-flow chromatography due to an increase in the surface area to volume ratio of the spray tip to the flow. As the spray tip radius decreases, the flow cross section decreases at a quadratic rate while the inner circumference of the tip decreases at a linear rate. This leads to a larger portion of the flowing liquid directly contacting the silica that makes up the tip. In my experiments, this resulted in salt crystal formation at the spray tip disrupting proper spray cone formation and ultimately blocking the system entirely. In early work I simply switched tips regularly, however the stochastic nature of this type of salt failure required an operator constantly keeping an eye on the instrument throughout the run. I found that

despite the claims from the synthetic oligonucleotide manufacturer that all products were HPLC purified, there was still substantial salt remaining in their nominally pure samples. To counteract this salt contamination, I used STop And Go Extraction (STAGE) tips to flush out as much of the salt as possible Yu et al. (2014).

RP-Ion-pair HPLC separates different miRNA sequences at the nanoflow scale. The majority of previous work on RNA MS has been performed using micro-flow chromatography. Using a nano-flow system allows for better sensitivity and smaller sample requirements. In this work, I demonstrate the capability of nano-flow MS/MS to accurately identify oligonucleotides. I am able to separate different oligonucleotides to allow individual identification of oligonucleotide species using nano-flow ion-pair reverse phase chromatography. Using my nanoLC-MS/MS platform, I was easily able to detect a clear chromatographic peak for the synthetic oligonucleotide Let-7, at low nanomolar concentrations (figure 14a). Mass spectra obtained on a Velos Pro Orbitrap instrument of the oligonucleotide showed a large variety of charge states (figure 14b). However, high resolution MS acquisition did demonstrate the presence of a series of salt adducts (K+ ions) on the oligonucleotide (figure 14b inset). Nonetheless, I was still able to acquire high-quality CID MS/MS scans, with fragment ions spanning the entire Let-7 sequence, demonstrating that I can distinguish between the extensive number of fragments that are created during MS2. (figure 14c). A demonstration of the output of the data processing step is seen in figure 19, which shows identifications (green boxes) of different charge states of Let-7 at picomole quantities.

One challenge I faced while optimizing these nanoLC-MS/MS methods was inconsistency in nanospray stability and quality. Salt build-up on the spray-tip further impaired spray stability and necessitated frequent replacement of the tip, and prevented me from effectively using fritless columns with integrated tip. Initially, I used commercial columns to perform separation; however, these columns needed to be replaced frequently because of clogging. To gain more control over the quality of spray and reduce costs, I pulled and packed columns

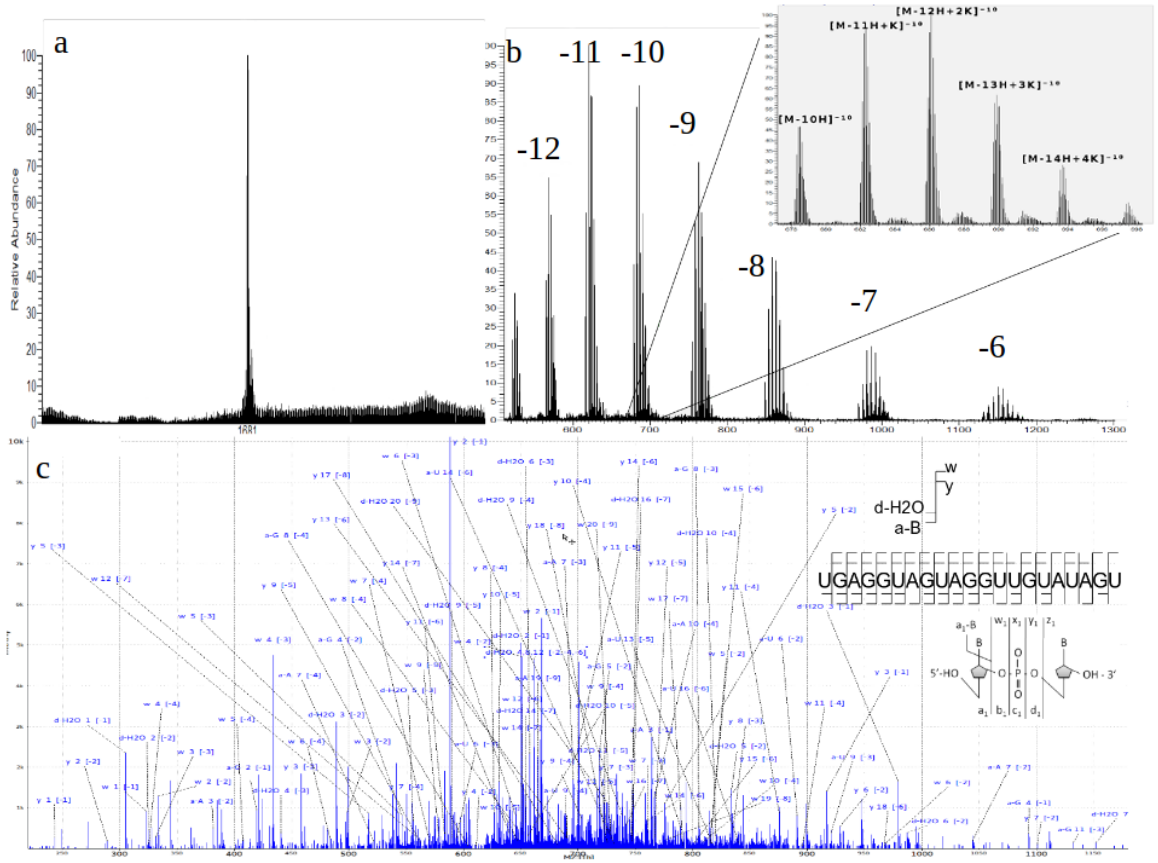

Figure 14: a) Chromatogram of Let-7, b)Mass Spectrum of Let-7 showing the same analyte at multiple charges. Inset shows the details of the analyte with varying number potassium adducts. Close examination shows the isotopic peaks. c) Tandem mass spectrum (CID) of Let-7, annotated by hand to show identified fragments of Let-7. RNA fragmentation produces a wider variety of fragment ion than peptide fragmentation, making annotation more difficult. Schema at right shows possible fragments, and sequence ladder shows detected fragments.

"in-house" specifically for nucleotide work. Using a 2-D HPLC with two gradients available has also allowed me to improve column life and provide more consistent results by including a "wash-cycle" with a simple aqueous (100% H2O) to organic (90% MeOH) gradient, after every 5 runs and at the end of an experiment. I found the addition of this step also decreased column back-pressure after a full cleaning cycle. The addition of a wash run omitting ion-pair reagents and cycling from 5% organic to 90% organic over 30 minutes has also decreased the incidence of clogging in the spray tips. I have also seen an improvement in signal, and a decrease in adduct formation by reducing the concentrations of ion-pairing reagents used in previous micro-flow work Lin et al. (2007). Nucleotides commonly form adducts with sodium or potassium ions, which decrease the total ion signal for the analyte and can damage the instrument (figure 14b). At the start of the method optimization, I saw significant adduct peaks, which I was able to suppress by adding 1mM EDTA as a chelating agent. The EDTA and attached salts were then washed out during the loading of the trap column.

2.3.1. Spray assist device

Working with negative mode ESI-MS is challenging, doubly so with the addition of ion-pair reagents. I found that in order to achieve consistent signal stability, it was necessary to modify an existing source to introduce sheath gas (that is, a gas introduced around, and in the same directions as, the electrospray tip). I followed existing research into adding a non-inline mixture of an inert gas and an organic solvent Huber and Krajete (2000). My experiments showed that adding the nebulized sheath liquid inline with the spray tip would require a less complicated hardware setup, as well as providing a more consistent distribution of the nebulized liquid around the tip. Design of the device took several iterations, with attention paid to minimizing the difficulty of setup and achieving an even mixture of liquid and gas. This mixing proved an engineering challenge, since commercially available parts for chromatography were either designed for gas or for liquid. The current device uses a commercially available mixing tee coupled to a check-valve to prevent back-pressure from

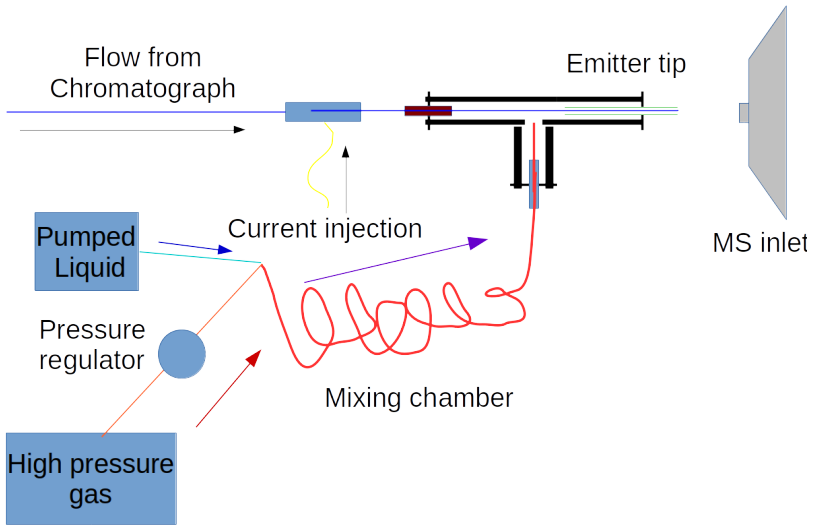

Figure 15: Schematic of the experimental setup including the sheath spray device. Sheath liquid, Acetonitrile in my experiments, is pumped from a syringe pump. Nitrogen flows from a high pressure ultra high purity nitrogen cylinder and flows through a pressure regulator. Gas and liquid mix in small diameter PEEK tubing to ensure an even gas liquid mixture. Voltage for electrospray is injected via an electrode between the column and the electrospray tip.

the gas side displacing liquid.

A custom built nano-ESI source was assembled to fit on a Thermo Orbitrap Classic. Attached to the source was a capillary containing a mix of nitrogen pumped from a gas cylinder, and Acetonitrile pumped by a HPLC pump. This mixture was then pumped into a tee containing the fused-silica electrospray tip. Flow from the analytical column entered the back end of the spray tip at a junction prior to the tee. The front end of the electrospray tip protruded out the opposite end of the tee, surrounded by a larger sheath capillary which provided space around the tip. I measured spray stability of nucleotides with and without sheath flow, as well as the signal produced in an Orbitrap mass spectrometer.

To assess spray stability across a range of sheath gas and sheath liquid conditions I attached syringe filled with standard calibration mix (calmix) to a pulled silica capillary. The capillary was inserted into the spray assist device, with the tip of the pulled capillary protruding approximately 1mm from the sheath capillary. The combined device was then attached to a

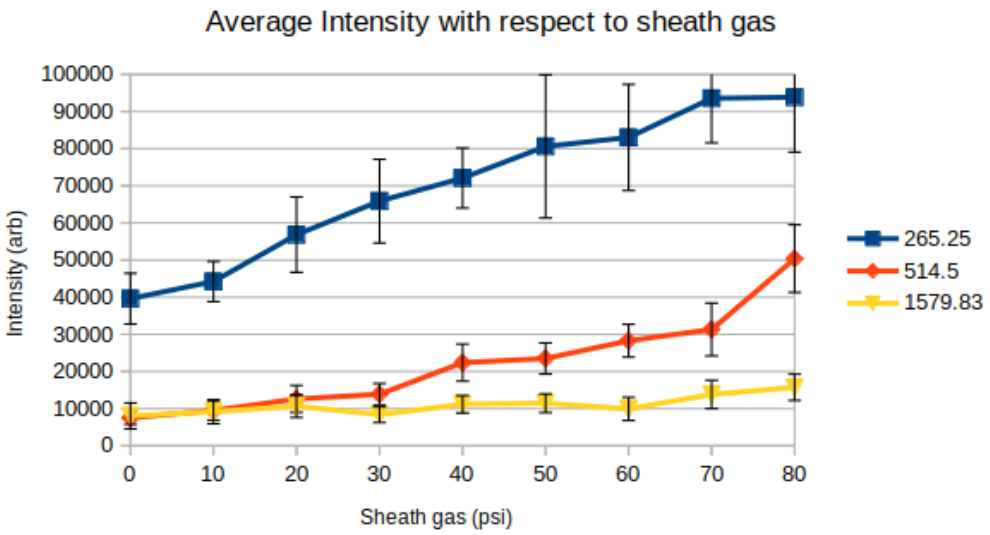

Figure 16: A graph showing the average intensity of the peaks for three different substances in the calibration mix, m/z of each are shown in the legend. Sheath gas pressure was measured at the regulator, and no sheath liquid flow was applied for this experiment. Intensity units on the Y axis are arbitrary but internally consistent in the instrument Error bars show the standard deviation between scans in average intensity.

Thermo Fisher Orbitrap LTQ. The spray voltage was set to 1.5kV and the syringe flow was set to 500nl/min. Sheath gas pressure was tested at 0, 10, 20, 30, 40, 50, 60, and 80 psi as measured at the gas regulator. Sheath liquid flow was set to 0 for these experiments. Next I set the sheath gas at 30psi (the setting I had used for nucleotide data acquisiton) and varied the sheath liquid flow to 250nl/min, 500nl/min, 1000nl/min, and 2000nl/min while holding the gas pressure steady at 30psi. I recorded data in FTMS mode for 60 seconds in each of the experimental conditions. Using in house software derived from OpenMS' EiceExtractor I measured average peak height, max peak height, total peak area and intensity variance at three different masses over the 60 second run. The masses chosen (265.25, 514.5 and 1579.83) were selected as being particularly abundant, and representing a wide range of m/z.

By all three metrics (average intensity, max intensity, and total peak area) the addition of sheath gas (independent of sheath liquid) significantly improved the signal. Interestingly

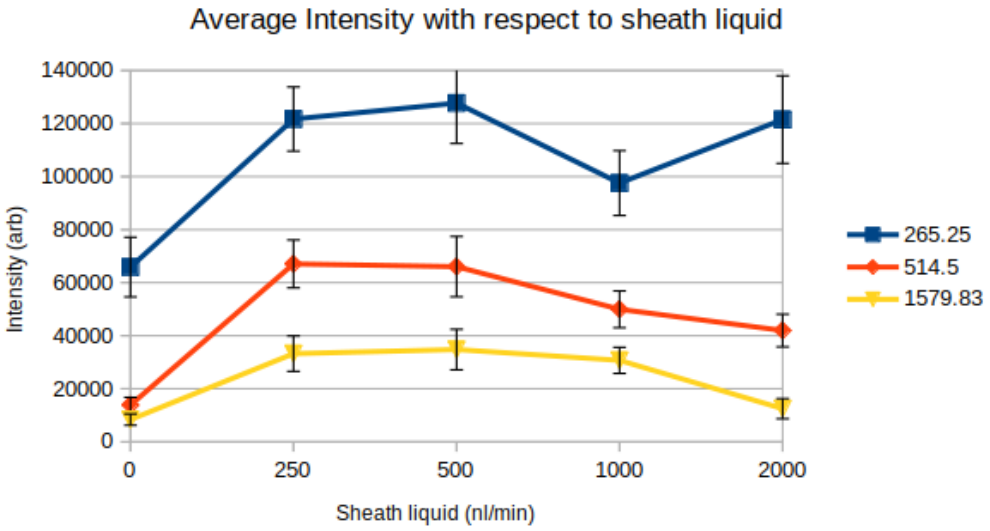

Figure 17: A graph shows the relationship between varying sheath liquid (actonitrile) flows and average peak intensity for the same three analytes as in the above graph. Sheath gas flow was held constant at 30PSI across all of the different liquid flows.

enough, in these experiments I saw a continued improvement in signal intensity as gas pressure was increased. This is in contrast to my previous experiments assessing signal stability using synthetic oligonucleotides, where increases in gas pressure above 40psi did not improve stability and appeared to be detrimental. I hypothesize that there is a link between the significantly greater size of the oligonucleotides (6000-8000 daltons) as compared to the calmix compounds decreasing their ability to be sufficiently ionized at higher gas flow rates.

