From whole-cell to compartment- and class-specific analysis of RNA post-transcriptional modifications

Ryan Quinn
University at Albany, State University of New York, rmquinn@albany.edu

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Chemistry Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/988

This Master's Thesis is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
From whole-cell to compartment- and class-specific analysis of RNA post-transcriptional modifications.

by
Ryan Quinn

A Thesis
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Master of Science

College of Arts & Sciences
Department of Chemistry
2013
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>iv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>iv</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>iv</td>
</tr>
<tr>
<td>Abstract</td>
<td>vi</td>
</tr>
<tr>
<td>Aims</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>14</td>
</tr>
<tr>
<td>Conclusions</td>
<td>46</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>48</td>
</tr>
<tr>
<td>References</td>
<td>62</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost I would like to thank my research advisor, Dr. Daniele Fabris, for his continual support and encouragement through my graduate studies.

I am indebted to the members of the Fabris Research Group, past and present, for their patience and support. In particular, I would like to thank Dr. John Mangrum, Dr. Maria Basanta Sanchez, and Rebecca Rose for their help with instrumentation and method development, and Jaclyn Sayre and Prince Jacob for their dedication and hard work.

I would like to thank our collaborators, past and present, for their help and support. In particular, Dr. Mario Canki was an invaluable resource in our virology work. Dr. Robert Osuna, Dr. Marlene Belfort, and Dr. Alexander Shekhtman also contributed necessary bacterial samples to my research.

The support of the University at Albany is also gratefully acknowledged. This work was supported by the National Institutes of Health (grant GM064328-11).
List of Figures
1. The Gower Distance (p. 12)
2. Full Spectrum of NMP/dNMP Standards in Ion Trap vs. Orbitrap (p.15)
3. Tandem Mass Spectrometric Analysis of AMP/dGMP Isobar Pair (p.17)
4. Tandem Mass Spectrometric Analysis of UMP/ΨMP Isobar Pair (p. 19)
5. Ion Mobility-Mass Spectrometric Analysis of Isobaric Species (p. 21)
6. Total Mobility Heat Map of NMP/dNMP Standards (p. 23)
7. Time-Alignment Parallel Fragmentation of NMP/dNMP Standards (p. 24)
8. Tandem Mass Spectrometric Analysis of Modified NMPs I (p. 26)
9. Tandem Mass Spectrometric Analysis of Modified NMPs II (p. 27)
10. Ion Mobility-Mass Spectrometric Analysis of Modified NMPs (p. 28)
11. High-Performance Liquid Chromatography Analysis of Isobaric NMPs (p. 30)
12. Assessment of Orbitrap Performance Versus Number of Ions Scanned (p. 32)
13. Escherichia Coli Full Mass Spectrum (p. 33)
14. Results of Ultracentrifugation Experiments (p. 40)
15. Isolation of RNA from Viral Samples (p. 41)
16. Development of an Affinity Capture Procedure (p. 42)
17. Full Mass Spectrum of HIV-1 Samples (p. 44)
18. Development of Cross-Linking Procedures (p. 45)
19. Example of Bacterial Phylogeny (p. 53)
20. Formation of Affinity Capture Beads (p. 59)

List of Tables
1. Exact Masses of Ribonucleotides and Deoxyribonucleotides (p. 14)
2. Exact Masses of Select Modified Ribonucleotides (p. 16)
3. List of Detected Ribonucleotide Modifications Under Select Metabolic Environments in Escherichia Coli (p. 36)
4. Relatedness of Gammaproteo Bacteria Samples Based on Ribonucleotide Modification Patterns (p. 37)
5. Modified Nucleotides Detected in the Human Immunodeficiency Virus (p. 43)
6. List of Bacterial and Yeast Species Investigated (p. 53)

List of Abbreviations
1. AGC: Automatic Gain Control
2. BME: Beta-mercapto ethanol
3. cDNA: Complementary Deoxyribonucleic Acid
4. CGE: Capillary Gel Electrophoresis
5. CID: Collisionally Induced Dissociation
6. CIP: Calf intestine phosphatase
7. CRM: Consecutive Reaction Monitoring
8. DEA: Diethylamine
9. DEPC: Diethylpyrocarbonate
10. DMEM: Dulbecco’s Modified Eagle’s Medium
11. DMF: Dimethyl formamide
12. DNA: Deoxyribonucleic Acid
13. DTNB: 5,5'-dithiobis-(2-nitrobenzoic acid)
14. DTT: Dithiothreitol
15. EDAC: 1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide
16. EDTA: Ethylene Diamine Tetraacetic Acid
17. FBS: Fetal Bovine Serum
18. FTMS: Fourier-Transform Mass Spectrometry
19. HCD: Higher-Energy Collisionally Induced Dissociation
20. HDMS: High-Definition Mass Spectrometry
21. HIV-1: Human Immunodeficiency Virus, Type 1
22. HPLC: High-Performance Liquid Chromatography
23. IM-MS: Ion Mobility-Mass Spectrometry
24. IMS: Ion Mobility Spectroscopy
25. IP: Immunoprecipitation
26. IPA: Isopropanol
27. IPTG: Isopropyl B-d-1-Thiogalactopyranoside
28. LOD: Limit of Detection
29. LTQ: Linear Trap Quadripole Mass Spectrometer
30. miRNA: Micro Ribonucleic Acid
31. mRNA: Messenger Ribonucleic Acid
32. MS: Mass Spectrometry
33. MS3D: 3-D Mass Spectrometry
34. NMP: Nucleoside Monophosphate
35. NMR: Nuclear Magnetic Resonance
36. NTP: Nucleoside Triphosphate
37. pAGC: Predictive Automatic Gain Control
38. PBS: Phosphate Buffered Saline
39. PQD: Pulsed-Q Dissociation
40. RISC: RNA-Induced Silencing Complex
41. RNA: Ribonucleic Acid
42. RNAP: Ribonucleic Acid Polymerase
43. RNase: Ribonuclease
44. RNP: Ribonucleic Protein Particle
45. RP-HPLC: Reversed-Phase High Performance Liquid Chromatography
46. rRNA: Ribosomal Ribonucleic Acid
47. RSD: Relative Standard Deviation
48. RT-PCR: Reverse-Transcriptase Polymerase Chain Reaction
49. SDS: Sodium Dodacyl Sulfate
50. siRNA: Small Interfering RNA
51. SNR: Signal-to-Noise Ratio
52. SRM: Selected Reaction Monitoring
53. TAE: Tris-Acetic Acid-EDTA Buffer
54. TAP: Time-Alignment Parallel Fragmentation
55. TBE: Tris-Boric Acid-EDTA Buffer
56. TE: Tris-EDTA Buffer
57. TOF: Time-of-Flight
58. tRNA: Transfer RNA
59. UCF: Ultracentrifugation
60. UTR: Untranslated Region
61. VSV-G: Vesicular Stomatitis Virus
Abstract

The genomic structure of the human immunodeficiency virus type 1 (HIV-1) has proven difficult to examine by traditional high-resolution techniques. Structural probing and in silico modeling have been turned to the task instead. Computer models are only as robust as the data used to create them. A number of techniques were explored in this investigation to expand the utility of in silico modeling as it relates to HIV-1: the study of nucleoside modifications, and the probing of in vivo conformations of the 5′-untranslated region (5′-UTR) by cross-linking and affinity capture.

The biological significance of RNA nucleoside modifications is poorly understood due to the dearth of analytical strategies for studying their cellular distribution and fate. Modified bases are not detectable by strategies based on the replication of template strands. In contrast, their presence can be readily recognized by mass spectrometry (MS) from their unique mass shifts. This analytical platform has played a major role in the identification of known RNA modifications and has shown potential for the investigation of their biological significance. The application of high-resolution mass spectrometry to the genome-wide analysis of modified ribonucleotides was explored in the context of HIV-1 structural elucidation and of diagnostic and phylogenetic purposes. The data indicate the applicability of MS to the transcriptome-level analysis of ribonucleotide modifications for all avenues explored. The techniques employed allowed for the identification of previously unknown modifications in HIV-1, and the differentiation of several human and bacterial samples.

Cross-linking and affinity purification were explored as a means to determine the structure of the HIV-1 5′-UTR, which is less prone to mutation than the rest of the virus due to the presence of many vital structural elements and thus presents a
favorable putative drug target. Preliminary data have shown only limited success of these techniques in a biological context.
Aims

The goal of this project was to develop a comprehensive experimental workflow for the detection and quantification of in vivo RNA covalent modifications, which employed mass spectrometry without front-end separation. The modifications considered were either natural, part of the normal cellular content, or man-made, generated by the application of footprinting and crosslinking agents. The significance of this approach rests on the possibility of using the former as biomarkers of disease and the latter as structural constrains for producing rigorous 3D models in silico. For this reason, we sought to obtain proof of principle by tackling the transcriptome-wide analysis of natural modifications in a variety of bacteria and eukaryotic cells. We have also initiated the exploration of methods for crosslinking and purifying HIV-1 RNA from viral particles and infected human lymphocytes, with the goal of enabling the structural elucidation of the HIV-1 5′-untranslated region (5′-UTR).

Introduction

The Role of RNA in Living Cells: Life on earth is a result of the complex interplay of large biopolymers and small molecules. Three of the most important biopolymers, DNA, RNA, and proteins, are linked by the Central Dogma of Molecular Biology. DNA is transcribed to RNA, which is translated into proteins, which are the functional agents of the cellular environment. In the early 1980s, the discovery of catalytically active RNA (ribozymes) challenged this idea [1], and helped to lead to the formation of the RNA world hypothesis [2] in 1986. It was later discovered the DNA could also perform catalytic functions, with the discovery of deoxyribozymes by Breaker in 1994 [3].

Like any polymer, the properties of nucleic acids are determined primarily by the nature of their constituent monomers. Nucleotides are the building blocks of
nucleic acids, and as such are the foundation for genetics and transcriptomics. Nucleotides can also serve as cofactors, energy carriers, metabolic regulators, second messengers, and vitamin components [4-7], making them vitally important to a host of metabolic processes. As such, nucleotides are often modified for a variety of different reasons. Deoxyribonucleotides are often modified by damage processes, exposure to mutagens, and as part of DNA replication [8-14], which makes the assessment of said modifications important for the monitoring of biological processes. RNA modification plays an even larger, albeit currently less diagnostic, role in cellular function. Over 110 ribonucleoside modifications have been catalogued to date [15]. Most are associated with stable RNAs (ribosomal and transfer), and with specialized transient RNA classes (such as siRNA) [16, 17]. As such these modifications are suspected to play a role in the structure [18, 19], function [17, 20, 21] and dynamics [22] of the parent RNA. In some cases, RNA modification is suspected to have been integral to the origin of life [23]. Much is still not understood of the significance and distribution of these modifications, however, due in large part to the dearth of techniques available to study them.

**Transcriptome Analysis:** Investigation of the transcriptome has conventionally been the purview of reverse-transcription polymerase chain reaction (RT-PCR) [24], which allows for the creation, amplification, and analysis of the complementary DNA (cDNA) sequences. RNA modification cannot be assessed by RT-PCR, as there are no set complementary bases for modified nucleosides. Indeed, it has been demonstrated [25] that RNA modification will cause RT-PCR to fail in many instances. Thus another technique is needed.

Mass spectrometry has been used to study nucleosides for 50 years [26], and its applicability to the field has only increased with the evolution of new ionization and detection techniques [27-30]. A common feature to all attempts to study
nucleic acid building blocks, however, has been the degradation of the polymer into its monomeric units by enzymatic hydrolysis (i.e. with exonucleases), and possible reduction to nucleosides, depending on the ionization techniques available [29, 31]. Typically, a method of front-end or off-line separation (high-performance liquid chromatography [HPLC] or capillary electrophoresis) has been employed to purify and concentrate samples [32, 33].

Recent advances in the field of mass spectrometry have provided a strong platform for analysis of nucleic acid monomers, even without front-end separation. High-resolution Fourier-transform mass spectrometry (FTMS) [34-36] and ion-mobility mass spectrometers (IM-MS) [37-39] provide complementary gas-phase data related to the identification of analytes of similar or identical chemical composition.

The high resolving power, low error (sub ppm) and very low limit of detection (picomolar) [36, 40] of the Orbitrap provides a means to investigate complex mixtures and unambiguously identify small molecules and metabolites [41, 42]. The ability of the Linear Trap Quadrupole (LTQ)-Orbitrap to perform multiple rounds of tandem MS allows for the direct investigation of the structure of target analytes [43].

IMS-MS instruments investigate the 3-D structure of a molecule by measuring the drift time of a molecule across a pressurized region of the instrument while in an electric field [44-48]. Since drift time is related to the number of interactions between the analyte and the background gas, which is in turn related to the average cross-section of the molecule, it can be used to differentiate different isomers or conformers in a mixture of isobaric species [48, 49]. Combined, these techniques afford a great potential for the differentiation of nucleic acid building blocks and the eventual analysis of small molecule biomarkers.
**Challenges to Working with Cellular RNA:** The study of RNA modification is, in general, only of interest inasmuch as it can be applied to living cells. In particular, it may have applicability in realms where other epigenetic methods have struggled, such as cancer biomarker analysis [50] and bacterial phylogeny [51-55].

Analysis of RNA from cellular extracts poses many challenges not seen in *in vitro* studies. The homogeneity, integrity, subcellular localization, and fractionation by type must all be considered.

RNA extraction from living cells has several well-established procedures associated with it, though there are many variations on the theme [56, 57]. All modern methods rely on a mild detergent to lyse cells and begin denaturing proteins (especially RNases), and a chaotrope to finish denaturing and solubilizing all macromolecular components (usually guanidinium isothiocyanate or urea). Many also use a second, stronger detergent (such as sodium dodecyl sulfate [SDS]), a chelating agent (usually ethylene diamine tetraacetic acid [EDTA]), and/or a reducing agent (usually 2-mercaptoethanol [BME]) to make sure that the RNA is protected and RNases are inactivated. These procedures are generally sufficient to cause lysis in most human and bacterial cells without adaptation, but are not sufficient to lyse plants, yeast, and the most robust of human (e.g. bone, cartilage) and bacterial cells (e.g. *mycobacterium tuberculosis*) [58, 59]. Such samples are protected by a robust cell wall that often requires enzymatic digestion (e.g. lyticase for yeast) [60] or grinding in liquid nitrogen. The resulting unprotected cells may be lysed by chemical or mechanical means. Human tissue samples from muscles or organs, which usually contain a modicum of extracellular matrix, are often lysed by mechanical stress in a Dounce homogenizer [61] while submerged in an RNA-protecting solution as described above.
The Homogeneity of Cellular Populations: This variable must also be considered, especially in regards to the sample type. Bacterial cells in liquid culture grow logarithmically until a limit of cell density is reached, at which time cell growth attenuates and cells begin to die. At each stage of growth, a variety of different RNAs and proteins are expressed in the cell, creating a population which varies greatly over time [62]. Adherent human cells in culture grow exponentially until they fill the plate on which they are grown, again leading to a population that changes over time. In human tissues there is a constant turnover of cells, but they are not growing logarithmically. Even within a population of cells at a given time, many will be in different phases of growth. For these reasons biological reproducibility poses a challenge to the investigation of nucleic acid building blocks in vivo.

Targeting Individual Compartments: Despite the proliferation and success of bacteria, they are simple organisms consisting of only the barest minimum of compartmentalization, i.e. separation between the intracellular environment and the outside world. Eukaryotes (such as humans), as implied by their name, are more complex, even at the level of single-celled organisms such as saccharomyces cerevisiae (baker's yeast). The hallmark of eukaryotic cells is sub-compartmentalization [63]. The eukaryotic cell is divided into many discrete sections that each handle a set of metabolic processes such as energy production, macromolecular degradation, neutralizing oxygen radicals, etc. [64]. Over the last 60 years, there has been great interest and substantial research into the nature of the compartments of the eukaryotic cell. Much of the early work in fractionation of organelles is attributed to de Duve [65], who used density gradient centrifugation to separate lysosomes, peroxysomes, and mitochondria for individual analysis. Subcellular fractionation is of particular interest in this line of inquiry because there
may be thus far unidentified ribozymes or ribonuclear protein complexes (RNPs) within organelles. Indeed, the differential distribution of proteins and metabolites has already been characterized [65-68], and some work has been done on the localization of micro RNAs as well [69].

Early work with subcellular fractionation was done with sucrose gradients and ultracentrifugation [65]. Sucrose solutions (typically 15-65% by mass) increase in density in a predictable manner with increased sucrose concentration [70]. The refractive index of sucrose solutions is linearly related to the density of the solution [71-73], and is the primary marker used to investigate sucrose gradients. To separate organelles, sucrose solutions are loaded as a gradient into ultracentrifuge tubes. The densest solutions are loaded first, providing a tube that increases in density as it descends. If used immediately, the sucrose solutions will not mix, providing a series of sharp changes in density (a “step gradient”). If left overnight, the solutions will begin to mix, providing a linear gradient.