The varied sheath liquid flow experiments also showed significant improvement in average intensity, max intensity and total peak area upon the introduction of sheath liquid to the device. In contrast to my gas-only experiments, I was able to find a maximum signal improvement at 500nl/min. Visual inspection of the spray tip showed an increase in liquid gathering at the sheath capillary tip when the liquid flow was increased to 1000 and 2000nl/min. I hypothesize that increasing the ratio of sheath liquid to sheath gas above a critical limit prevents even mixing of the liquid and gas, resulting in a decreased positive impact on analyte signal. Further experiments will be necessary to determine whether or not there is a consistent critical ratio above which sheath liquid is detrimental.

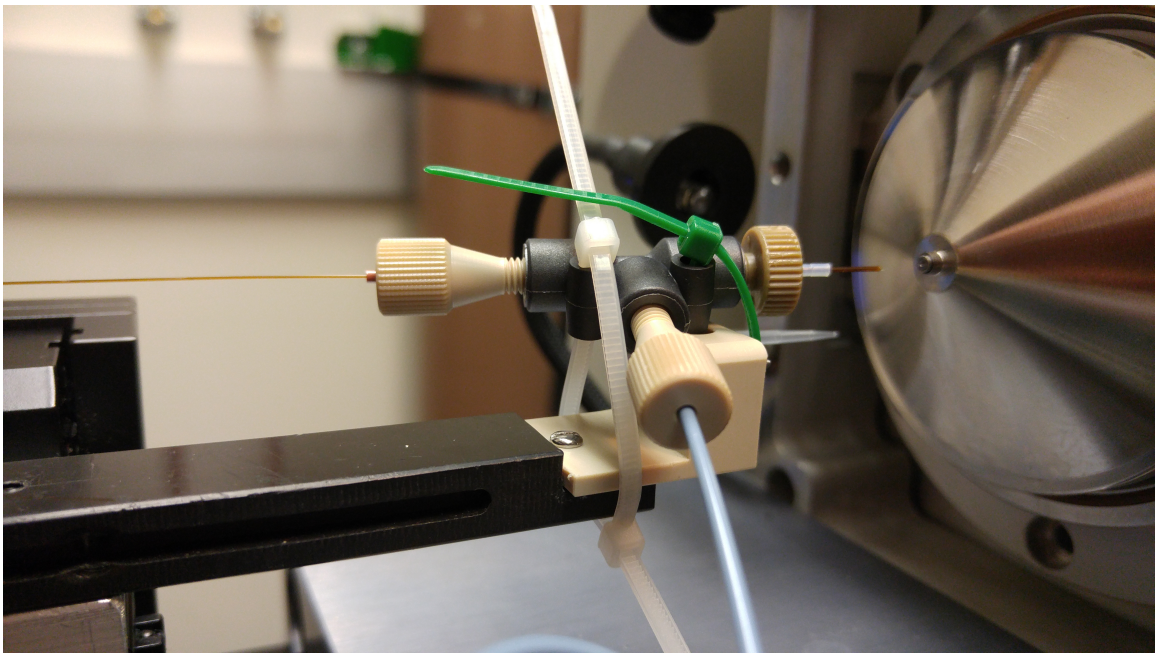

Figure 18: A picture of an early iteration of the coaxial sheath spray assist device.

Another open question is whether the optimal sheath liquid and gas settings are dependent on the m/z of the compound. In my experiments with adjusting sheath liquid, the higher m/z analyte shows much more improvement (4.06-fold) at a sheath liquid flow of 500nl/min than the lowest m/z compound (1.75 fold). I hope to expand my experiments to include a wider variety of m/z compounds to determine whether this trend holds. In my sheath gas tests, I did not observe a clear trend in fold intensity increase dependent on m/z. Additional future work is needed to confirm or deny this independence.

My sheath spray device offers substantial improvement to analytes run in negative mode nano-flow mass spectrometry. I demonstrate this qualitatively by looking at my improved ability to observe oligonucleotides, as well as quantitatively in my assessments of the calmix compounds at a variety of sheath gas and liquid flows.

I can distinguish modified and unmodified nucleic acids. Using my sheath spray assist device along with reverse-phase ion-pair chromatography, I am able to differentiate between different miRNA sequences and between methlyated and unmethylated species

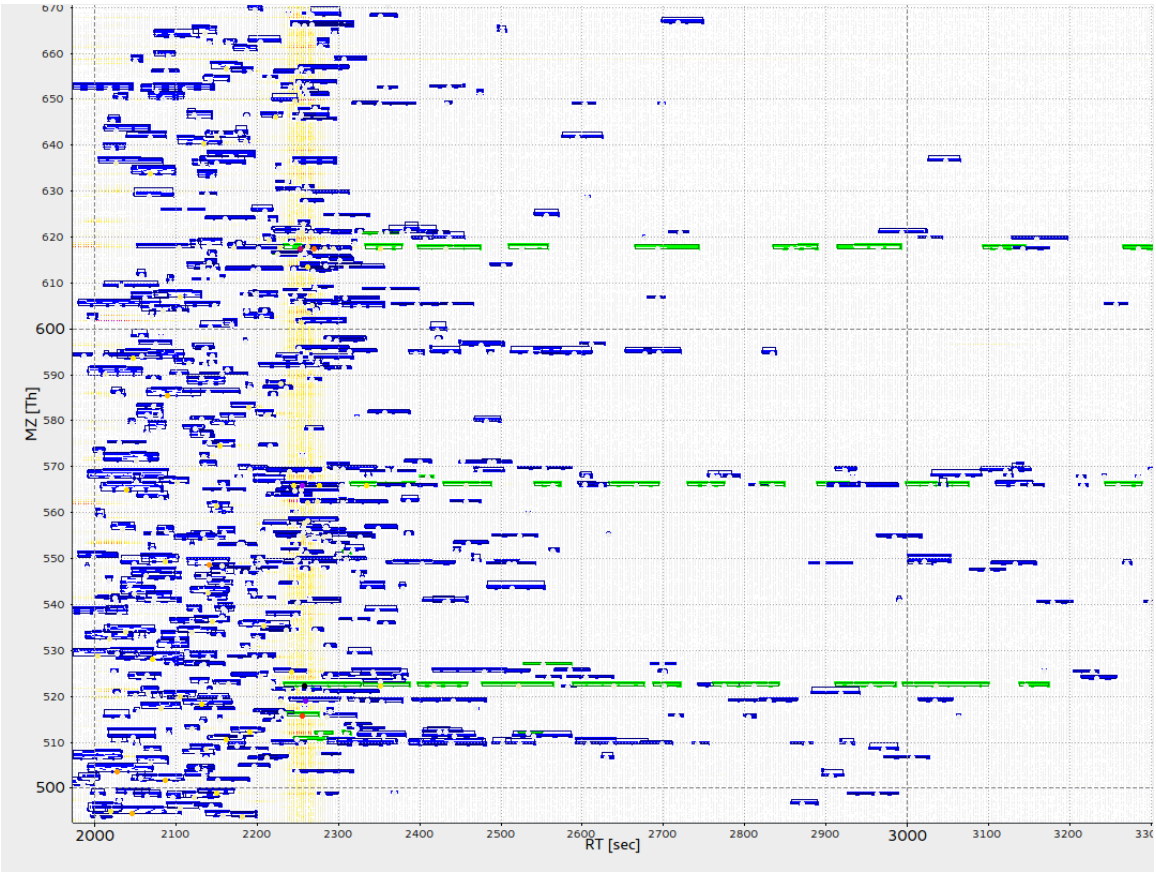

Figure 19: Identified miRNAs and features displayed over a 2d representation of an MS experiment using OpenMS' TOPPView. The X-axis is retention time, and the Y-axis is m/z. Intensity of peaks is represented by color. Blue rectangles are sets of peaks identified by the featurefinder algorithm. Green rectangles are features with an assigned identification.

observed across the m/z range collected by the Orbitrap. The observable charge states of the miRNAs are between negative 6 and negative 15. The resolution on the instrument was sufficient to distinguish between m/z differences as low as .01 dalton, which is high enough to differentiate between methylated and unmethylated sequences. Figure 20 demonstrates the clear separation of the methylated and unmethylated forms of the oligonucleotide. Panel 20a shows separation in both retention time (x-axis) and m/z (y-axis) between the two labeled green identified features, panel 20b shows the chromatograph demonstrating separation of retention time, and panel 20c shows the shifted peaks in the mass spectra of the methylated form compared to the mass spectra of the unmethylated form.

I can achieve sufficient coverage of the fragment spectra of nucleic acids to be able to definitively identify individual nucleic acids. In contrast to peptides, which produce predominantly two types of ions during fragmentation, nucleotides can produce up to nine different fragmentation types with varying abundances. Figure 21 demonstrates the complexity that this generates. The MS2 spectrum shown is 3' 2'-O-methylated Let-7 all of the peaks corresponding to fragments of the oligonucleotide are labeled, and a ladder of identifications is shown in the inset. Starting out in this project I generated a list of theoretical fragments for a given RNA sequence using the web-applet MongoOligo[1]. I then progressed through most abundant MS/MS spectrum, or spectra, from the experiment and hand annotated peaks which matched the theoretical fragments. This process was arduous and so for the later part of this work, I developed a theoretical spectrum generator to produce model fragments for nucleotides, and score their matches to the experimental spectrum, to produce a score for all potential assignments. This work can be seen in my `NucleotideID` node for OpenMS[2] . In both manual and software alignments, I am able to identify enough fragments to identify between different sequences of miRNA. Identical sequences modified at different locations cannot be distinguished by parent mass alone.

[1] http://mods.rna.albany.edu/masspec/Mongo-Oligo
[2] https://github.com/poshul/OpenMS/tree/feature/calculate_RNA_masses

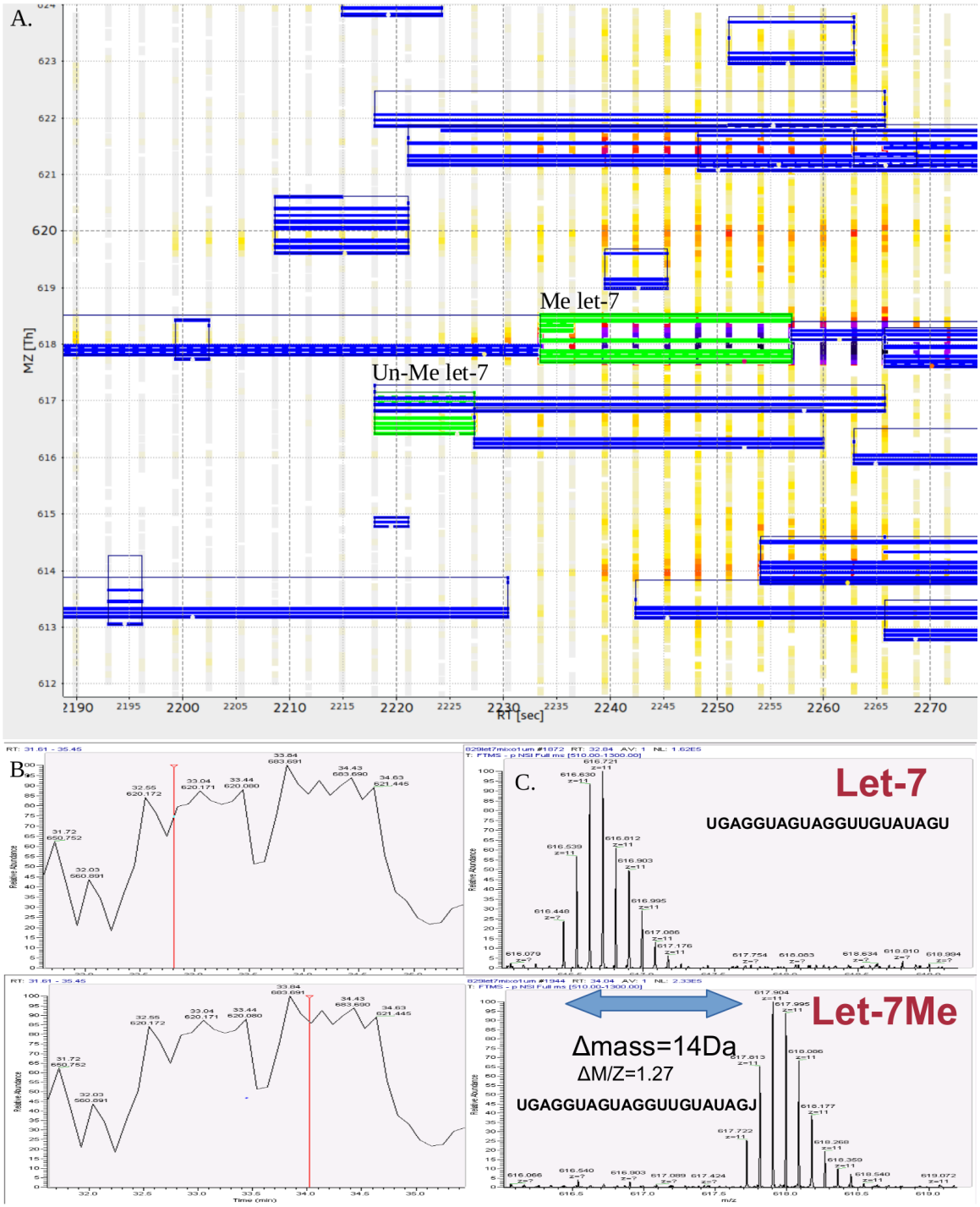

Figure 20: Separation between 2'-O-methylated and unmethylated Let7. A) A two dimensional view of the experiment. X-axis is retention time, and y-axis is m/z. Potential features are marked as blue boxes with each line outlining an isotopic peak in successive scans. Identified features are marked in green. B) Selected chromatograms and C) spectra. Separation between unmodified and modified species is noticeable in both retention time and m/z.

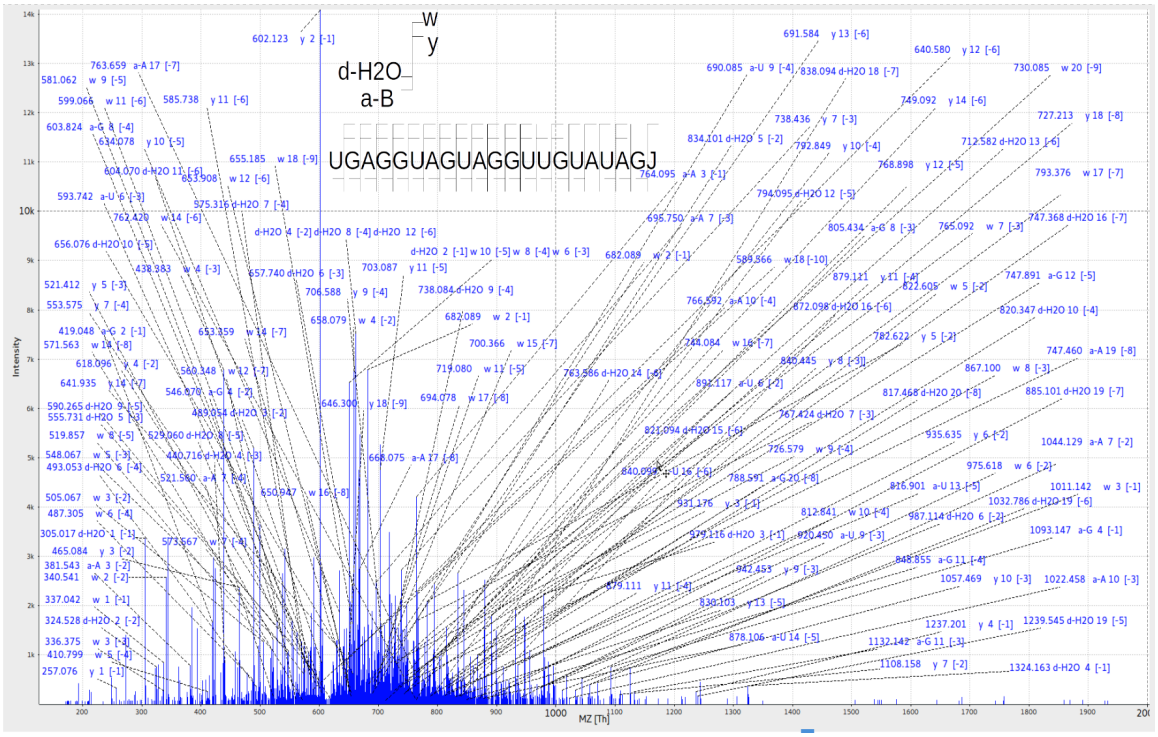

Figure 21: MS2 spectra of 3' 2'-O-methylated Let7. Identification ladder is inset. There are at least two different ion types detected at each position in the sequence, showing good and confident annotation.

I am able to detect oligonucleotide presence down to a concentration of 5 femtomoles per sample. I generated samples of my synthetic oligonucleotides with 1pMol, 500fMol, 100fMol, 50fMol, 10fMol, 5fMol and 1fMol each per sample. From these serial dilution experiments I can confidently detect oligonucleotide presence as low as 5fMol per sample. This corresponds to a signal to noise ratio of at least 4. I hope to continue improving my spray stability to increase my ability to detect even at lower concentrations. Future work identifying oligonucleotides at low concentration will likely be aided by limiting my detection programs to identifying peaks with high charge states.

2.4. Conclusions

I have developed a functional platform which helps automate the computational analysis of oligonucleotides. Through innovations in both my laboratory methods and my updated software, I am able to separate and identify a variety of oligonucleotides. This work has

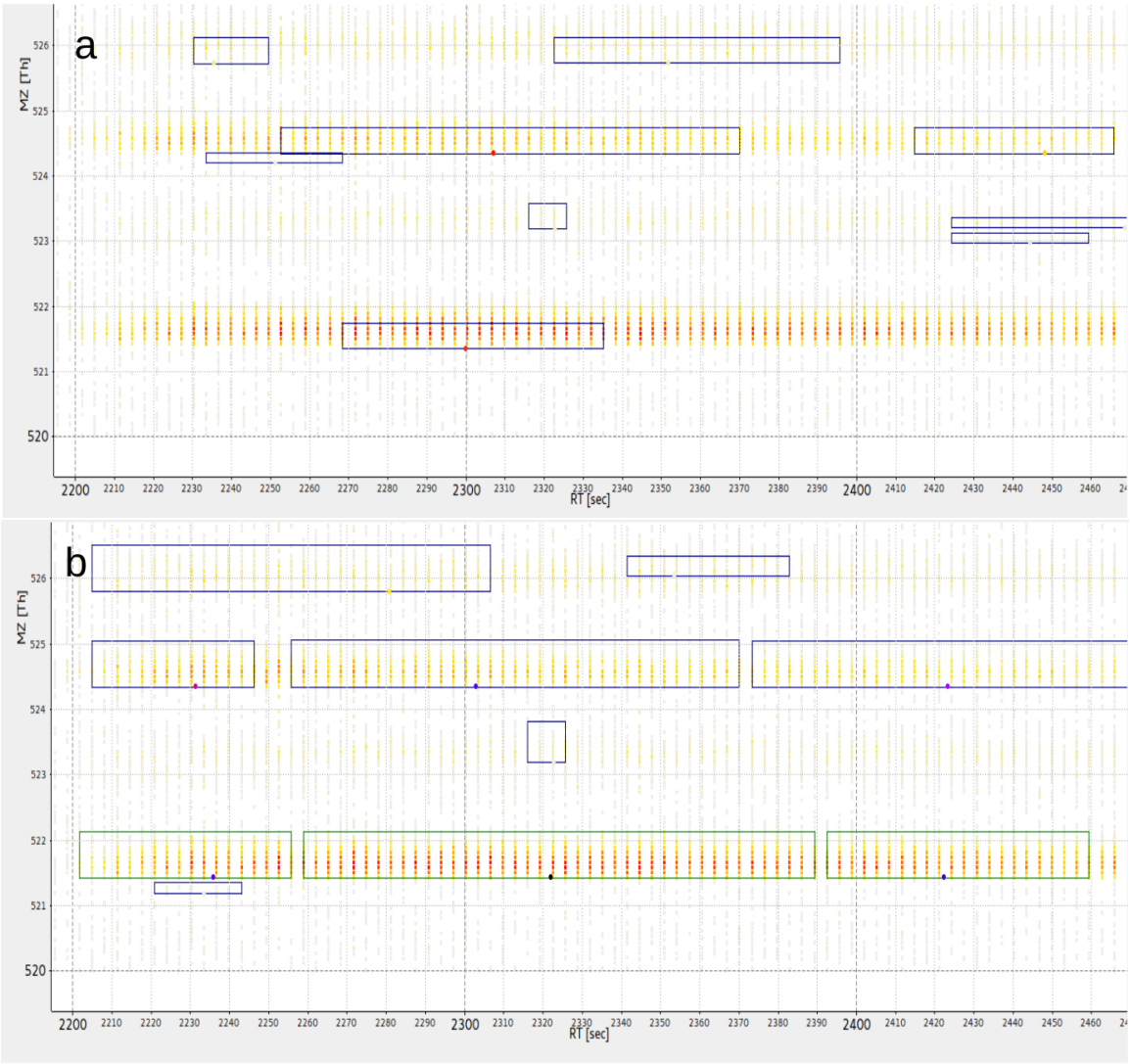

Figure 22: Features generated from the same 1pMol let-7 experiment data file using: a) feature selection using builtin average isotope distribution model (averagine). b) feature selection using user-defined RNA averagine. Since the atomic composition of nucleotides is substantially different than the atomic composition for amino acids, many valid feature identifications are discarded by the algorithm for not matching the predicted isotopic abundance. Providing the algorithm with a corrected average monomer composition results in much better selection of both let-7 (green boxes), and a mono-adducted form (blue boxes above the green in b).

shown that I can use MS to effectively identify nucleotides in low level quantities to be worthwhile for use in the analysis of small non-coding RNAs in general and miRNAs in specific. Continuing this research, I will work on identifying modifications that have not been seen before on miRNA. I also see potential applications of the software I have developed to a wide range of nucleic acid analysis, including but not limited to exploring DNA methylation products and characterizing natural modifications in viral genomes.

The initial setup this system required specialized equipment and expertise, as my experiments have progressed I have been able to simplify the lab setup. Moving to the EASY-Spray column and source has been the largest step toward this, allowing this method to be conducted with commercially available equipment.

CHAPTER 3 : Development and testing of the Software platform

3.1. Introduction

The computational portion of the platform went through two distinct phases. In the first, I used a feature detection based approach to select which sets of peaks were likely to be RNA. This approach worked well in testing with simple samples, and was very useful in the development of the analytical method (as described in chapter 2), however the time taken to search more complex data sets (such as the ones discussed later in this chapter) quickly became prohibitive. I then reevaluated whether to use parent or daughter (that is MS1 or MS2) spectra to identify whether a peak was nucleic acid or simply a look alike. Switching to searching for MS2 spectra with parent masses which were feasibly nucleic acids, and then looking at their parent spectra greatly decreased search complexity, and allowed the final tool (`NucleicAcidSearchEngine`) to be much faster, making analysis of biological samples feasible. In this chapter I discuss the development of `NucleotideID` and the later `NucleicAcidSearchEngine` (NASE).

3.2. Methods

3.2.1. Data Processing with NucleotideID

Raw files from Thermo Fisher's mass spectrometry software suite Xcalibur were converted to the open and standardized MzML format using MSConvert (a component of ProteoWizard[1]). I detect, group and quantify the mass traces of eluting (oligo-)nucleotides using a feature detection algorithm. I build upon an existing algorithm originally developed for label-free and labeled proteomics data available in the OpenMS framework for computational proteomics and metabolomics Rost et al. (2016). In order to make it applicable to (oligo-)nucleotides, I extended both parts of the OpenMS core library as well as the `FeatureFinderMultiplex` tool to support the characteristic isotopic envelopes (i.e. the ratio of heavy to light isotopes characteristic of the sample) of (oligo-)nucleotides in the

[1] http://proteowizard.sourceforge.net/ version used:3.0.4833

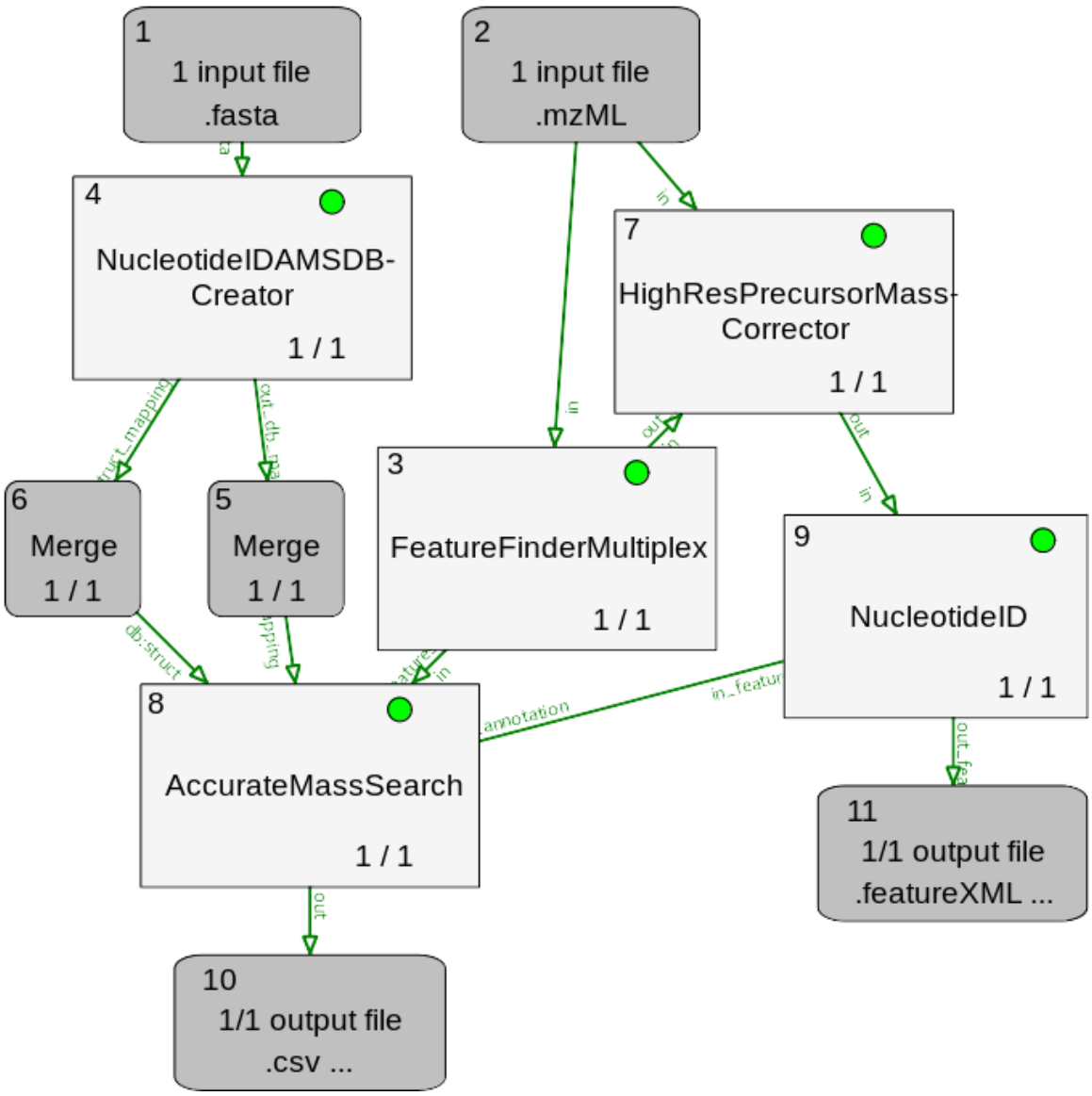

Figure 23: A schematic of the computational workflow. Nodes are referred to by their number in the upper left. 1) Input FASTA file, contains all of the known miRNA sequences source species of the sample. 2) Input mzML file, the MS experiments. 3) FeatureFinderMultiplex, takes a MS experiment as input, locates and annotates peaks as features which are likely the signal from Oligonucleotides, and outputs the detected features. 4) NucleotideIDAMSDBCreator transforms the input FASTA sequences into a database used by AccurateMassSearch. 5,6) Merge the two outputs from 4 into 8. 7) HighResPrecursorMassCorrector, takes the features from 3, and the experiment from 2, and corrects the m/z of peaks within each feature. 8) AccurateMassSearch annotates MS1 features with miRNA sequences that match the feature's mass. A list of putative modifications are stored in output 10. 9) NucleotideID, takes the corrected MS experiment and annotated features as input. Theoretical MS2 spectra are generated for each miRNA identified in 8, and compared to experimental tandem mass spectra. The results are scored and stored in output 11.

detection process. `FeatureFinderMultiplex` performs a filtering step to detect isotopic envelopes in single spectra, a clustering step that combines isotopic clusters over the chromatographic elution, and an (optional) linear fitting step required to determine relative quantification for isotopically labeled data. Since the experiments did not involve isotopic labeling, the linear fitting step was disabled by configuring the tool to perform label free analysis.

Several properties determine if mass peaks are considered part of an isotopic envelope of an analyte. For one, a user-specified minimum number of isotopes with intensities above a specified threshold is required. For another, the relative intensities of the isotopic peaks need to agree with a theoretical average monomer model (averagine). Averagine is a model of the average elemental composition of a theoretical monomer (single amino or nucleic acid). Previous to this work, the averagine model was fixed to proteomics data. I added a user definable parameter to define the averagine model to use for feature selection. Since the atomic composition of amino acids and nucleic acids differ, this addition greatly improved feature detection. The clustering step groups the data points in the isotopic envelopes over retention time, using a hierarchical clustering algorithm. The intensities of each group of data points are summarized into a single abundance value. This abundance value along with the mass traces of the analyte form a feature.

Feature identification, i.e. annotating a nucleotide sequence to each feature, was performed using existing tools of the OpenMS framework. Identification of each spectrum was done using a two-step process. First, the parent ion mass was compared to all possible miRNA masses for the species of interest (*Drosophila melanogaster*) based on a library created from a FASTA file using the `NucleotideIDAMSDB` tool. The `AccurateMassSearch` tool in the OpenMS framework compares the mass of each feature to all of the masses of the miRNAs in a user supplied database, and for each match, the feature is annotated with a candidate sequence. In the second step, this list is passed to another tool which generates a spectral library of theoretical fragment spectra based on the previously determined candidates. The

theoretical fragment spectra were then compared to the experimental MS2 spectrum using the OpenMS spectral library search tool `MetaboliteSpectralMatching`. The highest matching score is reported as best fit. More information about the tools in OpenMS and their parameters can be found online at http://ftp.mi.fu-berlin.de/pub/OpenMS/release-documentation/html/index.html. See Figure 23 for details of the NucleotideID workflow.

Improvements to the feature finding algorithm drastically improve feature identification for oligonucleotides. The OpenMS `FeatureFinderMultiplex` algorithm relies in part on a comparison between the experimental and theoretical isotopic distributions for a molecule. The algorithm was developed for use with peptides, which have a characteristic average monomer mass, or "Averagine". Since the atomic composition of nucleotides varies significantly from that of peptides, the theoretical isotopic distributions that the feature finder produced do not match experimental distributions. By adding a parameter to define the atomic composition of the theoretical average nucleotide monomer, I have been able to improve the definition of the theoretical isotopic envelope, thereby enhancing identification of nucleic acid features in the data. This can be seen in Figure 22, where panel 22a demonstrates the low number of identifications generated using a feature finder with the incorrect isotope settings. Panel 22b shows the results of the same feature finder algorithm but with the correct isotope settings. Many more features are detected (blue boxes) and identification of the nucleotide is successful (green boxes).