Ultracentrifugation is a preparative or analytical technique that relies on the application of high ($10^5$-$10^6$) g-forces to a sample by rapid spinning [74]. Particles in solution will sediment according to their size and density [74] until they reach the bottom of the tube or a point where the density of the solution is equal to the density of the particle (the “isopycnic point”). Isopycnic density gradient centrifugation is often used to separate organelles, and is the original technique used by de Duve [65].

Sucrose solutions pose several problems to the separation of organelles. They are viscous and difficult to work with, and are easily contaminated by bacteria. Sucrose solutions are also hyperosmotic as compared to organelles, which causes the organelles to dehydrate and assume an artificially high buoyant density [75].
Ultracentrifugation has been advanced in recent years by the development of new methods for creating density gradients, namely the use of colloidal solutions. One of the most common, Percoll, is a solution of colloidal silica coated in polyvinylpyrrolidone [75]. The colloidal nature of Percoll changes several fundamental facets of the ultracentrifugation procedure [76]. The density curve is sigmoidal instead of linear. Percoll gradients are self-forming and do not require preparation or layering in most cases. Percoll will sediment rapidly in a swinging bucket rotor. To circumvent this limitation, fixed-angle and near-vertical rotors are used instead, which may cause damage to organelles as they are pushed against the wall of the tube while sedimenting. Despite these limitations, Percoll solutions have several advantages over traditional density gradients, such as physiological pH and osmolarity and reduced run times and speeds [76], and were for these reasons investigated in this study.

While there are several other preparative techniques for the separation of organelles, such as immunoprecipitation, capillary electrophoresis, and isoelectric focusing, they were not explored in this investigation.

Classification of RNA: A new interest in RNA editing and modification has begun in recent years with the discovery of new classes of RNA [77] that greatly expand the role of the polymer within a cell [78]. These new RNAs, such as micro RNA (miRNA) and small interfering RNA (siRNA), are often heavily modified after transcription [17]. Enzymes such as Dicer [79-81] and the RNA-induced silencing complex (RISC) [82, 83] produce and utilize RNA oligomers to control gene expression at the post-transcription level. The existence of so many classes of RNA (nearly 10); all with distinct structures, dynamics, and functions within the cell; indicates that RNA modification may play an appreciable and distinct role within each class.
RNA is, overall, a homogenous polymer, despite any modifications present. In each class, the nature of the sugar-phosphate backbone is a constant, providing for little chemical differentiation. Despite these limitations, there are biological aspects to RNA which enable the fractionation of whole-cell RNA extracts by class.

Messenger RNA, despite being a key component of the Central Dogma, comprises only a small fraction of the total RNA in a cell (~4%) [84]. Fortunately, eukaryotic mRNA is heavily modified by the addition of a 3′-poly A tail. This modification allows mRNA to be purified by affinity capture with oligo dT resins [84]. Prokaryotic mRNA is not modified in this way, and has been difficult to isolate until recently [85].

Ribosomal RNA is in many ways the easiest class to isolate, as it is bound up with proteins in the cell. After the mechanical lysis of the cells, all cellular debris larger than ribosomes is sedimented by normal rate zonal centrifugation. Ribosomes are then isolated from smaller subcellular particles by rate zonal ultracentrifugation, as they are in many cases the only RNPs large enough to be sedimented under normal conditions [86].

Transfer RNA is much smaller than the other main classes of RNA, which allows it to be isolated from total RNA by differential precipitation with salt and isopropanol. This method is well-characterized in the literature [87]. Briefly, larger RNA is precipitated under high-salt conditions and the supernatant removed. The supernatant is then treated with isopropanol to precipitate tRNA from smaller RNAs and the supernatant discarded. The resulting RNA is of comparable purity to chromatography on DEA-cellulose.

Other classes of RNA can be isolated as well, but were not explored in this investigation.
**Target Biological Systems:** One of the primary interests of our research program is the development of new strategies for the structure/function relationships in retroviruses. HIV-1 is a deadly and successful pathogen. Indeed, it is one of the most vexing of the modern era. HIV-1 has routinely stymied efforts to develop drugs at combat it, since the virus is capable of rapid mutation to both the genome and resulting proteins [Reviewed in 88]. Some regions of the HIV-1 genome, however, are far less capable of mutation. One of these, the Psi region of the 5′-UTR of the genome, which contains the dimerization initiation site and packaging signal [89-95], is of particular interest due to its inability to mutate rapidly [96]. Like many RNA targets, however, the 5′-UTR presents many problems for analysis.

Typical high-resolution structural techniques have been turned on the 5′-UTR of HIV-1, but to little avail. The 5′-UTR is too large for nuclear magnetic resonance (NMR) analysis of more than a small portion such as a single stem loop, and the structure is too malleable to crystalize properly. Structural probing represents a valid, although low resolution, approach for pursuing the structural elucidation of systems that are not directly amenable to the typical high resolution techniques. Mass spectrometry represents a very powerful platform for the characterization of RNA modifications introduced by structural probes. The combination of probing and MS-detection has been dubbed MS3D [97-100]. It uses primarily bifunctional cross-linkers of a known size to form covalent linkages between nearby parts of a folded macromolecule. These linkages cause information of the 3D structure to be retained after digestion of the target molecule when the resulting fragments are examined by MS. Information gleaned by cross-linking studies can be used as a set of modeling constraints for *in silico* modeling with results similar to those obtained by traditional techniques [99]. The resulting models can be used for drug design.
In this investigation, RNA-RNA and RNA-protein cross-linking with the subsequent purification of the product from heterologous RNA were again examined with an eye towards mapping the structure (or structures) of the HIV-1 5′-UTR in vivo. Information gained from the investigation of nucleoside modifications will complement cross-linking data for in silico modeling.

Despite the sensitivity of mass spectrometry as an analytical technique [40], it cannot compete with replicative techniques such as RT-PCR for sample requirements, necessitating a (relatively) large amount of virus to be collected. The standard work-horse of molecular biology, E. coli, can be used to express recombinant HIV 5′-UTR, but without the benefit of native HIV proteins and modifications. Eventually, the virus had to be purified from human cells.

The most efficient way to express HIV-1 ex vivo is by transient transfection [101] of immortal eukaryotic cells. A common cell line used for this procedure is the 293T line of human embryonic kidney cells [102], due to their propensity for successful transfection and immortality. Viral yields can be increased further through a process known as pseudotyping [101].

Pseudotyping involves the simultaneous expression of the proteins from two of more viruses in a single cell. Virions are packaged using any available compatible proteins, possibly including those from another virus. In the case of HIV-1, using a coat protein from the vesicular stomatitis virus (VSV-G) removes the dependency of HIV-1 on the CD4 receptor as an infection pathway (as it can now infect by direct fusion with the plasma membrane), and removes the dependence of infection success on pH and the Nef protein (indicating that virions are no longer brought into the cell by endocytosis). A side effect of pseudotyping, besides a tenfold increase in infectivity, is a significant increase in the production of viral particles, by up to a factor of 10 [101].
HIV-1 is known to package several RNAs within the virion when it exits a cell [103]. A method is necessary to purify viral RNA from heterologous RNA. After reviewing the efficacy of HPLC, capillary gel electrophoresis (CGE), and immunoprecipitation (IP) for this task, affinity capture was employed as described below due to its increased size limitations (vs. HPLC and CGE) and high specificity (vs. IP).

Analysis of Data from Biological Samples: Investigation of the nucleic acid building blocks of biological samples generates large amounts of data. Various software tools and statistical methods have been employed during our analysis, specifically for the comparison of technical and biological replicates of bacterial samples and the analysis of tandem MS data.

In proteomics, database searching algorithms such as MASCOT [104, 105] and X!Tandem [106] can be used to readily employed to search HDMS data against different databases to identify proteins from peptide digestion products in a complex mixture. Unfortunately, these tools have not been adapted to enable the positive identification of RNA components. Recently, new software called ProVerB [107] has been published, which uses a novel probability algorithm to assess the strength not only of the exact mass match, but also of the tandem MS data. Unambiguous identification of nucleotide modifications by gas-phase techniques requires the use of tandem MS data, and conclusions based on these data are strengthened by a quantitative method for analyzing the resultant spectra. The ProVerB algorithm can be adapted to this purpose, and it provides a measure of the congruence of the experimental spectrum to a reference file based on the presence of peaks, the sequence of peaks, and the height of each peak relative to a theoretical value. As such, operator skill is removed from the equation in regards to successful analysis of CRM data.
Unlike traditional mass spectrometry, where the signal is returned as a sine wave, high-definition mass spectrometry (HDMS) data is often reported without any background, so a different method is required for the calculation of signal-to-noise ratio (SNR). As well, a method is needed to compare quantitative data across multiple replicate data sets. In high-definition optics the background of a picture is taken to be zero, and all measurements are based off of that assumption \[108\], similar to HDMS. Or course, this makes the signal-to-noise ratio effectively infinite when assessed through conventional means. The SNR is then calculated based on the variation of the signal, as such:

“SNR = average/STDev”

The relative intensity of each species detected is averaged across five technical replicates, and the SNR is calculated. A cutoff of SNR=2 was used for qualitative data. This means that some modifications not present in all samples are reported as extant, as their lack in one or more samples is considered statistically insignificant.