Given the relatively limited number of miRNAs for a given organism (˜2000 in humans Hammond (2015)), the database-based approach works well at unambiguously identifying a given oligonucleotide. Unlike the large size of the proteome, the search space for a miRNA is much smaller, allowing the identification approach to be extremely successful in definitively identifying the miRNAs despite their more complicated fragmentation profiles. The comparatively small search space also makes a combinatorial approach to identifying modifications possible; a modern desktop computer has the power to search experimental data against all known RNA modifications for all known miRNAs.

3.2.2. Data Processing with NASE

Here, I present my work on creating a fast, scalable database-matching tool called `NucleicAcidSearchEngine` (NASE) for the identification of RNA oligonucleotide tandem mass spectra. This new software was developed within the OpenMS framework. NASE will be fully integrated into the primary distribution of OpenMS in the upcoming version 2.5, and will then be accessible to interested users who download the software from the website (https://www.openms.de). In the meantime builds of NASE in the OpenMS framework are available at https://www.openms.de/comp/nase/.

The field has an urgent need for enhanced RNA characterization tools. The field of proteomics has been able to advance rapidly in large part due to its ability to analyze data in a consistent and rational manner. Proteomics has been able to take advantage of another important feature of mass spectrometric data: it is inherently quantifiable, making it able to be compared between different conditions in the same experiment. The software which I introduce here also enables this, making it the first nucleotide mass spectrometry solution to enable label-free quantification of data based on MS feature integration. Here I demonstrate in three different case studies that my software, NASE, is capable of reliably identifying a variety of RNA types from different sources.

3.2.3. Liquid chromatography-tandem mass spectrometry

All of the RNA samples in these experiments were separated by reversed-phase ion-pair liquid chromatography (using 200 mM HFIP + 8.5 mM TEA in H2O as eluent A, and 100 mM HFIP + 4.25 mM TEA in methanol as eluent B) and characterised by negative ion MS/MS in a hybrid quadrupole-orbitrap mass spectrometer (Q Exactive HF, Thermo Fisher). A gradient of 2.5 to 25% eluent B eluted oligonucleotides from various lengths of nanoflow Acclaim PepMap C18 solid phase (Thermo Fisher) at 200 nL/min. The length of gradient was varied according to the complexity of the sample. Precursor ion spectra were collected at a scan range of 600 to 3500 m/z at 120k resolution in data-dependent mode,

with the top five MS1 species selected for fragmentation and MS2 at 60k resolution.

3.2.4. RNA samples

A variety of RNA samples were characterised by nanoflow LC-MS and sequence analysis performed using NASE. Initial work was carried out on a mature Drosophila let-7 sequence that was prepared by solid-phase synthesis. This sequence is a 21nt long microRNA that was among the first miRNAs to be characterised Reinhart et al. (2000). The RNA was chemically synthesised in unmethylated and methylated forms (with or without a 2'-O-methyluridine (Um) at position 21). A sample was prepared by mixing both forms, and was characterised by nLC-MS without further processing, but with varying normalised collision energy (NCE) settings to give different levels of precursor fragmentation.

Subsequent experiments were carried out on NME1, a 340 nt long Saccharomyces lncRNA. NME1 RNA was generated by in vitro transcription, and two samples with and without NCL1 enzyme treatment were prepared. NCL1 is a yeast RNA methyltransferase that catalyses the 5-methylcytidine (m5C) modification Motorin and Grosjean (1999). RNA was extracted and digested with RNase T1 prior to nLC-MS. This endonuclease generates shorter oligonucleotides by cleaving immediately after guanosine residues.

The most complicated sample was a solution of digested crude human cellular tRNA, which was isolated from HAP1 tissue culture using an RNeasy kit (Qiagen) as according to the manufacturer's instructions. Briefly, RNAs can be fractionated by length by differential elution, with RNAs less than 200 nucleotides mostly made up of tRNA, and the larger fraction being mostly rRNA. The shorter RNA fraction was digested with RNase T1, and the resultant oligonucleotides were characterised by nLC-MS.

In all cases of internal RNA digestion by RNAse T1, oligonucleotides are generated with a 5' OH and a 3' phosphate. In the RNA sequence notation used throughout this paper, a p at the end of a sequence represents the 3' phosphate.

Sequence database searches For NASE analyses, all proprietary raw files were converted to mzML format Martens et al. (2011) without compression and with vendor peak-picking using MSConvert Chambers et al. (2012) (https://github.com/ProteoWizard). The full list of fragment ion types (a-B, a, b, c, d, w, x, y, z) was considered for peak matching. Precursor and fragment mass tolerance were both set to 3 parts per million. For precursor mass correction, the monoisotopic up to the fifth (+4 neutrons) isotopologue peak were considered.

The synthetic let-7 data was searched with NASE using unspecific cleavage to account for incomplete RNA synthesis products. An extensive set of potential adducts (Na+, K+, Na22+, K22+, NaK2+, Na33+, K33+, Na2K3+, NaK23+) was used because of the substantial salt that remained from the RNA synthesis reactions. Two copies of the let-7 sequence, one with a fixed 2'O-methylation of uridine (Um) at the 5' position, were specified in the FASTA sequence file. The small size of the sequence database prevented the use of a target-decoy approach for FDR estimation. I thus used a stringent hyperscore cutoff of 150 (corresponding to the 1% FDR in the tRNA sample, see below) to define a high-confidence set of results.

The NME1 data analysis used RNase T1 digestion with one allowed missed cleavage. m5C was set as a variable modification; up to two modifications per oligonucleotide were considered. Na+ was specified as a potential adduct. The sequence database contained the NME1 (target) sequence as well as a shuffled decoy sequence. Using these parameters I successfully identified NME1.

In my search of the tRNA data, 26 variable modifications (based on previous findings in yeast and human tRNA available in the Modomics database at http://modomics.genesilico.pl/) were specified, at a maximum of three modifications per oligonucleotide. See Table 1 for the full list of modifications. Na+ was specified as a potential adduct. The FASTA file contained 420 human tRNA sequences collected from the tRNA sequence database tRNAdb Jhling et al. (2009)(http://trna.bioinf.uni-leipzig.de) plus the same number of reversed decoy

sequences. The digestion parameters were set to RNase T1 with up to two missed cleavages.

Search engine comparison The NME1 data was processed with two other publicly available RNA identification engines, in addition to NASE: Ariadne Nakayama et al. (2009) and RNAModMapper Yu et al. (2017). To this end, the raw files were converted to MGF format using MSConvert. Cleavage and variable modification settings in the searches were the same as for NASE and appropriate for the samples. For Ariadne, the online version at http://ariadne.riken.jp was used in October 2018. The "Calc as partial modifications" option was enabled. The precursor and fragment mass tolerances were left at their default values (5 and 20 ppm). Alternatively, using the parameters from Taoka et al. (2016b) publication (20 and 50 ppm) made no appreciable difference for Ariadne's performance in tests. For RNAModMapper, a program version from July 2018 was used with settings recommended by the author, Ningxi Yu. All available ion types (a-B, w, c, y) were enabled; precursor and fragment mass tolerance were set to 0.02 and 0.1 Da, respectively.

Label-free Quantification In order to perform label-free quantification on the NME1 dataset, target coordinates (chemical sum formulas, charge states, median retention times) for oligonucleotides identified at 1% FDR were exported from NASE. Based on these coordinates, feature detection in the LC-MS raw data (mzML files) was carried out with the OpenMS tool `FeatureFinderMetaboIdent`. The results were exported to text format using OpenMS' `TextExporter`, for subsequent processing and visualization in R 3.5.1 Team (2018). Results from both NME1 samples were merged and feature intensities for oligonucleotides were summed up over multiple charge and adduct states, where available. To ensure comparability, manual adjustments were made in a few cases where modified oligonucleotides had been identified with different m5C localizations in the two samples.

3.3. Results

3.3.1. NucleicAcidSearchEngine

I developed a sequence database search engine for the identification of (modified) RNA sequences based on tandem mass spectra. The software, termed `NucleicAcidSearchEngine (NASE)`, was implemented in C++ within the OpenMS framework. The OpenMS library was extended with classes representing (modified) ribonucleotides (based on data from the MODOMICS database Boccaletto et al. (2018)), RNA sequences, and riboendonucleases. A new generalized data structure for spectrum identification results (supporting peptides/proteins, nucleic acid sequences, and small molecules) and an algorithm for theoretical spectrum generation of RNAs were added as well. The new executable tool NASE combines this and existing OpenMS functionality (e.g. for data input/output, filtering, and FDR estimation - see below).

Data processing with NASE works as follows: Inputs are an RNA sequence database (FASTA format) and a mass spectrometry data file (mzML format). RNA sequences are digested in silico using enzyme-specific cleavage rules for the user-specified RNase. Tandem mass spectra are pre-processed (intensity filtering, deisotoping) and mapped to oligonucleotides based on precursor masses. Theoretical spectra of the oligonucleotides in the relevant charge states are generated and compared to the experimental spectra; matches are scored using a variant of the hyperscore algorithm (Feny and Beavis 2003). If the sequence database contains decoy entries, the resulting oligonucleotide-spectrum matches can be statistically validated through the automatic calculation of q-values, a measure of the FDR Kll et al. (2008). Supported output formats are an mzTab-like text file Griss et al. (2014), suitable for further analysis, and an XML file, suitable for visualisation in OpenMS' interactive viewer, TOPPView Sturm and Kohlbacher (2009).

In addition to the built-in FDR calculation, NASE provides other features that set it apart from alternative tools that are currently available. Even with extensive preparation, nu-

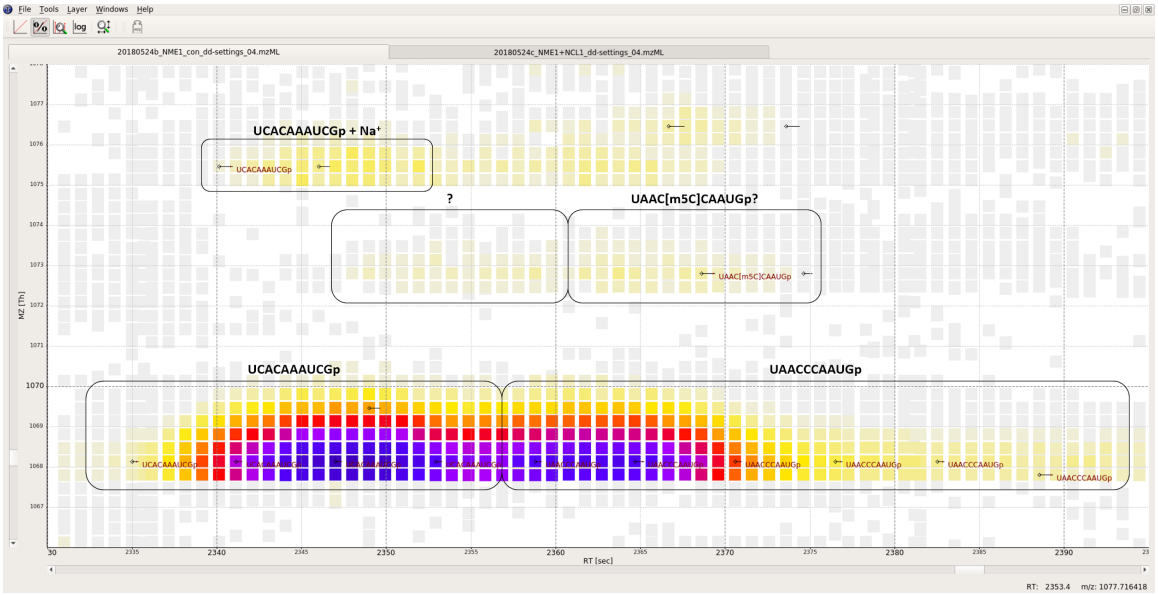

Figure 24: Annotated screenshot from TOPPView showing data from the NME1 control sample, corresponding to the NCL1-treated data shown in Figure 4a. Note the loss of signal intensity and sequence identifications for the methylated oligonucleotides, compared to Figure 4a. Due to a lower-quality MS2 spectrum, the m5C site in UAACCCAUGp has here been localized to the second, not third cytidine.

cleotide samples frequently contain salt adducts (in the form of cations attached to the phosphate backbone). NASE searches can take this into account, by allowing users to specify chemical formulas of adducts to consider in the precursor mass comparisons.

Furthermore, NASE supports the correction of precursor masses for MS2 spectra that were sampled from isotopologue peaks other than the monoisotopic one. Especially for longer sequences, MS2 precursor ions are often picked from heavier isotopologues by the mass spectrometer's data-dependent acquisition software, because the monoisotopic peaks are of comparatively lower intensity. Without correction, the MS2 precursor masses would not closely match the theoretical (monoisotopic) masses of the correct oligonucleotides, leading to no assignment or incorrect matches. This feature thus greatly increases NASE's ability to identify oligonucleotides with longer sequences.

Finally, through the OpenMS toolbox NASE provides basic support for label-free quantification of identified oligonucleotides. The core step of the quantitative workflow, the

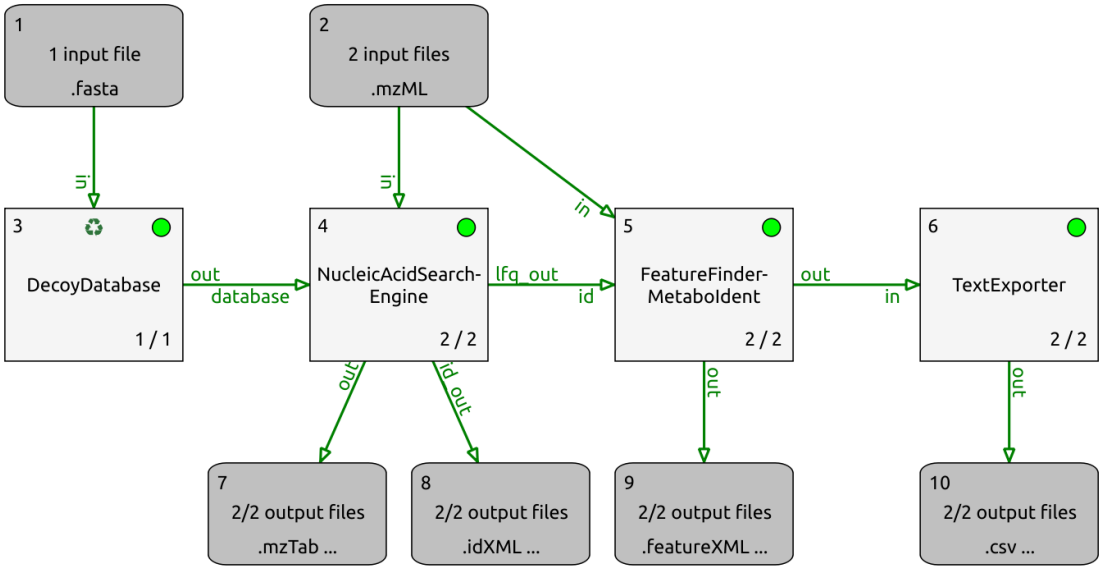

Figure 25: Data analysis pipeline for the NME1 data, comprising target/decoy database generation, database search (incl. FDR estimation and filtering), targeted feature detection and data export. Screenshot from TOPPAS, the OpenMS workflow editor. The whole pipeline ran in only 12 seconds (single-threaded) on my server.

detection of chromatographic features in the LC-MS data, is handled by the OpenMS tool FeatureFinderMetaboIdent (FFMetId). FFMetId is a variant of the proteomics tool FeatureFinderIdentification Weisser and Choudhary (2017) which provides targeted feature detection for arbitrary chemical compounds. NASE can write an output file with all relevant information about the oligonucleotides it identified, which is directly suitable as an input file for FFMetId. This allows seamless label-free quantification of the oligonucleotides that were identified in a sample. Through the inclusion of a graphical workflow editor, OpenMS makes it very easy to create and run data analysis pipelines Junker et al. (2012); an example pipeline from my analysis of the NME1 data is shown in Fig. 25.