Biological replicates thus far have been compared largely on a qualitative basis, i.e. whether or not an RNA modification is present or absent in the sample. There are many statistical models for comparing disparate data sets, but one of the simplest and most robust is the Gower distance \[109\]. The Gower distance is essentially a Boolean comparison of two data

\[
S_{ij} = \frac{\sum_{k=1}^{r} s_{ijk}}{\sum_{k=1}^{r} t_{ijk}}.
\]

**Scores and validity of dichotomous character comparisons:**

<table>
<thead>
<tr>
<th>Individual $i$</th>
<th>Values of character $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j$</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

$g_{ijk}$

$g_{ijk}$

1 0 0 0
1 1 1 0

**Figure 1:** a representation of the mathematics and logic behind the Gower Distance. Reprinted from reference \[109\]
sets, where a data point present in both sets indicates similarity, a difference between the sets indicates distance, and a data point present in neither set does not contribute to the similarity or difference of the sets [Figure 1]. The Gower distance was chosen for this analysis due to strong and current support in the literature [110, 111] and malleability: the Gower distance is a purely mathematical construct and as such is only as constrained and valid as the method in which it is applied to the data.

Finally, it is important to note that the rather disparate biological systems employed in the study were specifically selected to pose the challenges necessary to drive the development of our approach, test its robustness according to different application demands, and demonstrate its broad applicability to a variety of themes. From the analytical chemistry point of view, there is no discrepancy between the analysis of natural modifications for pseudotyping purposes and that of man-made crosslinks meant to reveal RNA structure in vivo. For this reason, the next sections describe the analysis of nucleoside modifications from a series of biological samples, as well as the development of methods for crosslinking, isolation, and purification of HIV-1 RNA. The methods explored here represent the different steps of a general workflow for the identification and characterization of any type of RNA modification in vivo.
Results and Discussion

Analysis of nucleic acid building blocks in the gas phase is hindered by the closely related masses and structures of the different ion species; many have similar structures or identical masses. The ability of orthogonal gas-phase analytical techniques to differentiate isobaric molecular species was examined to secure the tools needed to assess nucleotide modification levels in whole cell RNA extracts. A mixture of the four canonical bases from DNA and RNA, as well as pseudouridine, the “fifth ribonucleotide [112],” was examined on both a high-resolution and an ion mobility mass spectrometer to determine their respective utilities in this matter. Exact masses and elemental compositions are listed in Table 1. As can be seen from the chart, for several species the mass differs by only a single Dalton (which leads to overlapping isotopic distributions), and even within these nine species there are two sets of isobars: UMP/ΨMP and AMP/dGMP. Isobaric species have historically presented a challenge to mass spectrometric analysis, but many of these limitations may be overcome with new techniques. A representative negative-ion spectrum from the LTQ is shown below in Figure 2a and all seven expected peaks are seen well above the background. All data below is for negative-ion mode, unless otherwise specified.

<table>
<thead>
<tr>
<th>Name</th>
<th>Elemental composition</th>
<th>Monoisotopic mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guanosine monophosphate (GMP)</td>
<td>C_{10}H_{14}N_{5}O_{8}P</td>
<td>363.05800</td>
</tr>
<tr>
<td>Adenosine monophosphate (AMP)</td>
<td>C_{10}H_{14}N_{5}O_{7}P</td>
<td>347.06308</td>
</tr>
<tr>
<td>Cytidine monophosphate (CMP)</td>
<td>C_{9}H_{14}N_{5}O_{8}P</td>
<td>323.05185</td>
</tr>
<tr>
<td>Uridine monophosphate (UMP)</td>
<td>C_{9}H_{13}N_{2}O_{9}P</td>
<td>324.03587</td>
</tr>
<tr>
<td>Pseudouridine monophosphate (ΨMP)</td>
<td>C_{9}H_{13}N_{2}O_{9}P</td>
<td>324.03587</td>
</tr>
<tr>
<td>Deoxyguanosine monophosphate (dGMP)</td>
<td>C_{10}H_{14}N_{5}O_{7}P</td>
<td>347.06308</td>
</tr>
<tr>
<td>Deoxycytidine monophosphate (dCMP)</td>
<td>C_{9}H_{14}N_{3}O_{7}P</td>
<td>307.05693</td>
</tr>
<tr>
<td>Deoxyadenosine monophosphate (dAMP)</td>
<td>C_{10}H_{14}N_{5}O_{7}P</td>
<td>331.06817</td>
</tr>
<tr>
<td>Deoxythymidine monophosphate (dTMP)</td>
<td>C_{10}H_{15}N_{2}O_{8}P</td>
<td>322.05660</td>
</tr>
</tbody>
</table>

Table 1: Monoisotopic neutral masses of mononucleotides used in this study, calculated from the elemental composition by the PNNL molecular mass calculator [113].
Use of HDMS and Tandem MS for dNMP/NMP Analysis: The Thermo Finnegan LTQ Orbitrap Velos is a hybrid mass spectrometer consisting of a versatile linear trap (LTQ) where ions are isolated and may be detected or subjected to gas phase chemistry reactions, and an electrostatic trap (Orbitrap) which is capable of high-resolution and high-accuracy mass measurements. With small molecules such as mononucleotides, the nature of the LTQ Orbitrap provides many unique analytical advantages.

Mass resolution refers to the ability of an instrument to separate ions of similar m/z values. A low resolution instrument, such as a linear trap, has sufficient resolution to differentiate nominally different masses, such as [CMP-H]$^-$ (322 m/z) and [UMP-H]$^-$ (323 m/z) [Figure 2b]. Since these molecules differ in mass by only a single Dalton, however, the LTQ cannot deconvolute the overlapping $^{13}$C isotope of [CMP-H]$^-$ (323.05191 m/z) from the monoisotopic peak of [UMP-H]$^-$ (323.02804 m/z) at the best resolution available (~2300 FWHH). The ability to resolve nominal isobars on the mass scale is important for whole-cell extract
analysis, as some modified nucleotides differ in mass by less than 0.1 Da. See Table 2 above for examples.

An electrostatic trap, such as an Orbitrap, is a high-resolution Fourier transform instrument, designed to compete with the ICR in terms of resolution and accuracy. Compared to the LTQ, Orbitrap spectra [Figure 2c] display very narrow peaks (resolution ~160000 FWHH), allowing nominal isobars to be distinguished. In the inset of Figure 2c, the ability of the Orbitrap to distinguish the $^{13}$C isotope of [CMP-H]$^-$ from the monoisotopic peak of [UMP-H]$^-$ can clearly be seen.

Another advantage of the LTQ Orbitrap is the high level of mass accuracy. An accuracy of up to 200 ppb can be obtained using an external calibration. For molecules of this size (<400 Da), this level of accuracy is sufficient to unambiguously determine the chemical formula of the analyte from the mass, using an elemental composition calculator such as that available from PNNL [113]. For example, the experimental m/z of UMP, 322.04406, was input and two formulae returned, $C_8H_7N_1O_3P$ and $C_9H_{13}N_3O_6P$, the latter of which is the correct formula for UMP and the former of which does not represent a viable molecule.