3.3.2. Synthetic let-7 analysis

There was a strong dependence of sequence coverage on the Normalized Collision Energy (NCE) value. Identical samples were run with NCE ranging from 5 to 55. The best results were obtained for an NCE of 20 (Figure 30). Subsequent LC-MS/MS analyses,

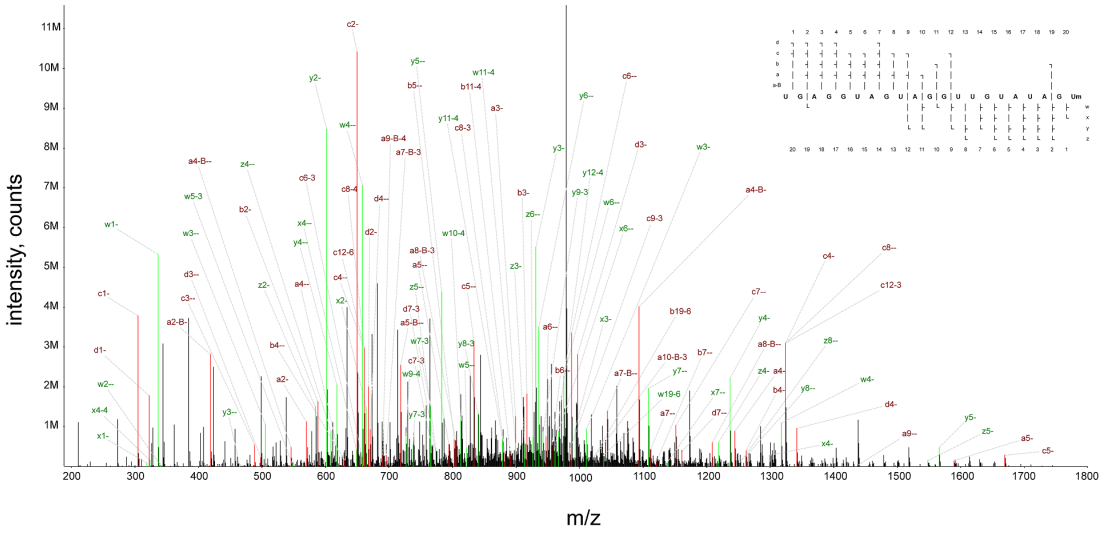

Figure 26: A tandem MS spectrum of let-7 denoting all of the assigned peaks. The primary ion was deprotonated seven times to give a charge state of -7 (m/z 971.55). The ion coverage plot in the upper right shows coverage for nine different types of fragment ion (based on the naming scheme of Mcluckey et al. (1992)).

including of the NME1 and tRNA samples, were thus carried out with this NCE setting. At the optimal NCE, both unmodified and modified RNA were detected, and the location of the modification could be determined with high confidence. 851 spectra were identified that passed my hyperscore cutoff, matching sequences of length 5-21 nt, including the full-length let-7. The shorter sequences correspond to artifacts of incomplete solid-phase RNA synthesis, which are easily detectable by nLC-MS. In the full 21-nt sequence I averaged over two-fold MS2 ion coverage of the let-7 sequence, with one or more forward (a-B/a/b/c/d) ion and one or more reverse (w/x/y/z) ion detected at each base (see figure 26). This demonstrates the software's ability to sequence even relatively long (longer than 20 nt) RNAs.

3.3.3. NME1 analysis

I processed the NME1 data using the three search engines Ariadne, RNAModMapper, and NASE. I compared the results of my target/decoy database searches in terms of: A, the number of identified spectra at different FDR thresholds; B, the sequence length distribution

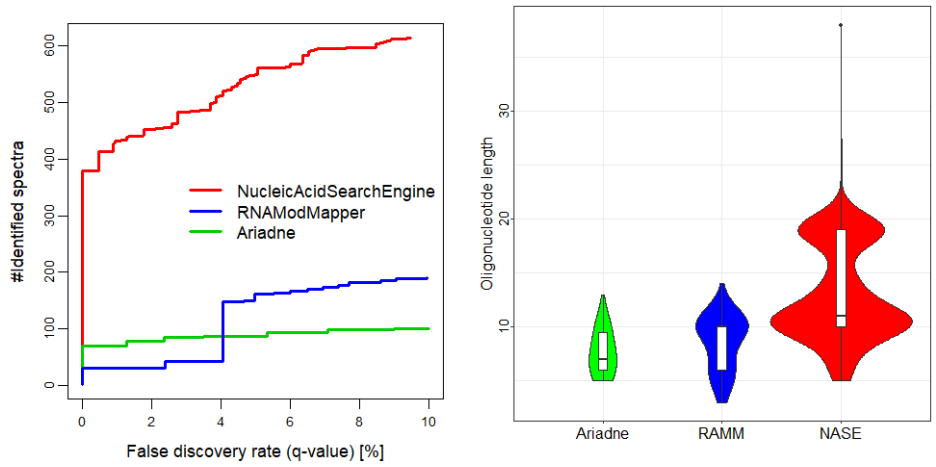

Figure 27: Performance comparison of RNA identification engines (Ariadne, RNAModMapper, NucleicAcidSearchEngine) based on searches of the NME1 data. Left: The number of successfully identified spectra plotted against the q-value, a measure of the false discovery rate, which was calculated from a target/decoy database search using each of the three tools. Right: The sequence length distribution of identified oligonucleotides for each tool at a confidence level of 5% FDR.

of the identified oligonucleotides at 5% FDR (Figure 27). NASE identified significantly more spectra at a given confidence level than the other tools. It also found longer oligonucleotides, which would be more informative for identifying RNAs in complex samples. About 8% of the oligonucleotide-spectrum matches generated by NASE at 1% FDR included sodium adducts (and would have been missed without the adduct search capabilities). Note that Ariadne's performance in this comparison was hampered by the fact that a recommended tool for data preprocessing, the commercial software SpiceCmd, was not available to us. RNAModMapper had previously been evaluated based on searches against expected sequences only (i.e. no decoys), followed by manual validation of spectral assignments Lobue et al. (2018).

To assess reproducibility and the performance of my software at detecting RNA modifications, I compared the NASE search results for the NME1 lncRNA with and without NCL1 incubation (Figure 28). I considered results at a high confidence level of 1% FDR; at this level, 74% sequence coverage was achieved for both the control and the NCL1-treated sample. As Figure 28 shows, there is good agreement between the unmodified oligonucleotides

that were identified in both samples. While a number of m5C-modified oligonucleotides were identified in the control sample, all except two of these false positives were observed in only a single oligonucleotide-spectrum match - in proteomic LC-MS experiments, such single hits would be commonly filtered out Cox and Mann (2008). I suspect that trace amounts of carry-over from earlier test runs of the NCL1 sample on the same chromatographic column may have caused these identifications in the control sample. Nonetheless, two modified oligonucleotides, UCACAAAU[m5C]G (at position 21-30 in the NME1 sequence) and UAACC[m5C]AAUG (positions 299-308), were identified only in the NCL1-treated sample, based on 5 and 4 spectra in multiple charge states (-2 to -4 and -3 to -4, respectively). These oligonucleotides thus provide strong evidence for true m5C modification sites (see also Figures 30 and 31). In a few cases, NASE was unable to uniquely assign the location of a cytidine methylation within an oligonucleotide sequence, reporting two alternative hits with identical scores and different m5C sites. This is a common limitation that affects all search engines (including Ariadne and RNAModMapper), caused by the absence of discriminating peaks in the corresponding mass spectrum.

3.3.4. tRNA analysis

Previous work on tRNA Pan (2018b) has shown that it is heavily modified. My analysis confirms this. I ran NASE on the short RNA fraction of a cell extract sample that had been digested with RNAse T1. I searched for 26 variable modifications with different molecular masses, which had previously been identified to be present in yeast or human tRNA (Machnicka et al. (2015),Jhling et al. (2009)). Most of these represent sets of isobaric modifications which I cannot distinguish, such as position-specific variants of the same modification; e.g. m1A was used to represent any singly-methylated adenosine (incl. Am, m6A etc.). Note that it was not feasible to search this dataset with this high number of variable modifications using other available database-matching tools (RNAModMapper, Ariadne). The full results are available in PRIDE (the PRoteomics IDentifications database) . At an FDR cutoff of 5%, 1341 spectra were matched to oligonucleotide sequences, giving

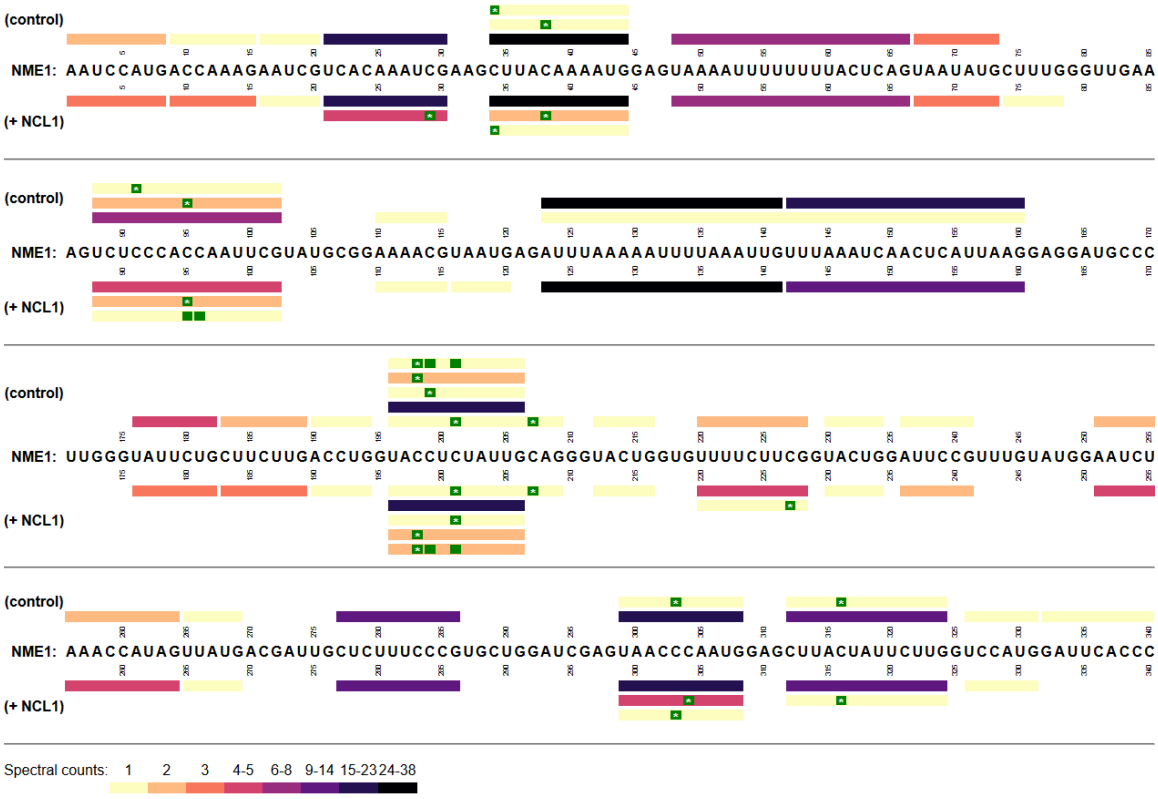

Figure 28: Coverage plot showing the NME1 RNA sequence and highlighting oligonucleotides identified using NASE in the control (top) and NCL1-treated sample (bottom), respectively. The bars corresponding to oligonucleotides are colored according to their number of identifications (spectral counts) at 1% FDR. Putative 5-methylcytidine (m5C) modification sites are marked in green. Sites with an asterisk (*) were uniquely localized, while blank sites indicate uncertainty between two possible locations.

rise to 236 different oligonucleotides. The sequences of human tRNAs share significant similarities, especially for tRNAs of one isotype, i.e. tRNAs that bind the same amino acid. Consequently, only 38 (16%) of the identified oligonucleotides map to a unique tRNA sequence; however, 225 (95%) map uniquely to a single tRNA isotype. Overall 385 of the 420 tRNA sequences in the search database had matching oligonucleotides. 35 tRNAs, including ten variants of tRNA-Gly, were not detected at all. The sequence coverage, when counting all matching oligonucleotides for each of the detected tRNA sequences, ranged from 8.1% up to 54.8% (see Figure 29), with a median coverage of 20.8%. Many of the oligonucleotides that were identified contained multiple modifications. In the search, up

to three modifications per oligonucleotide were allowed, to limit the combinatorial space of modified sequences that needed to be explored. Of the unique oligonucleotides identified at 5% FDR, 11% were unmodified (accounting for 16% of the identified spectra), while 36% carried one, 26% carried two, and 25% carried three modifications (accounting for 45%, 23% and 16% of the identified spectra, respectively) the preponderance of modifications were on the tRNA loops, which is in agreement with previous experiments. All modifications included in the search, except queuosine, wybutosine and their derivatives, were detected as part of identified oligonucleotides. However, the prevalences of different modifications differed widely - see Table 1 for details.

Existing data on the modification landscape of human cytosolic tRNAs is incomplete (e.g. tRNAdb lists information for 20 genes covering 15 isotypes) and at least some modifications are differentially regulated, complicating comparisons. I will focus on cytosine monomethylation (mC, represented by m5C in my search) as one example that has been studied more thoroughly, e.g. via bisulfite sequencing to detect m5C Gilbert et al. (2016). At 5% FDR I identified 24 unique oligonucleotides with unambiguous assignments of mC. The oligonucleotides contained one or two mC sites each and were supported by a total of 173 identified spectra. Each mC-containing oligonucleotide was associated with one unique or predominant (more matching genes) tRNA isotype. At the level of these isotypes, a total of 18 unique mC sites were identified. Seven of these sites agree with the canonical m5C sites in the VL junction at consensus sequence positions 48-50 Blanco et al. (2014). Cytosine methylation at position 34 in the anticodon, previously reported as m5C for tRNA-LeuCAA Brzezicha et al. (2006) and 2'-O-methylcytidine (Cm) for tRNA-Met in tRNAdb, was here observed for tRNA-Met and tRNA-Trp. Methylation at C32 was found in a several tRNAs (tRNA-GlnCTG, tRNA-LeuTAA, tRNA-PheGAA, tRNA-Trp, tRNA-ValCAC); correspondingly, Cm is reported at this position for tRNA-Gln and tRNA-Phe in tRNAdb. I find that my results generally recapitulate annotated modifications in tRNAdb, in regions where I have sequence coverage and with the caveat that I cannot distinguish between isobaric modifications (including uridine/pseudouridine). Known recurring modifications that

Search mod. (short code)	**Represented isobaric modification(s)**	**Spectrum matches**	**Unique oligo-nucleotides**
m1A	Adenosine monomethylation	391	55
m5U	Uridine or pseudouridine monomethylation	247	46
t6A	N6-threonylcarbamoyladenosine	222	29
m5C	Cytidine monomethylation	173	24
m2G	Guanosine monomethylation	114	26
D	Dihydrouridine	107	26
I	Inosine	78	26
acp3U	3-(3-amino-3-carboxypropyl)uridine or -pseudouridine	63	18
i6A	N6-isopentenyladenosine	55	13
m2,2G	Guanosine dimethylation	48	13
ncm5s2U	5-carbamoylmethyl-2-thiouridine	43	17
ac4C	N4-acetylcytidine or 5-formyl-2'-O-methylcytidine (f5Cm)	37	14
Ar(p)	2'-O-ribosyladenosine (phosphate)	26	14
mnm5U	5-methylaminomethyluridine	23	5
m1I	Inosine monomethylation	20	7
m1Im	Inosine dimethylation	10	6
f5C	5-formylcytidine	9	7
io6A	N6-(cis-hydroxyisopentenyl)adenosine	5	4
m5Cm	Cytidine dimethylation	3	3
m5D	Dihydrouridine monomethylation	2	2
m6Am	Adenosine dimethylation	2	1
m5Um	Uridine or pseudouridine dimethylation	1	1
yW	Wybutosine	0	0
Q	Queuosine	0	0
o2yW	Peroxywybutosine	0	0
galQ	Galactosyl- or mannosyl-queuosine (manQ)	0	0

Table 1: Summary of modifications detected in the HAP1 tRNA data using NASE at a 5% FDR level. Columns: 1. Short code of the modification specified as a search parameter. 2. The set of modifications implied by the corresponding mass shift, since e.g. position-specific variants of a modification (Am, m1A, m6A etc.) generally cannot be distinguished. 3. Number of identified oligonucleotide-spectrum matches with at least one instance of the corresponding modification in the sequence. 4. Number of unique oligonucleotides with at least one corresponding modification among the search results.

I detect in several tRNAs include monomethylation at G10, mono- or dimethylation at G26, and monomethylation at A57. In many cases I find additional, alternatively modified (or unmodified) variants of expected oligonucleotides.

3.3.5. Data visualization

TOPPView is the interactive viewer of the OpenMS suite Sturm and Kohlbacher (2009) and allows multidimensional visualisation of tandem MS data. I extended TOPPView with capabilities for visualizing RNA identification results obtained using `NucleicAcidSearchEngine`. These extensions mirror and augment existing TOPPView functionality for visualizing peptide identifications in proteomics experiments. First, identified sequences can be displayed in the context of MS1 signal intensities, positioned according to their MS2 precursors in a two-dimensional RT-by-m/z LC-MS map. Figure 30 shows an example from the NME1 data, focusing on the two (isobaric) oligonucleotides highlighted in the discussion of that dataset and showing unmodified, adducted, and modified variants of them in charge state -3 in the NCL1-treated sample. A corresponding image showing the loss of signal for the modified oligonucleotides in the control sample can be seen in Figure 30.

Second, oligonucleotide-spectrum matches can be visualized by showing MS2 spectra with annotations for matching peaks in the corresponding theoretical spectrum, analogous to the identification view for peptide-spectrum matches. I augmented this view by adding an ion coverage diagram in the top-right corner of the main window, indicating which of the possible fragment ions were matched. Figure 31 shows an example comparing spectrum matches for the two aforementioned modified oligonucleotides from the NME1 dataset.

3.3.6. Label-free quantification

I quantified the identified oligonucleotides in the two NME1 samples, using a label-free, feature detection-based approach. Figure 32 summarizes the results. Although all oligonucleotides come from the same RNA, they were quantified with signal intensities spanning several orders of magnitude. This is indicative of widely varying ionization efficiencies dur-

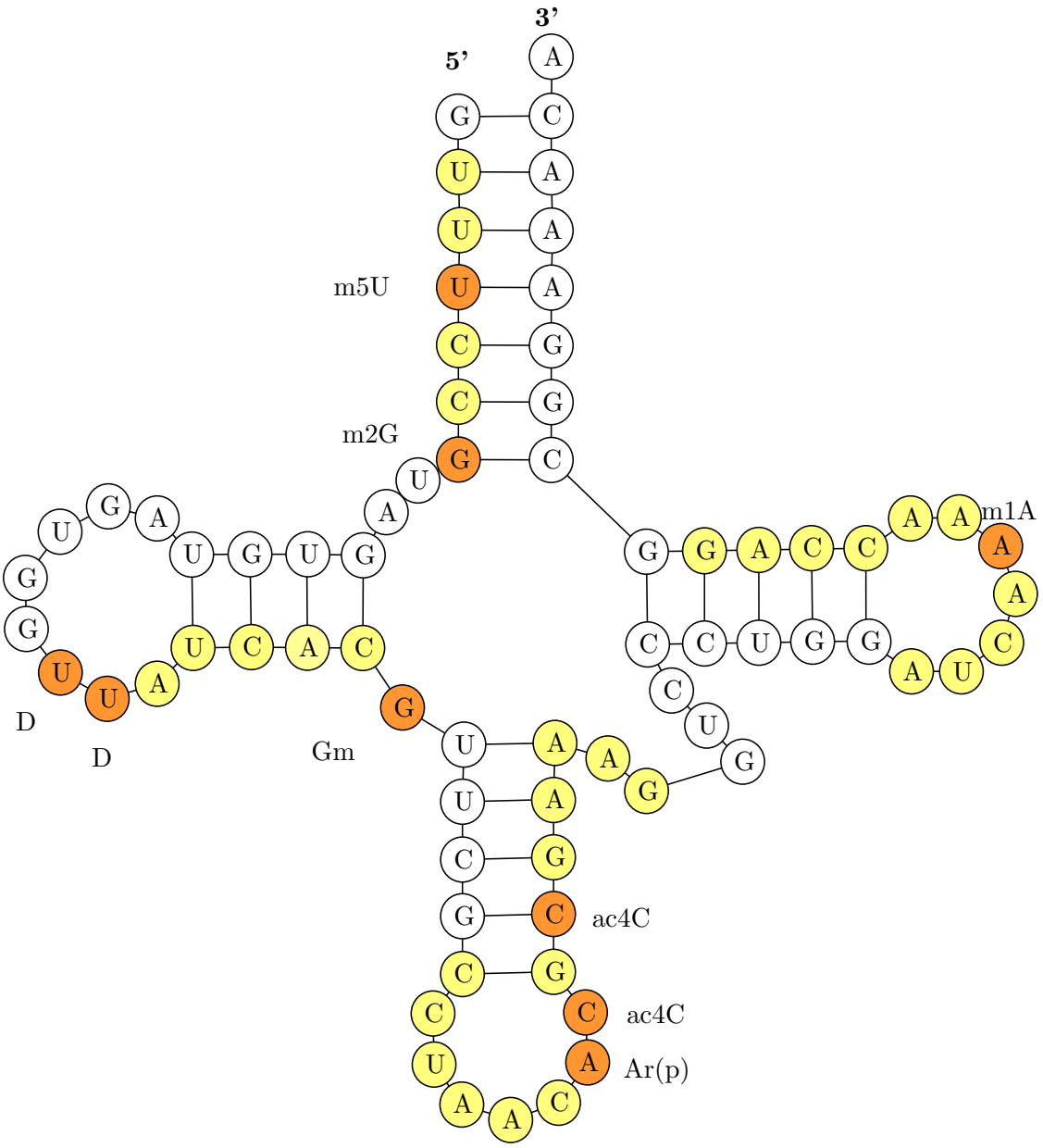

Figure 29: A schematic depiction of Homo sapiens Val-AAC-3-1 tRNA. Sequences which I detected at 5% FDR are highlighted in yellow for unmodified, and orange for modified residues. Total coverage is 54.8%. The tRNAdb entry for tRNA-Val agrees with my findings, except for the methylation at U4 (based on four identified spectra) and the three modifications in the anticodon loop and stem (bottom right, based on two identified spectra).

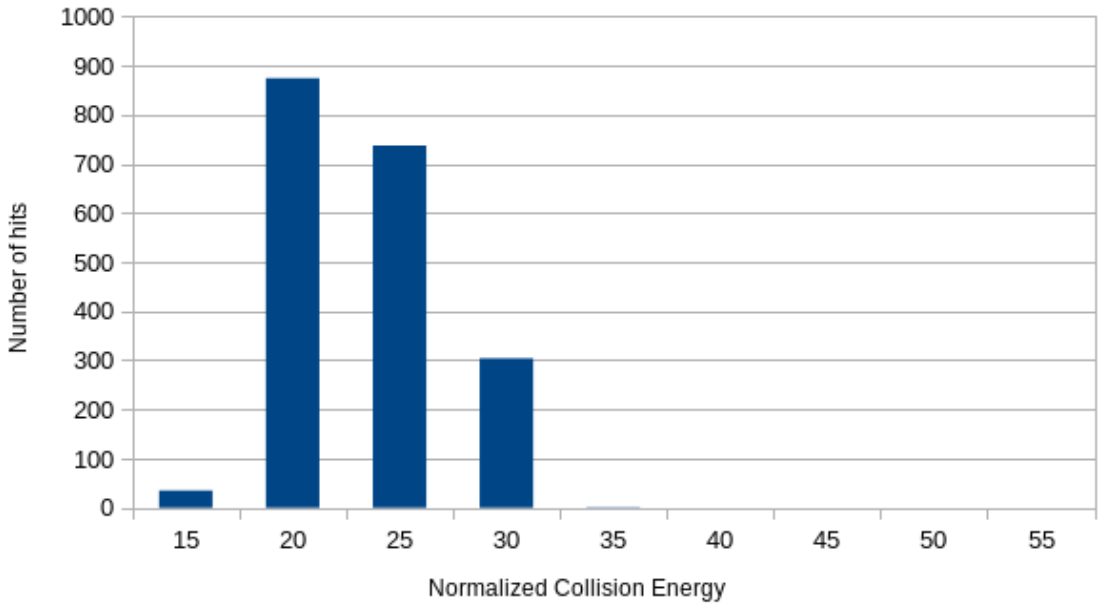

Figure 30: A histogram showing the number of spectra identified as portions of Let7 at various normalized collision energies between 15 and 55. Normalized Collision Energy is an arbitrary value with no associated units. It is therefore likely to vary between different instruments. The rapid drop-off in the number of hits above NCE 20 shows that controlling fragmentation is very important for properly identifying RNA

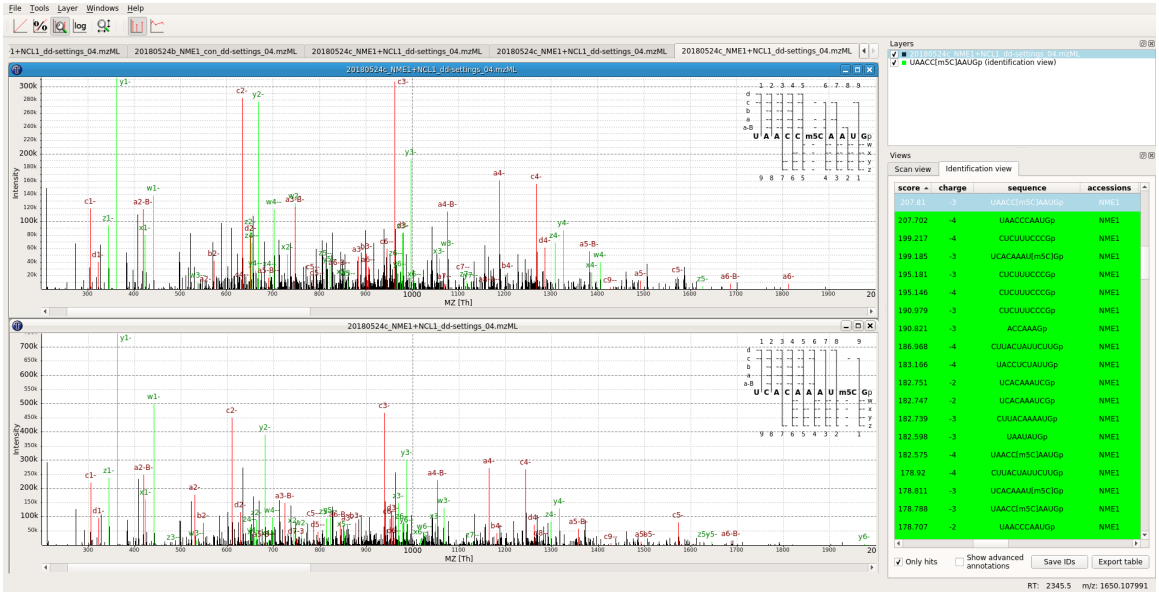

Figure 31: Screenshot from TOPPView's identification view. Two MS2 spectra from the NCL1-treated NME1 data, identified by NASE as the sequences UAACC[m5C]AAUGp and UCACAAAU[m5C]Gp (cf. Figure 4), are compared. Matching peaks between the acquired and theoretical spectrum are annotated and highlighted in red and green. In the top-right corner of each spectrum plot, an ion coverage diagram shows which of the theoretical fragment ions of the sequence were matched in the MS2 spectrum (in any charge state).

ing MS analysis, a common caveat that generally limits label-free quantification to relative comparisons between similar samples. Of 37 and 40 oligonucleotides that were identified at 1% FDR in the control and NCL1-treated sample, respectively, 34 could be quantified in each sample (corresponding to 92% and 85% success rates). Most oligonucleotides were quantified at similar levels in both NME1 samples, with putative m5C-modified oligonucleotides generally found at lower intensities. The notable exception is the pair of modified oligonucleotides UCACAAAU[m5C]G/UAACC[m5C]AAUG already discussed above. While the chromatographic peaks for the unmodified oligonucleotides UCACAAAUCG and UAACCCAAUG were distinct enough to allow separate quantification of each, their modified variants could only be quantified together. The difference in signal intensities for these modified oligonucleotides between the control and NCL1-treated sample is clearly visible in Figure 30 and Fig. 30. This difference is exacerbated in the label-free analysis by the fact that only one corresponding identification was made in the control sample, while multiple charge states were identified, quantified and aggregated in the NCL1-treated sample. (The other obvious outlier, with the sequence AUUUAAAAAUUUUAAAUUG, was eluted for a long period at the end of the chromatographic gradient and thus could not be quantified reliably.) More advanced capabilities for LC-MS-based quantification, including retention time alignment, inference of identified analytes across samples, and labelling approaches, are already available in OpenMS for proteomics experiments. With future improvements to the support for nucleic acids in the framework, these features will become available for RNA analyses as well.

3.4. Discussion

NASE is a new open-source database search engine for RNA, optimised for high-resolution MS data. It supports arbitrary modifications, salt adducts, and FDR estimation through a target/decoy search strategy. Moreover, integration with the OpenMS toolbox enables high-quality data visualisation, e.g. for manual validation of spectral assignments, and label-free quantification of RNA oligonucleotides. I have tested NASE against a range of sample types

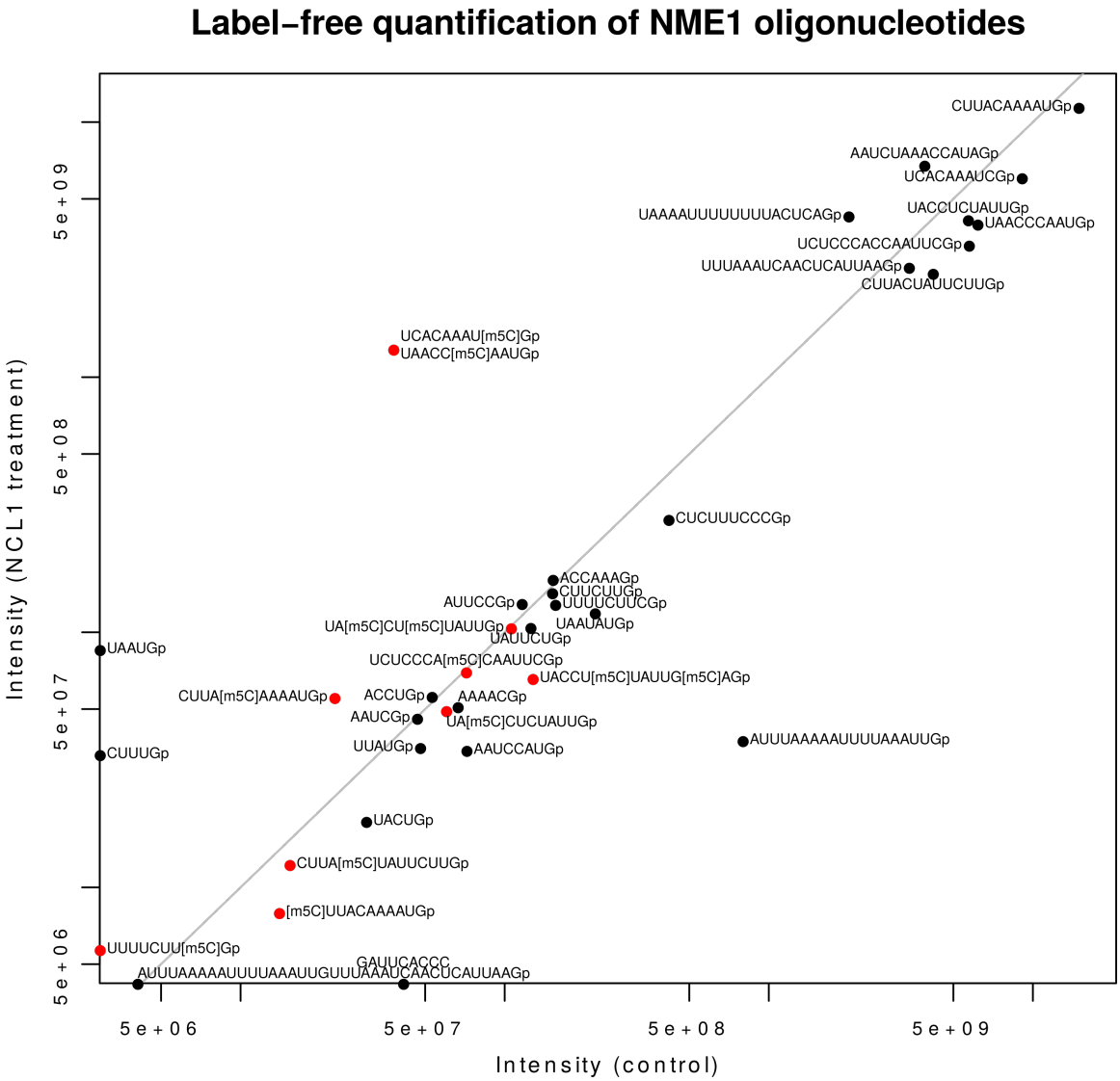

Figure 32: Label-free quantification results for oligonucleotides identified in the NME1 dataset, comparing signal intensities in the control and NCL1-treated sample. Intensities were aggregated over multiple charge and adduct states, where applicable. m5C-modified oligonucleotides are marked in red. Oligonucleotides that were quantified in only one of the samples are shown directly on the x and y axis, respectively. The grey diagonal line represents equal intensity in both samples.

and complexities, spanning synthetic nucleic acids, in vitro-transcribed sequences, and cell extracts. In all of these experiments I have been able to effectively identify RNA sequences and their modifications.