True isobars, such as UMP/$\Psi$MP and AMP/dGMP cannot be differentiated by MS, regardless of the resolving power. Instead, they must be examined by tandem MS or IM-MS.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Chemical Formula</th>
<th>Exact Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-thiouridine</td>
<td>$s^4U$</td>
<td>$C_9H_9N_2O_5S$</td>
<td>339.005198</td>
</tr>
<tr>
<td>5-hydroxyuridine</td>
<td>$ho^5U$</td>
<td>$C_9H_9N_2O_5$</td>
<td>339.022956</td>
</tr>
<tr>
<td>5-methylidihydrouridine</td>
<td>$m^5D$</td>
<td>$C_{10}H_{15}N_2O_5$</td>
<td>350.059342</td>
</tr>
<tr>
<td>5-formylycytidine</td>
<td>$f^5C$</td>
<td>$C_9H_{13}N_3O_6$</td>
<td>350.08941</td>
</tr>
<tr>
<td>5,2'-O-dimethylyctidine</td>
<td>$m^5Cm$</td>
<td>$C_{10}H_{17}N_3O_5$</td>
<td>350.075326</td>
</tr>
</tbody>
</table>

Table 2: Examples of known nucleotide modifications which are nominal isobars. High mass-resolution will be necessary to differentiate their masses. Masses were calculated with the molecular mass calculator available from PNNL [113].
For nucleic acid building blocks, gas phase fragmentation provides sufficient information for positive identification of each species [26-29]. The LTQ region of the LTQ-Orbitrap is capable of up to 10 rounds of fragmentation by a variety of techniques. For this investigation, collisionally-induced dissociation (CID) was used (see below). CID occurs under low energy, high pressure, and long time-scales. Ions activated by CID distribute the energy gained throughout the molecule, causing the weakest bond to break first. In all nucleotides, save ΨMP (see below), the N-glycosidic bond is the most labile [30, 114]; thus in negative-ion mode the most abundant first generation fragment ion is the ribose-5′-phosphate (211 m/z) or deoxyribose-5′-phosphate (195 m/z) [Figure 3a]. Other fragments are visible as well, corresponding to loss of water an outright cleavage of the ribose ring structure. MS³ data [Figure 3b] of the predominant first-generation ion product produces a series of diagnostic fragments which are traceable back to the parent ion.

For an isobaric pair such as AMP and dGMP, MS³ can provide structural information to enable their differentiation. As mentioned above, both ribonucleotides and their deoxy versions will fragment in the same location. The resulting fragments
will differ in mass by that of the 2’-oxygen. It is therefore possible to differentiate
AMP and dGMP by their first-generation fragments [Figure 3c].

Interestingly, no difference was observed between the NMPs and dNMPs in
fragmentation of the N-glycosidic bond in the LTQ, which is in direct conflict with
previously reported results [30, 114]. This same trend was not observed on the ion
mobility instrument, however (see below).

Positive-ion spectra of the NMP/dNMP mixture showed strong signals for all
species present. Tandem MS under similar energy and activation time schemes as in
negative ion mode readily produced fragmentation of the N-glycosidic bond for all
species save ΨMP. For all but dTMP and UMP, the primary fragment observed
among the first generation ion products was the protonated nucleobase, [B+2H]^+. These fragments allowed strong corroboration of the data collected in negative-ion
mode, and the differentiation of AMP and dGMP once again by the difference in mass
of the nucleobases [Figure 3d]. dTMP and UMP do not show the [B+2H]^+ fragment,
a phenomenon which may be ascribed to their relatively low proton affinity versus
the other nucleobases (i.e. G > C > A >> T ≈ U) [115-117].

Most ribonucleotide modifications that produce isobaric species do not afford
the same breadth of structural differences that AMP and dGMP have. In this vein,
the ability of tandem MS to differentiate UMP and ΨMP, which differ only by the
orientation of the nucleobase, was investigated.

In negative-ion mode, UMP produces similar first-generation ion products as
AMP [Figure 4a], save for the addition of the loss of CONH from the nucleobase as
well. Because ΨMP and its derivatives are the only NMPs, modified or otherwise, to
replace the N-glycosidic bond with a carbon-carbon bond, the base is not the most
labile moiety. Instead, ΨMP first loses the phosphate moiety, leaving a nucleoside
Figure 4: Negative ion mode MS/MS spectra obtained from a) a pure UMP sample; b) pure ΨMP; c) a UMP/ΨMP mixture.

For this reason it is also possible to differentiate UMP and ΨMP by their first-generation fragments [Figure 4c].

Evaluation of Fragmentation Techniques: CID is one of the most common fragmentation techniques, but it is not without limitations. Notably, ion traps cannot retain fragments with an m/z of less than ~1/3 of the parent ion. For most NMPs and dNMPs, this limit precludes the detection of two key negative-ion diagnostic fragments, phosphite (79 m/z) and phosphate (97 m/z). For this reason both high-energy CID (HCD) and pulsed-Q dissociation (PQD) were investigated as methods to circumvent this problem. HCD is similar in principle to CID, but with a different energy scheme [118] and slightly different fragmentation patterns expected [119]. It also requires the use of the Orbitrap analyzer. While HCD allows for the low-mass diagnostic fragments to be detected (data not shown), the high electrical noise in the Orbitrap versus the LTQ makes interpretation of some spectra difficult. PQD is a Thermo proprietary fragmentation technique relying on pulsed electrical fields [120]. It was specifically designed to counteract the dynamic range problems of CID, but is not well supported in the literature and produces far more noisy and random fragmentation patterns (data not shown). For these reasons
both techniques were abandoned in favor of MS$^3$ CID, which produces clearer spectra despite requiring an additional fragmentation step.

**Investigation of Limit of Detection:** For biological samples, the ability to quantitate any present modifications and compare the relative intensity (versus the canonical nucleotides) to data expected from biological experiments. Since many modifications are present only in trace amounts relative to the canonical bases, it will also become necessary to detect ions at very low concentrations. As such, the limit of detection (LOD) of the Orbitrap was assessed for dTMP, which displayed the weakest signal in an equimolar mixture. Dilutions from the original concentration of 1 µM down to 200 pM were analyzed for 16 orbitrap scans over ~30 seconds. A 200 pM solution displayed a 5:1 S/N ratio for dTMP, and the signal was absent at lower concentrations. Since nano-ESI flow rates average ~20 nL/min [1121], only 2$^{-18}$ (attomole) moles of sample were consumed in the run. LOD was tested again with selected reaction monitoring of the fragmentation of dTMP to ribose-5'-phosphate, which allowed a concentration of only 5 pM to produce a signal with an S/N ratio of at least 2:1. LOD data for the Orbitrap collected here rivals that reported previously for the instrument. In positive-ion mode, an abundance drop of ~30% was seen for all species.

Even at equimolar concentrations, NMPs display markedly different abundances. It is expected that in negative-ion mode the ion abundance will follow the order of gas proton affinities (i.e., AMP > GMP > CMP > UMP > TMP [1117, 122]), and the reverse order of their solution pK$_a$s (i.e., TMP > UMP > CMP > AMP > GMP [123]). Following this line of reasoning, the reverse order should be expected for spectra acquired in positive-ion mode. This was not the case, however, as the ions showed the same abundance pattern in both modes. Ion abundance for dNTP samples has previously been shown to be a function of the hydrophobicity of the
Figure 5: Ion mobility profiles obtained from the precursor ions at m/z 346.17 a) and 324.04 c). Panel b) and d) are the respective curve-fitted plots generated by PeakFit 4.1 (see Experimental). Analyte identity was assigned by comparing the apparent drift time ($t_D$) of the deconvoluted signals with those obtained in separate experiments from isolated nucleotides.

Use of IM-MS for dNMP and NMP Analysis: Ion mobility is a gas phase technique for separating ions based on collisional cross-section. Ions are moved through a high-pressure nitrogen-filled cell by a pulsed electric field, called a “traveling wave” on the Synapt G2. Each wave moves the ions based on molecular mass and cross-section as it passes. The ions are delivered to the TOF analyzer on a millisecond timescale. dNMPs have been modeled previously [125], with the data confirmed by drift-time IMS-MS. It was shown that the addition of bulky groups, such as the 5-methyl group on dTMP, increase the drift time over otherwise similar molecules (such as dCMP), as expected. However, it was demonstrated that even small molecules such as dNMPs can travel in different conformations in the gas phase, and dGMP travels in the syn conformation instead of the anti, causing it to travel more quickly than dAMP. It is expected, then, that no such trend was observed for any NTP sample, even under identical conditions. Clearly further investigation is required into these trends.
dNMPs will migrate more quickly than the corresponding NMP, UMP will migrate faster than CMP, and AMP will travel more slowly than GMP.

IMS provides complementary data to tandem MS in that it investigates the placement and structure of functional groups and units on a molecule rather than the identity of said groups. Data collected by the IMS (in the form of drift time) provides information of the cross-sectional area of a molecule, related to its structure and conformation rather than the identity of functional groups.

For example, the isobaric pair of AMP/dGMP at 346.17 m/z was isolated in the mass-selection quadrupole and injected into the mobility cell of the instrument. The corresponding mobility chromatogram of total ion intensity versus time shows two distinct populations with different interactions with the background gas (i.e. different conformations) [Figure 5a]. When a Gaussian curve-fitting algorithm was applied (see experimental section for details), the presence of two normally-distributed populations is even more apparent [Figure 5b]. The apex of each population correspond to the drift times of AMP and dGMP (6.08 and 5.74 ms, respectively), as determined from pure samples.

Mobility profiles for the more closely related UMP/ΨMP pair also display a separation, albeit not as well resolved [Figure 5c]. In this case curve-fitting software plays an even larger part in differentiating the two populations [Figure 5d]. Drift times for each peak correspond to those attained for each species in a pure sample (5.24 and 5.04 ms for deprotonated UMP and ΨMP, respectively). Given the ability of IMS to distinguish UMP from ΨMP, it is likely that it will possess sufficient resolution to differentiate modified nucleosides in a cellular extract.
Mobility data can be obtained individually for each species of interest, or in an automated sequence within the MassLynx software. Alternatively, all mobility data collected in an experiment can be assessed in parallel. Each “pulse” of the Tri-WAVE element delivers a new population of ions to the TOF detector, which analyzes the mass and intensity of all ions in the sample as normal. Each pulse (“bin”) is analyzed in series, providing a 3-D dataset for each sample. The DriftScope software (see Experimental) can display this information as a 2-D graph of ion m/z versus drift time, with the color of each point representing the intensity of the peak. As can be seen in Figure 6, the resulting “heat map” displays not only the ions of interest (i.e. the NMP/dNMP mixture), but also all salt adducts and unidentified contaminants. Enlarged in Figure 6 is the region containing the NMP/dNMP mixture, in which all analytes can be seen at a distinct mass or mobility peak. AMP and dGMP, while

![Figure 6: Comprehensive representation of IMS-MS data obtained from the entire nucleotide mixture. In the heat-map, the m/z of each ion is plotted against its apparent tD, whereas its relative intensity is expressed by progressively lighter colors for the more abundant species. Red dots mark the maxima corresponding to the various nucleotides. The enlarged inset highlights the separation between mixture components, which was achieved in the tD and m/z dimension of the plot. A variety of salt adducts and unidentified background contaminants were also detected. No signal corresponding to typical nucleotide fragments was recognized in experiments completed without ion activation.](image-url)
sharing the same horizontal parallel (m/z value), are separated in the time dimension owing to their differing conformations. UMP and ΨMP, due to their greater similarity, are not fully resolved. In such a case it is possible to extract a mobility profile similar to Figure 5c directly from the heat map, and apply curve-fitting algorithms as in Figure 5d.

The Synapt G2 has the ability to perform tandem MS in multiple locations, including after the IMS cell. In this setup, ions traveling through the IMS cell are separated by mobility before being fragmented, and thus each bin of ions delivered to the TOF should contain the first generation ions products of each bin, but with an identical drift time to the precursor ion. This is referred to as “time alignment parallel fragmentation (TAP)” [126]. Due to the broad distribution of mobility values

![Figure 7: Heat-map representation of time aligned parallel (TAP) dissociation products provided by the entire nucleotide mixture. The data were collected by using a transfer voltage of 25 V (see Experimental). The inset represents the mass spectrum extracted at 3.6 ms of the plot (white solid line). The spectrum contains pseudo-molecular ions that shared the same $t_D$ across the Tri-WAVE, as well as their fragment ions produced in the transfer region before the TOF analyzer. Such fragments were not observed in analogous experiments performed. Asterisks mark unidentified background contaminants. without activation (Fig. 5).](image)
for each ion population [Figures 5a and 5c], TAP spectra are likely to contain overlapping fragments. For example, a drift time transect at 3.6 ms (the vertical line in Figure 7) shows not only the fragments for AMP, but also fragments from the tail-end of the dAMP and dCMP peaks. This can be seen in the spectrum the inset of Figure 7, which shows both the ribose-5’-phosphate and the deoxy ribose-5’-phosphate peak. TAP data display fragmentation patterns similar to previous spectra of mixed precursors [Figures 3c, 3d and 4c]. As such the interpretation of TAP data will benefit from the same knowledge and procedures as the previous tandem MS experiments.

Contrary to the pattern seen in the LTQ-Orbitrap for gas-phase fragmentation, an investigation of the effect of activation energy on the presence of fragmentation products showed a difference between the stability of the N-glycosidic bond in NMPs versus dNMPs (data not shown). This may indicate that fragmentation schemes in the Synapt versus the LTQ are sufficiently different to provide different information.

**Development of Procedures for Cellular Work:** The eventual goal of this line of investigation is to apply the methods developed above to biological samples. Differences in nucleotide modification levels may be present between different species of microorganisms, between different human tissue types, and between healthy and cancerous human cells [127, 128]. In order to examine any of these samples, many challenges need to be addressed. Namely, the techniques outlined thus far must be adapted to far more complicated samples, and they must be assessed for reproducibility between technical and biological replicates of similar samples to establish what constitutes a “significant difference” between two cell types.
Application of techniques to biological samples: CID provides a method for investigating the presence of isobars in an unknown mixture of NMPs from total RNA extracts. One of the primary challenges to this method is the possible presence of multiple collections of isobaric species, each with 2-5 possibilities. As an example, for a species such as methyl GMP, there may be up to four different isomeric species present in the sample: $\text{m}_1^1\text{GMP}$, $\text{m}_2^2\text{GMP}$, $\text{m}_7^7\text{GMP}$, and GmMP. As noted above, any 2′-O-methylated nucleotide can be distinguished from the other species by its first generation ion products in negative-ion mode. Instead of producing the 211 m/z ribose-5′-phosphate ion, it will instead produce the 225 m/z 2′-O-ribose-5′-phosphate [Figure 8a]. The other species will produce the same ion products in both negative- and positive-ion mode, and must be subjected to additional rounds of fragmentation for positive identification. Since all information as
to the location of the modification is on the nucleobase, MS^3 was performed in positive ion mode. The characteristic MS^3 pattern of GMP is shown in Figure 8b. Similar to Figure 3d, MS^2 in positive mode shows two prominent ion products, methylated guanine, and unmethylated guanine [Figure 9a]. From here it is further possible to differentiate m^1GMP and m^2GMP from m^7GMP based on the fragmentation pattern, as seen in Figure 9b.

One second-generation ion product at 109 m/z corresponds to m^7GMP, as the modification is lost in the fragmentation. Another primary ion product is present at 123 m/z, which has the same loss as unmethylated guanine. There is, however, no way to differentiate m^1GMP and m^2GMP. The third generation ion product was not abundant enough for another round of fragmentation. This problem is also observed for mAMP and mCMP. mUMP brings the additional problem of not ionizing well in
positive mode, precluding MS\textsuperscript{3} at the concentrations available in most biological samples. Even though some isobaric combinations, such as m\textsuperscript{2,2}GMP and preQ1 could be fully differentiated by MS3 [Figure 9c] (despite problems with signal intensity and quality), these difficulties lead to the use of IM-MS to attempt to circumvent these limitations.

Use of IM-MS to Examine Isobars in a Complex Mixture: The ability of the Synapt to differentiate isobaric modified nucleotides was investigated using a standard mix of methylcytidine monophosphate species: 2'-O-methyl CMP, 3-methyl CMP, and N\textsuperscript{4}-methyl CMP. As can be seen above [Figure 10a], IMS cannot separate methylated CMP derivatives in negative-ion mode (a). Certain heavily modified GMP derivatives can be separated by IMS, such as m\textsuperscript{2,2}GMP and PreQ\textsubscript{1}MP (b). It is not known which mobility peak corresponds to which species. Fragmentation before the IMS cell allows unmodified adenine bases to be separated from methylated bases (c).

mAMP were similarly impossible to separate (data not shown). Preliminary theoretical calculations of the collisional cross-sections of these modified nucleotides showed very little difference, and increases IMS resolution is necessary to separate

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure10.png}
\caption{IMS cannot differentiate methylated CMP derivatives in negative-ion mode (a). Certain heavily modified GMP derivatives can be separated by IMS, such as m\textsuperscript{2,2}GMP and PreQ\textsubscript{1}MP (b). It is not known which mobility peak corresponds to which species. Fragmentation before the IMS cell allows unmodified adenine bases to be separated from methylated bases (c).}
\end{figure}
then in the gas phase. Certain isobaric mixtures, such as m²GMP and preQ1 were readily differentiable by IMS [Figure 10b], but this did not provide an advantage over tandem MS.