NASE contains many unique functionalities that are not currently realized in other database search tools for RNA. To my knowledge, no other tools account for precursor mass defect resulting from instrumental selection of higher isotopologue peaks. This functionality is a major contributor to the excellent performance of NASE in identifying longer oligonucleotides compared to other database-matching tools. In addition, despite extensive desalting steps, cation adduction is always present to some extent in oligonucleotide mass spectra - NASE provides powerful correction for cation adduction events, which lessens the impact of sodium and potassium ions on sequence characterisation. In addition, OpenMS in general and NASE specifically were designed to be fast. My search times for complex samples are orders of magnitude faster than other tools. The searches on the NME1 and let-7 data take seconds; the much more complicated 20-modifications search of the tRNA dataset took 14 hours in multithreaded mode (20 parallel threads) on my server. For comparison, an analogous search using RNAModMapper was not feasible, with an estimated running time of one month. Equivalent searches with Ariadne did not return any modified oligonucleotides.

Not least, the open-source nature of OpenMS and NASE enables users to modify the software to fit their specific needs, to extend the existing functionality, and to create new interoperating programs. Already, many analysis tools have been implemented within the OpenMS framework for mass spectrometry-based proteomics and metabolomics experiments. The present work gives a foretaste of the power of leveraging these methods for the analysis of nucleic acid data. Future developments will streamline the use of OpenMS tools and algorithms, e.g. for quantification and comparisons across many samples, in the field of RNA epigenetics. In general, the development of NASE is an important step towards the large-scale analysis of RNA by mass spectrometry.

CHAPTER 4 : Putting it all together

4.1. Current directions

My technology is currently being put to good use in a couple of projects. Since these provide a good blueprint for where these techniques and software can go I will give a basic overview of the two ongoing projects which I am associated with.

Storm Therapeutics uses NASE to screen small-molecule RNA modifying enzyme inhibitors I developed NASE in concert with Hendrik Weisser from Storm Therapeutics. They needed a way to analyze the effectiveness of small-molecule inhibitors to RNA modifying enzymes. Existing technologies for assessing RNA modifications were limited to specific modifications, and many of them lacked the ability to determine what base of the RNA was being modified. Mass spectrometry serves their needs well, and NASE is part of their ongoing research.

The Garcia lab is looking for previously unidentified modifications to tRNA The tRNA data that we collected as part of testing NASE has been the subject of ongoing research into new, and previously unidentified tRNA modifications. By analysis of the data for mass shifts corresponding ot theorized but unconfirmed modification sites we are able to locate both new modifications and sites that were known to be modified, but for which the type of modification was unknown. This work is ongoing and we hope to be able to publish our results in the near future.

I am currently in discussion with Thermo Fisher Scientific to integrate NASE Thermo Fisher has expressed interest in building NASE into their commercial offerings.

4.2. Future directions

The work that I detail in this dissertation is a solid start for the field of high throughput RNA-modomics by mass spectrometry. From here there are a number of immediate ap-

plications for the system, as well as a number of improvements that I plan on making, to extend the abilities of the system.

Stable isotope labeling of nucleic acids Extending NASE to recognize pairs of heavy and light nucleic acids offers a potential new field of study. Pairs of samples, one labeled with a heavy isotope of a metabolic precursor, one not, is a well established experiment in metabolomics. So-called SILAC (stable isotope labeling of amino acids in cell culture) has been a valuable tool for proteomics and metabolomics in plumbing the depths of various metabolic pathways. Indeed we have a much deeper understanding of which metabolic pathways lead to which peptide modifications through SILAC mass spectrometry. An analogous method of study is potentially hugely valuable for determining the metabolic pathways tied to specific nucleic acid modifications.

Knock-out studies Another potential future direction for investigation is pursuing studies knocking out various RNA modifying enzymes, and then using my techniques and software to determine what modifications were caused by the enzymes. Through this method researchers would be able to determine whether specific enzymes have different location preferences for modifying RNA, and what modifications, if any are catalyzed by multiple different modifying enzymes. This technique could be further abstracted to working with knocking out other enzymes or proteins which are suspected to be involved in the metabolism of modification precursors. Knockouts could even be combined with stable isotope labeling for further targeting of our understanding of where modifications come from biochemically, and what proteins are involved in their formation.

Measuring modification stoichometery In chapter 3 I talk briefly about NASE's ability to generate label free quantification of modifications. This ability opens up a number of potential research paths. Previously to these innovations I was only able to determine whether a modification is present. Now I am able to determine both the presence of the modification, and how the proportion modified versus unmodified sequences of total nucleic

acid species changes between different samples. Equivalent studies in the field of histone epigenetic markers have shown that being able to determine the fold change in modification quantity gives key insight into what experimental conditions are associated with what modification changes.

Parallelization improvements NASE performs very well and quickly on samples with moderate complexity, unfortunately the addition of more potential modifications to be searched causes an exponential growth in the amount of work the software needs to perform. During the early analysis of the tRNA samples with more than 20 potential modifications and up to three modifications on each sequence, runs could last up to two weeks running single threaded. The current versions of NASE have multithreading support which decreases the time to complete the above search in a matter of hours. Thus far I have not explored the option to further spread out the processing of exceptionally large datasets to multiple processors, or further to be able to run on cluster computers. As researchers use this software for larger and more complicated datasets I predict that implementing multiprocessor and cluster support in NASE will improve its overall usability and allow for even more cutting edge experiments.

Moving away from ion-pair chromatography While I have managed to use ion-pair chromatography to separate different nucleic acids it remains a less than optimal separation mechanism, both because of instrument contamination as well as ion suppression. Experiments by Garcia lab member Richard Lauman in using charged porous-graphitic-carbon columns to separate nucleic acids have been encouraging and demonstrate a potential alternative to ion-pair chromatography. In brief, we have found that nucleic acids can be retained on a porous-graphitic-carbon column without the aid of ion-pair reagents. This is the first substrate that we are aware of which retains nucleic acids. We found that the nucleic acids can then be eluted from the column through a combination of applying an electrical charge to the column, and flushing with an organic solvent. Initial experiments have suffered from breakdown of nucleic acids on the column into dimers and trimers, making this technique

not quite ready for use in my applications, however we hope that by adjusting the voltage across the column that we will be able to preserve longer nucleic acids. Given that using porous-graphitic-caron columns with an attached power supply is a lot less expensive than dedicating a high-performance liquid chromatograph and a dedicated mass spectrometer I am hopeful that further study of this technique will increase the accessibility of nucleic acid mass spectrometry to many more labs and researchers.

Better monoisotopic peak selection The mass spectrometers used in my experiments can select ions of interest for fragmentation in a number of different ways. For all of the work here I used a so-called data dependent acquisition (DDA), where the most intense ions for which a charge state could be determined were collected for fragmentation and acquisiton of tandem mass spectra. The instrument software provided along with our instruments allows for a couple of alternative acquisition methods. The simplest of these is simply feeding a list of masses to fragment when ions of that mass are detected. This selection method is useful when searching against a small database of known nucleic acid sequences, but it does not scale to samples for which the experimenter doesn't know all of the sequences ahead of time. For future work I propose adding DDA mode which takes both the charge state of the precursor as well as the isotopic abundance of the peaks in the isotopic envelope. Since nucleic acids have an elemental distribution which is known, and each element has an isotopic distribution which is known, an approximate theoretical isotopic distribution for a nucleic acid of a known mass can be calculated. I have already done this as part of the my additions to `FeatureFinderMultiplex`. An improvement for the future would be to work with the instrument manufacturers to include this feature as part of the selection criteria for tandem mass spectra. I anticipate that such an improvement would greatly help with selecting nucleic acids as opposed to other background in complicated samples.

4.3. Conclusions

Mass spectrometry is a viable and rich platform for the analysis of nucleic acid modifications. Separation and ionization efficiency issues plagued the early iterations of the system,

thankfully the use of the commercially available easySpray system coupled with having a dedicated instrument for RNA have made the analytical method repeatable, stable, and easy for other groups to replicate. The software platform has matured to the point at which it is poised for integration into the standard OpenMS build. The experiments that I have conducted have shown the software to be functional across a wide variety of nucleic acid inputs, and able to handle the complexity of cellular samples. All of these parts taken together speak to a bright future for nucleic acid mass spectrometry. By making these tools available I hope that more labs will be able to partake in this interesting and growing field. Likewise I think this project has great potential to continue to be expanded in the directions I have mentioned above (among others) and I look forward to seeing what this nascent field has in store for it.

BIBLIOGRAPHY

M. Abe, A. Naqvi, G.-J. Hendriks, V. Feltzin, Y. Zhu, A. Grigoriev, and N. Bonini. Impact of age-associated increase in 2'-o-methylation of miRNAs on aging and neurodegeneration in drosophila. *Genes and Development*, 28(1):44–57, 2014. ISSN 0890-9369. doi: 10.1101/gad.226654.113.

S. S. Abedini, K. Kahrizi, L. R. de Pouplana, and H. Najmabadi. trna methyltransferase defects and intellectual disability. *Archives of Iranian Medicine (AIM)*, 21(10), 2018.

K. A. Afonin, M. Kireeva, W. W. Grabow, M. Kashlev, L. Jaeger, and B. A. Shapiro. Co-transcriptional assembly of chemically modified rna nanoparticles functionalized with sirnas. *Nano letters*, 12(10):5192–5195, 2012.

A. Apffel, J. A. Chakel, S. Fischer, K. Lichtenwalter, and W. S. Hancock. Analysis of Oligonucleotides by HPLCElectrospray Ionization Mass Spectrometry. *Analytical Chemistry*, 69(7):1320–1325, Apr. 1997. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac960916h. URL `http://pubs.acs.org/doi/abs/10.1021/ac960916h`.

L. Balcaen, E. Bolea-Fernandez, M. Resano, and F. Vanhaecke. Inductively coupled plasma–tandem mass spectrometry (icp-ms/ms): A powerful and universal tool for the interference-free determination of (ultra) trace elements–a tutorial review. *Analytica chimica acta*, 894:7–19, 2015.

M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*, 404(4):939–965, 2012.

D. P. Bartel. Microrna: genomics, biogenesis, mechanism, and function. *Cell*, 116(2): 281–297, 2004.

A. Bertero, S. Brown, P. Madrigal, A. Osnato, D. Ortmann, L. Yiangou, J. Kadiwala, N. C. Hubner, I. R. de los Mozos, C. Sade, A.-S. Lenaerts, S. Nakanoh, R. Grandy, E. Farnell, J. Ule, H. G. Stunnenberg, S. Mendjan, and L. Vallier. The SMAD2/3 interactome reveals that TGF controls m6a mRNA methylation in pluripotency. *Nature*, 555(7695): 256–259, Feb. 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature25784. URL `http://www.nature.com/doifinder/10.1038/nature25784`.

S. Blanco, S. Dietmann, J. V. Flores, S. Hussain, C. Kutter, P. Humphreys, M. Lukk, P. Lombard, L. Treps, M. Popis, S. Kellner, S. M. Holter, L. Garrett, W. Wurst, L. Becker, T. Klopstock, H. Fuchs, V. Gailus-Durner, M. Hrabe de Angelis, R. T. Karadottir, M. Helm, J. Ule, J. G. Gleeson, D. T. Odom, and M. Frye. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *The EMBO Journal*, 33 (18):2020–2039, Sept. 2014. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.201489282. URL `http://emboj.embopress.org/cgi/doi/10.15252/embj.201489282`.

P. Boccaletto, M. A. Machnicka, E. Purta, P. Pitkowski, B. Bagiski, T. K. Wirecki,

V. de Crcy-Lagard, R. Ross, P. A. Limbach, A. Kotter, M. Helm, and J. M. Bujnicki. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Research*, 46(D1):D303–D307, Jan. 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1030. URL https://academic.oup.com/nar/article/46/D1/D303/4584632.

E. Borek and P. R. Srinivasan. The Methylation of Nucleic Acids. *Annual Review of Biochemistry*, 35(1):275–298, June 1966. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev.bi.35.070166.001423. URL http://www.annualreviews.org/doi/10.1146/annurev.bi.35.070166.001423.

B. Brzezicha, M. Schmidt, I. Makaowska, A. Jarmoowski, J. Piekowska, and Z. Szweykowska-Kuliska. Identification of human tRNA:m5c methyltransferase catalysing intron-dependent m5c formation in the first position of the anticodon of the pre-tRNA(CAA)Leu. *Nucleic Acids Research*, 34(20):6034–6043, Nov. 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl765. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1635329/.

S. Bustin, V. Benes, T. Nolan, and M. Pfaffl. Quantitative real-time rt-pcr–a perspective. *Journal of molecular endocrinology*, 34(3):597–601, 2005.

M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick. A Cross-platform Toolkit for Mass Spectrometry and Proteomics. *Nature biotechnology*, 30(10):918–920, Oct. 2012. ISSN 1087-0156. doi: 10.1038/nbt.2377. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3471674/.

C. S. Chow, T. N. Lamichhane, and S. K. Mahto. Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS chemical biology*, 2(9):610–619, 2007.

J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, Dec. 2008. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.1511. URL http://www.nature.com/articles/nbt.1511.

F. F. Davis and F. W. Allen. Ribonucleic acids from yeast which contain a fifth nucleotide. *The Journal of Biological Chemistry*, 227(2):907–915, Aug. 1957. ISSN 0021-9258.

W. A. Decatur and M. J. Fournier. rrna modifications and ribosome function. *Trends in biochemical sciences*, 27(7):344–351, 2002.

A. El-Aneed, A. Cohen, and J. Banoub. Mass spectrometry, review of the basics: electro-

spray, maldi, and commonly used mass analyzers. *Applied Spectroscopy Reviews*, 44(3): 210–230, 2009.

J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

W. V. Gilbert, T. A. Bell, and C. Schaening. Messenger rna modifications: form, distribution, and function. *Science*, 352(6292):1408–1412, 2016.

L. C. Gillet, A. Leitner, and R. Aebersold. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annual Review of Analytical Chemistry*, 9(1):449–472, 2016. doi: 10.1146/annurev-anchem-071015-041535. URL `https://doi.org/10.1146/annurev-anchem-071015-041535`.