To circumvent these limitations, a variety of fragmentation schemes were implemented. TAP fragmentation after the mobility cell did not assist in identification, as it lowers the mobility resolution of the IMS. Fragmentation before the IMS cell allows the modified nucleobases to travel through the cell without the attached sugar, which was expected to increase their separation based on preliminary modeling with Jmol. As can be seen above [Figure 10c], this allows for the separation of modifications on the base from those on the sugar, but not any improvement in function over the Orbitrap. By increasing the source pressure and energy, it is possible to cause fragmentation in the source of the Synapt, before the mass selection quadrupole. This allows for the isolation of only the modified nucleobase, which can then be fragmented before the IMS cell, and the MS³ products separated by mobility. This approach has not been successful (data not shown) due to low ion abundance and a lack of IMS resolution.

The failure of ion mobility to differentiate many isobaric nucleotides may be a function of the highly homogenous nature of many NMPs, i.e. that 2/3 of the molecule consists of the sugar phosphate moiety, as Jmol predictions showed similar cross-sections for methyl AMP derivatives. It may also be attributable planar nature of the molecules once the ribose-5'-phosphate is removed. Though preliminary modeling indicates that there is a difference in the collisional cross-section of the methylated nucleobases, they may travel such that the “edge” of the base leads through the mobility cell instead of the “face,” which would reduce the perceived collisional cross-section.
Use of HPLC to Examine Isobars in a Complex Mixture: In a final effort to differentiate isobaric species, a mixture of methylated cytidine derivatives (m\(^3\)C, m\(^4\)C, and Cm) were tested on a reversed-phase C18 column with a methanol gradient. Cytidine derivatives were delivered as triphosphates and digested with either CIP or BAP until all phosphates were removed. They were loaded on the column without further purification. Partial separation was achieved in less than 30 minutes [Figure 11]. For preliminary identification of modified nucleosides, this method was abandoned as the resolution and sensitivity of the Orbitrap were deemed sufficient for identification without front-end separation. For diagnostic purposes where identification of difficult isobars is crucial, HPLC will be revisited.

Reproducibility Studies 1 - Technical Reproducibility: In order to assess the differences between disparate samples, the reproducibility of the method for technical and biological replicates was examined. Technical reproducibility can be defined using different criteria. One of them considers the reproducibility of the mass assignments for repeated determinations of the same sample. However, the accuracy afforded by orbitrap MS and, if necessary, the possibility of employing internal calibrants, limit the usefulness of such criterion. A more meaningful measure of reproducibility considers the signal intensity of selected species, which can be correlated to their respective abundance in the sample. Given that analyte

![Figure 11: Methyl cytidine derivatives can be separated by RP-HPLC. The non-retained solute (mostly undigested di- and tri-phosphates) elutes between 7.5-10 minutes. Methylated cytidines are well separated, eluting at 20 minutes, 27 minutes, and 30 minutes. A 5-65% methanol gradient with 100 mM ammonium acetate was used to elute the samples from a C-18 column. Identity of experimental peaks was not confirmed by mass spectrometry.](image)
abundance can be affected by losses at any possible step of the analysis workflow, a true assessment of technical reproducibility can be obtained only by repeating the entire sample preparation procedure that precedes MS analysis. In contrast, biological reproducibility can be assessed by growing in parallel multiple cultures of a model microorganism, under exactly the same conditions. This type of determination can sense putative variations between individual cultures that may be more difficult to control, but have greater biological significance. Such variations may be very subtle and, thus, a good handle on technical reproducibility is a must to correctly attribute the source of observed variability.

The reproducibility study was carried out by using *E. coli* as a model, due to the fact that it is the most well-studied and characterized bacterium to date [Reviewed in 129]. It is also non-pathogenic and grows rapidly, traits all of which make it an ideal starting point for method standardization. Five technical replicates each of five biological replicates of *E. coli* were examined to standardize the method.

Technical replicates provided several hurdles for analysis, namely that several low-abundance modifications were absent from the final analysis of some samples that were present in others. Since there were three data filtration steps (charge, mass accuracy, and signal-to-noise ratio), all of which are related to the ability of the instrument to accurately and reproducibly (on a scan-to-scan basis) detect ions, the nature of the sensitivity and accuracy of the Orbitrap instrument was assessed.

Since each quartz needle for nano-ESI is different, the parameters of the instrument were adjusted to compensate. The LTQ Orbitrap Velos is equipped with a predictive ACG control (pAGC), which modulates ion accumulation time depending on the rate of ion accumulation from the previous scan. When the number of ions in the trap reaches the pre-set limit, or a manually determined time limit has passed, the ion trap will close and the ions will be exported to the Orbitrap for analysis. For
scans performed in the LTQ, ion accumulation is the rate-limiting step in scan time. In contrast, scan times in the FT-Orbitrap are determined primarily by the time taken to acquire the FID transient. This allows the maximum inject time can be raised so that the ion trap always fills to capacity, as determined by the AGC, without significantly reducing the number of scans per unit time. It has previously been reported [130] that the accuracy of the Orbitrap mass measurements increases (from \(~5\) ppm to \(<1\) ppm) with number of ions scanned, up the preset limit of the instrument (\(~10^6\)). Further investigation with our instrument has shown similar results [Figure 12a]. Sensitivity did not show a correlation with number of ions scanned [Figure 12b].

Allowing the Orbitrap to always reach the AGC limit, increasing the number of microscans per analytical scan, and increasing the number of analytical scans performed enhanced reproducibility between technical MS replicates. Increasing
sample concentration did not increase the reproducibility of the samples (data not shown), as still there were still species present in some but not all technical MS replicates. In fact, several modifications observed at the lower concentration are not observed at the higher concentration, likely due to the ion trap shutting earlier when the AGC limit is reached.

Final HDMS spectra collected on the LTQ-Orbitrap [Figure 13] look similar to those gathered from NMP standards [Figure 2a]. There are a significant number of peaks present in the experimental sample not present in the standard, mostly of low abundance. The XCalibur software is capable of exporting full lists of these peaks with many other attached data points. As noted in the Methods section, peaks were accepted as possible modifications if they were reported with a charge, and were within the error limit of the canonical bases.

Final confidence in the identity of each modification was established by CRM in negative ion mode. As noted in the introduction, nucleotides produce a distinctive MS$^3$ pattern that can be compared to a reference file. The ProVerB [107] algorithm, a freely available statistical package for quantitatively assessing tandem MS data in
proteomics, provides a quantitative method for this comparison. Any spectra with a ProVerB score above the established cutoff, defined as the 1% FDR limit, will accepted as true, and all others discarded. At this point, however, analysis of tandem MS spectra is performed manually.

Any isobaric species in the resulting list of accepted extant modifications were examined by MS$^3$ in positive ion mode or IM-MS as applicable.

To quantitatively assess the homogeneity of technical replicates, a second method of signal-to-noise ratio was used, as noted in the Methods section. Despite some modifications existing in only certain technical replicates, the statistical methods employed provided a homogenous data set for each biological sample.

Due to time limitations, this study considered only the MS technical reproducibility. Given that the overall variance ($s^2$) of an analytical workflow is the sum of the variances observed at each individual step, we expect that cell lysis, nucleic acid extraction, RNA digestion, and any other sample handling operation performed before MS analysis will contribute in different ways to the overall technical variance.

Reproducibility Studies 2 - Biological Reproducibility: *E. coli*, similar to most cells, display a changing variety of gene expression throughout their life cycle [62]. Biological samples were originally collected at saturation, but the diversity of cell populations (dividing, nascent, aging, dead) provided for poor reproducibility. Samples collected at mid-log phase proved similarly diverse, as dividing cells display a similarly wide array of metabolic states depending on where in the division process they fall [62]. Based on growth-curve experiments for reproducibility (data not shown), the *E. coli* was collected at OD$_{600}$ of 1.2-1.5, which corresponds to late logarithmic growth. In this phase cells display a more homogenous population; this has led to better reproducibility of biological samples. Biological samples collected at
times of diverse populations show a great difference in the number and identity of nucleoside modifications detected, with no discernible pattern to the changes. Collecting samples at late logarithmic phase has raised reproducibility to a measureable level, as shown in the following section. Biological reproducibility studies have not yet been completed.