R. Green and H. F. Noller. In vitro complementation analysis localizes 23s rrna post-transcriptional modifications that are required for escherichia coli 50s ribosomal subunit assembly and function. *Rna*, 2(10):1011–1021, 1996.

R. I. Gregory, K.-p. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. The microprocessor complex mediates the genesis of micrornas. *Nature*, 432(7014):235, 2004.

J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q.-W. Xu, N. d. Toro, Y. Prez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcano, and H. Hermjakob. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775, Oct. 2014. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.O113.036681. URL `http://www.mcponline.org/content/13/10/2765`.

S. Å. Gustavsson, J. Samskog, K. E. Markides, and B. Långström. Studies of signal suppression in liquid chromatography–electrospray ionization mass spectrometry using volatile ion-pairing reagents. *Journal of Chromatography A*, 937(1-2):41–47, 2001.

S. M. Hammond. An overview of microrna. *Advanced drug delivery reviews*, 87:3–14, 2015.

J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim. The drosha-dgcr8 complex in primary microrna processing. *Genes & development*, 18(24):3016–3027, 2004.

M. Helm and Y. Motorin. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews. Genetics*, 18(5):275–291, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.169.

M. Hossain and P. A. Limbach. Mass spectrometry-based detection of transfer rnas by their signature endonuclease digestion products. *RNA*, 13(2):295–303, 2007.

C. G. Huber and A. Krajete. Sheath liquid effects in capillary high-performance liquid chromatography–electrospray mass spectrometry of oligonucleotides. *Journal of Chromatography A*, 870(1-2):413–424, 2000.

C. G. Huber and H. Oberacher. Analysis of nucleic acids by on-line liquid chromatographyMass spectrometry. *Mass Spectrometry Reviews*, 20(5):310–343, Jan. 2001. ISSN 1098-2787. doi: 10.1002/mas.10011. URL `http://onlinelibrary.wiley.com/doi/10.1002/mas.10011/abstract`.

P. Hupé. Mass spectrometry protocol, 2012. URL `https://commons.wikimedia.org/wiki/File:Mass_spectrometry_protocol.svg`. (CC BY-SA 3.0).

G. Jia, Y. Fu, and C. He. Reversible RNA adenosine methylation in biological regulation. *Trends in Genetics*, 29(2):108–115, Feb. 2013. ISSN 01689525. doi: 10.1016/j.tig.2012.11.003. URL `http://linkinghub.elsevier.com/retrieve/pii/S0168952512001874`.

J. Junker, C. Bielow, A. Bertsch, M. Sturm, K. Reinert, and O. Kohlbacher. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *Journal of Proteome Research*, 11(7):3914–3920, July 2012. ISSN 1535-3907. doi: 10.1021/pr300187f.

E. Junowicz and J. H. Spencer. Rapid separation of nucleosides and nucleotides by cation-exchange column chromatography. *Journal of Chromatography A*, 44:342–348, Jan. 1969. ISSN 0021-9673. doi: 10.1016/S0021-9673(01)92545-2. URL `http://www.sciencedirect.com/science/article/pii/S0021967301925452`.

F. Jhling, M. Mrl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Ptz. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(suppl_1): D159–D162, Jan. 2009. ISSN 0305-1048. doi: 10.1093/nar/gkn772. URL `https://academic.oup.com/nar/article/37/suppl_1/D159/1009888`.

D. Kierzkowski, M. Kmieciak, P. Piontek, P. Wojtaszek, Z. Szweykowska-Kulinska, and A. Jarmolowski. The arabidopsis cbp20 targets the cap-binding complex to the nucleus, and is stabilized by cbp80. *The Plant Journal*, 59(5):814–825, 2009.

Y. Kirino, T. Yasukawa, S. Ohta, S. Akira, K. Ishihara, K. Watanabe, and T. Suzuki. Codon-specific translational defect caused by a wobble modification deficiency in mutant tRNA from a human mitochondrial disease. *Proceedings of the National Academy of Sciences*, 101(42):15070–15075, Oct. 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0405173101. URL `http://www.pnas.org/content/101/42/15070`.

M. Kullolli, E. Knouf, M. Arampatzidou, M. Tewari, and S. J. Pitteri. Intact MicroRNA analysis using high resolution mass spectrometry. *Journal of The American Society for Mass Spectrometry*, 25(1):80–87, 2014. ISSN 1044-0305, 1879-1123. doi: 10.1007/s13361-013-0759-x. URL `http://link.springer.com/article/10.1007/s13361-013-0759-x`.

L. Kll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning Significance to Peptides

Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, Jan. 2008. ISSN 1535-3893, 1535-3907. doi: 10.1021/pr700600n. URL http://pubs.acs.org/doi/abs/10.1021/pr700600n.

R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *cell*, 75(5):843–854, 1993.

J. Li, Z. Yang, B. Yu, J. Liu, and X. Chen. Methylation protects mirnas and sirnas from a 3-end uridylation activity in arabidopsis. *Current biology*, 15(16):1501–1507, 2005.

X. Li, X. Xiong, and C. Yi. Epitranscriptome sequencing technologies: decoding RNA modifications. *Nature Methods*, 14(1):23–31, Dec. 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4110. URL http://www.nature.com/doifinder/10.1038/nmeth.4110.

Z. J. Lin, W. Li, and G. Dai. Application of LCMS for quantitative analysis and metabolite identification of therapeutic oligonucleotides. *Journal of Pharmaceutical and Biomedical Analysis*, 44(2):330–341, 2007. ISSN 0731-7085. doi: 10.1016/j.jpba.2007.01.042. URL http://www.sciencedirect.com/science/article/pii/S0731708507000787.

B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey. Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome. *Nature Methods*, 12(8):767–772, Aug. 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3453.

C. Llave, K. D. Kasschau, M. A. Rector, and J. C. Carrington. Endogenous and silencing-associated small rnas in plants. *The Plant Cell*, 14(7):1605–1619, 2002.

P. A. Lobue, N. Yu, M. Jora, S. Abernathy, and P. A. Limbach. Improved application of RNAModMapper - An RNA modification mapping software tool - For analysis of liquid chromatography tandem mass spectrometry (LC-MS/MS) data. *Methods (San Diego, Calif.)*, Oct. 2018. ISSN 1095-9130. doi: 10.1016/j.ymeth.2018.10.012.

A. F. Lovejoy, D. P. Riordan, and P. O. Brown. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mrnas in s. cerevisiae. *PloS one*, 9 (10):e110799, 2014.

M. A. Machnicka, A. Olchowik, H. Grosjean, and J. M. Bujnicki. Distribution and frequencies of post-transcriptional modifications in tRNAs. *RNA Biology*, 11(12):1619–1629, Jan. 2015. ISSN 1547-6286. doi: 10.4161/15476286.2014.992273. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4615829/.

I. J. MacRae, K. Zhou, and J. A. Doudna. Structural determinants of rna recognition and cleavage by dicer. *Nature structural & molecular biology*, 14(10):934, 2007.

L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rmpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch. mzMLa Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics*, 10(1):

R110.000133, Jan. 2011. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.R110.000133. URL http://www.mcponline.org/content/10/1/R110.000133.

S. A. Mcluckey, G. J. Berkel, and G. L. Glish. Tandem Mass Spectrometry of Small, Multiply Charged Oligonucleotides. *Journal of the American Society for Mass Spectrometry*, 3(1): 60–70, Jan. 1992. ISSN 1044-0305, 1879-1123. doi: 10.1016/1044-0305(92)85019-G. URL http://link.springer.com/10.1016/1044-0305(92)85019-G.

Y. Motorin and H. Grosjean. Multisite-specific tRNA:m5c-methyltransferase (Trm4) in yeast Saccharomyces cerevisiae: identification of the gene and substrate specificity of the enzyme. *RNA (New York, N.Y.)*, 5(8):1105–1118, Aug. 1999. ISSN 1355-8382.

H. Nakayama, M. Akiyama, M. Taoka, Y. Yamauchi, Y. Nobe, H. Ishikawa, N. Takahashi, and T. Isobe. Ariadne: a database search engine for identification and chemical analysis of RNA using tandem mass spectrometry data. *Nucleic Acids Research*, 37(6):e47–e47, Apr. 2009. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkp099. URL http://nar.oxfordjournals.org/content/37/6/e47.

Narayanese. Mirna-biogenesis, 2012. URL https://en.wikipedia.org/wiki/MicroRNA#/media/File:MiRNA-biogenesis.jpg. (CC BY-SA 3.0).

H. F. Noller. Structure of ribosomal rna. *Annual review of biochemistry*, 53(1):119–162, 1984.

A. Nyakas, L. C. Blum, S. R. Stucki, J.-L. Reymond, and S. Schrch. OMA and OPASoftware-Supported Mass Spectra Analysis of Native and Modified Nucleic Acids. *Journal of The American Society for Mass Spectrometry*, 24(2):249–256, Dec. 2012. ISSN 1044-0305, 1879-1123. doi: 10.1007/s13361-012-0529-1. URL http://link.springer.com/article/10.1007/s13361-012-0529-1.

T. Pan. Modifications and functional genomics of human transfer rna. *Cell research*, page 1, 2018a.

T. Pan. Modifications and functional genomics of human transfer RNA. *Cell Research*, 28(4):395–404, Apr. 2018b. ISSN 1748-7838. doi: 10.1038/s41422-018-0013-y. URL https://www.nature.com/articles/s41422-018-0013-y.

B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906, Feb. 2000. ISSN 0028-0836. doi: 10.1038/35002607.

H. L. Rost, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmstrom, R. Aebersold, K. Reinert, and O. Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrom-

etry data analysis. *Nature Methods*, 13(9):741–748, 2016. ISSN 1548-7091. URL http://dx.doi.org/10.1038/nmeth.3959.

J. Rozenski and J. A. McCloskey. SOS: A simple interactive program for ab initio oligonucleotide sequencing by mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 13(3):200–203, Mar. 2002. ISSN 1044-0305, 1879-1123. doi: 10.1016/S1044-0305(01)00354-3. URL http://link.springer.com/article/10.1016/S1044-0305%2801%2900354-3.

W. Sahib. Ribonucleic acid chemical structure, 2014. URL https://commons.wikimedia.org/wiki/File:Ribonucleic_acid_chemical_structure.svg. CC BY-SA 3.0.

Sakurambo. Stem-loop, 2006. URL https://commons.wikimedia.org/wiki/File:Stem-loop.svg. CC BY-SA 3.0.

S. J. Sharp, J. Schaack, L. Cooley, D. J. Burke, and D. Sll. Structure and transcription of eukaryotic tRNA genes. *CRC critical reviews in biochemistry*, 19(2):107–144, 1985. ISSN 0045-6411.

S. K. Shenouda and S. K. Alahari. Microrna function in cancer: oncogene or a tumor suppressor? *Cancer and Metastasis Reviews*, 28(3-4):369, 2009.

M. Sturm and O. Kohlbacher. TOPPView: An Open-Source Viewer for Mass Spectrometry Data. *Journal of Proteome Research*, 8(7):3760–3763, July 2009. ISSN 1535-3893. doi: 10.1021/pr900171m. URL https://doi.org/10.1021/pr900171m.

D. Su, C. T. Y. Chan, C. Gu, K. S. Lim, Y. H. Chionh, M. E. McBee, B. S. Russell, I. R. Babu, T. J. Begley, and P. C. Dedon. Quantitative analysis of ribonucleoside modifications in tRNA by HPLC-coupled mass spectrometry. *Nature Protocols*, 9(4):828–841, Apr. 2014. ISSN 1754-2189. doi: 10.1038/nprot.2014.047. URL http://www.nature.com/nprot/journal/v9/n4/full/nprot.2014.047.html.

P. Svoboda and A. D. Cara. Hairpin rna: a secondary structure of primary importance. *Cellular and Molecular Life Sciences CMLS*, 63(7-8):901–908, 2006.

M. Taoka, Y. Nobe, M. Hori, A. Takeuchi, S. Masaki, Y. Yamauchi, H. Nakayama, N. Takahashi, and T. Isobe. A mass spectrometry-based method for comprehensive quantitative determination of post-transcriptional RNA modifications: the complete chemical structure of schizosaccharomyces pombe ribosomal RNAs. *Nucleic Acids Research*, page gkv560, 2015. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv560. URL http://nar.oxfordjournals.org/content/early/2015/05/26/nar.gkv560.

M. Taoka, Y. Nobe, Y. Yamaki, Y. Yamauchi, H. Ishikawa, N. Takahashi, H. Nakayama, and T. Isobe. The complete chemical structure of saccharomyces cerevisiae rrna: partial pseudouridylation of u2345 in 25s rrna by snorna snr9. *Nucleic acids research*, 44(18):8951–8961, 2016a.

M. Taoka, Y. Nobe, Y. Yamaki, Y. Yamauchi, H. Ishikawa, N. Takahashi, H. Nakayama, and T. Isobe. The complete chemical structure of *Saccharomyces cerevisiae* rRNA: partial pseudouridylation of U2345 in 25s rRNA by snoRNA snR9. *Nucleic Acids Research*, 44 (18):8951–8961, Oct. 2016b. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkw564. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw564.

R. C. Team. R: A language and environment for statistical computing, 2018. URL https://www.R-project.org/.

J. J. Thomson. *Rays of positive electricity and their application to chemical analyses*, volume 1. Longmans, Green and Company, 1921.

Vossman. Ribosome shape, 2009. URL https://commons.wikimedia.org/wiki/File:Ribosome_shape.png. (CC BY-SA 3.0).

J.-X. Wang, J. Gao, S.-L. Ding, K. Wang, J.-Q. Jiao, Y. Wang, T. Sun, L.-Y. Zhou, B. Long, X.-J. Zhang, Q. Li, J.-P. Liu, C. Feng, J. Liu, Y. Gong, Z. Zhou, and P.-F. Li. Oxidative modification of miR-184 enables it to target bcl-xL and bcl-w. *Molecular Cell*, 59(1):50–61, 2015. ISSN 1097-2765. doi: 10.1016/j.molcel.2015.05.003. URL http://www.sciencedirect.com/science/article/pii/S1097276515003391.

S. Wein. Reverse transcription polymerase chain reaction, 2019. URL https://upload.wikimedia.org/wikipedia/commons/1/18/Reverse_transcription_polymerase_chain_reaction.svg. Adapted from Jpark623 (CC BY-SA 3.0).

H. Weisser and J. S. Choudhary. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *Journal of Proteome Research*, 16(8):2964–2974, Aug. 2017. ISSN 1535-3907. doi: 10.1021/acs.jproteome.7b00248.

Y.-L. Weng, X. Wang, R. An, J. Cassin, C. Vissers, Y. Liu, Y. Liu, T. Xu, X. Wang, S. Z. H. Wong, J. Joseph, L. C. Dore, Q. Dong, W. Zheng, P. Jin, H. Wu, B. Shen, X. Zhuang, C. He, K. Liu, H. Song, and G.-l. Ming. Epitranscriptomic m 6 A Regulation of Axon Regeneration in the Adult Mammalian Nervous System. *Neuron*, 97(2):313–325.e6, Jan. 2018. ISSN 08966273. doi: 10.1016/j.neuron.2017.12.036. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627317311844.

J. A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, 52(2-3):337–349, 1983.

Yikrazuul. Trna-phe yeast en, 2010. URL https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_en.svg. (CC BY-SA 3.0).

B. Yu, Z. Yang, J. Li, S. Minakhina, M. Yang, R. W. Padgett, R. Steward, and X. Chen. Methylation as a crucial step in plant microrna biogenesis. *Science*, 307(5711):932–935, 2005.

N. Yu, P. A. Lobue, X. Cao, and P. A. Limbach. RNAModMapper: RNA Modification

Mapping Software for Analysis of Liquid Chromatography Tandem Mass Spectrometry Data. *Analytical Chemistry*, 89(20):10744–10752, Oct. 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b01780. URL https://doi.org/10.1021/acs.analchem.7b01780.

Y. Yu, M. Smith, and R. Pieper. A spinnable and automatable stagetip for high throughput peptide desalting and proteomics. *Protocol Exchange*, 2014.