Reproducibility Studies 3- Assessing Reproducibility: The preliminary lists of modifications were compared based on the presence or absence of a modification (i.e. qualitatively), by Gower’s coefficient of similarity [109] as described in the Methods section. Briefly, an agreement between all technical replicates was considered a point of similarity, and a disagreement between sets as to the existence of a modification was considered a point of dissimilarity. The number of similar points was divided by the total number of modifications identified in the study. Based on this metric, technical replicates have a similarity coefficient of 0.9 or greater, while biological replicates show 85% congruity. These distances establish a baseline for future comparisons, such as between different bacterial species, different human tissue types, and healthy versus cancerous human cells. Later technical replicate results were analyzed by averaging signal (as a percentage of the combined response for the four canonical nucleotides) over variation (see Methods for details). This procedure provided a quantitative means for averaging the technical replicates, raising the similarity coefficient to unity. The increase in stringency reduced the number of modifications positively identified across the biological replicates to only seven.
To ensure further the robustness of the method, the effects of metabolic environment on RNA modification in E. coli were examined. Three different growth conditions, including stringent (glucose minimal media), complete (lysogeny broth) and rich (terrific broth) were examined. In cells under stringent control, there is a smaller pool of stable RNA \([131, 132]\), and synthesis of more is down-regulated compared to a healthy cell, where up to 60% of transcription involves stable RNA. It was thus expected that RNA from cells under stringent control would contain lower levels of modification than RNA from cells in complete and rich environments. Preliminary data support this hypothesis, with three modifications associated with stable RNA [Table 3] being absent from the stringent cells. Only one additional modification was observed in cells grown in rich media. At this point there is insufficient data to determine if RNA from stringent cells is hypomodified, or if the missing modifications merely fell under the limit of detection due to depletion of the stable RNA pool. Future work will aim at tightening up the experimental conditions to obtain a firm answer. At the same time, we will consider healthy and diseased cells to evaluate whether their state is reflected by the full complement of RNA modifications. It will be important to establish whether any variations of RNA modification, though no conclusions can be drawn at this point. Importantly, this data suggests that

<table>
<thead>
<tr>
<th>Negative Mode m/z</th>
<th>Normal</th>
<th>Stringent</th>
<th>Rich</th>
</tr>
</thead>
<tbody>
<tr>
<td>322.04403</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>323.02804</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>325.04369</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>336.05968</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>337.04369</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>339.0052</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>346.05526</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>347.03927</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>360.07091</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>362.05017</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>364.05459</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>376.06582</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>378.07024</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>382.0474</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>397.02844</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>397.02844</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>414.11786</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>460.10558</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>488.11825</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>491.09277</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Presence of nucleotide modifications in E. coli grown under normal, stringent, and nutrient-rich conditions. Data are qualitative and derived from exact mass and CRM data. An entry of one (1) indicates an extant modification, while an entry of zero (0) indicates an absence.
disruptions to RNA modification may be indicative of a large changes within the cell (such as cancer), and not due to transient changes (such as metabolic environment).

**Application of the Method 1- Bacterial Samples:** Over the years many methods have been used for classifying and differentiating bacteria on the basis of phenotypic and genetic information. Much of this data is nebulous at best: for example, the type strain of the bacterium *Proteus vulgaris* was recently reclassified as an entirely different species, *Proteus hauseri*, due to its lack of phenotypic and genetic similarity to any other strain of *P. vulgaris* [133]. Genetic methods, such as 16s rDNA sequence and DNA-DNA re-association have fallen under criticism over the years as they are either based on arbitrary cut-off points or provide a vastly different criterion for speciation that is used for plants and animals [Reviewed in 134, 135]. For these reasons, it was posited that the identification and classification of nucleoside modifications could assist in the phylogenetic differentiation of bacterial samples.

Bacterial samples from eight different strains and species of bacteria (see Methods section) were analyzed for nucleoside modifications by the same procedure as the *E. coli* above. Similarity between biological samples was qualified using the Gower distance [109]. Based on the reproducibility data gathered from *E. coli*, a similarity coefficient of 0.85 was taken to be the cut-off for similarity between species (See *Reproducibility Studies 3 - Assessing Reproducibility*, above). That is, samples would be considered to be of separate species if they had less than 85% congruence by the Gower method. A variety of bacterial species in the γ-proteobacteria were analyzed (see Methods section),

<table>
<thead>
<tr>
<th>Relatedness</th>
<th>Percent Congruity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Replicates</td>
<td>&gt;85%</td>
</tr>
<tr>
<td>Different Strain</td>
<td>53%</td>
</tr>
<tr>
<td>γ-Proteobacteria</td>
<td>42-44%</td>
</tr>
<tr>
<td>Unrelated</td>
<td>&lt;40%</td>
</tr>
</tbody>
</table>

*Table 4:* Relatedness of gammaproteobacteria based on a qualitative assessment of the presence or absence of nucleotide modifications. Two strains of *E. coli*, MG1655 and XL1B were investigated, and other samples were chosen based on their location in the phylogenetic tree (see *Experimental*).
with *Bacillus subtilis* and *Saccharomyces cerevisiae* serving as outgroups. Preliminary results show a wide difference between strains of the same species, and an extant, though slim, increase in difference between samples less closely related [Table 4].

Due to the small differences in relatedness between the example species (*E. coli*) and other γ-proteobacteria examined and the other γ-proteobacteria and each other, it is unlikely that this method will provide sufficient information to be used as a method for characterizing and placing new bacterial species within the existing phylogenetic tree. It is, however, capable of identifying known bacterial species and strengthening the assignment of sameness or differentness between them.

**Application of the Method 2 - Yeast Samples:** *Saccharomyces cerevisiae* is one of the simplest and most well-characterized eukaryotic organisms in existence. As such, it provides a convenient model for experimentation on genetics, epigenetics, etc. when a bacterial system is insufficient. For development of this method, yeast was used as a model for RNA fractionation, organelle fractionation, and viral isolation.

Isolation of total RNA from yeast was more difficult than expected. The cell wall proved resistant to detergents and all chaotropes tested, as well as vortexing and sonication (data not shown). In the end digestion with lyticase and lysis by French pressure in a mild detergent were required to fully lyse cells [136].

Yeast RNA modifications served as an out-group for the bacterial samples examined above.

RNA modification was also investigated by class of RNA. Ribosomal and transfer RNA were isolated from *Saccharomyces cerevisiae* for proof of principle, as the yeast has ribosomes of similar size and density to humans. The original procedure developed for plants by Cross [86] is able to isolate large quantities of
rRNA from yeast (data not shown). Due to the relatively low priority of this experiment, isolated rRNA has not yet been examined for modifications by mass spectrometry. Isolation of tRNA [87] was also explored with a modicum of success, but no MS data was recorded.

Subcellular fractionation techniques were investigated to localize modified RNA within the cellular environment, as there may be ribozymes within catalytic compartments such as lysosomes and the Golgi apparatus. Since bacteria do not have organelles, yeast was used as a model instead.

In retrospect, it is likely that the methods used for cellular disruption prior to subcellular fractionation were too vigorous. Tests for marker enzymes for all organelles examined show a distribution of enzyme that is concentrated at the top of the tube and attenuates down the first third of the gradient [Figure 14a]. It is likely that organelles were damaged from the repeated bouts of lysis by French pressure. Care must be taken in the future to lyse cells by more gentle methods such as mechanical homogenization in a Dounce homogenizer [61].
Isopycnic centrifugation with Percoll showed better separation of enzymatic activity [Figure 14b] than UCF with a sucrose gradient, likely due to the slower speeds, shorter run times, and more favorable osmotic environment. The sigmoidal curve of Percoll density in UCF [Figure 14c] proved to be a problem for zonal resolution, however, as appreciable separation of organelles was never attained. In future work, vertical and near-vertical centrifuge rotors, which have been shown to increase resolution of Percoll gradients [76] will be examined.

**Application of the Method**

3 - Viral Samples: RNA viruses comprise a large and diverse class of intracellular pathogens such as the ebolavirus and HIV. There is little known, however, about editing and modification of viral RNA in host cells. Editing of the sequence of RNA viral genomes has been investigated several times
[137, 138], but covalent modification has not. For viral genomes such as HIV where only in silico models based on NMR and MS3D constraints are available, RNA modification has significant implications for the veracity of these models. The presence or absence of RNA modifications in the HIV genome was investigated by the same method used for bacterial samples.

In order to establish a working method for scarce and infectious HIV samples, the yeast dsRNA virus L-A and its satellite M [60] were used. The intention was to gain experience working with RNA viruses that can be grown on a large scale before moving to time and resource intensive HIV. However, because L-A does not have an extracellular phase, it requires a much more thorough work-up than HIV and presented many more problems with handling than the virus for which it was meant to stand in. Early work with the L-A virus resulted in incomplete lysis of the yeast cells in several instances, but the prevalent problem was the tendency of L-A to aggregate irreversibly in oxidizing conditions. In the end, DTT concentrations of 2-10 times those reported in the literature were necessary for efficient preparation of samples. Despite eventually being able to isolate both L-A and M on a gel [Figure 15a], no mass spectrometry data was obtained from either virus. The most important information gained from the investigation of L-A is that traditional RNA extraction procedures can remove capsid proteins from T=1 icosahedral viruses such as L-A and HIV-1 [Figure 15a, lane 3]
HIV-1, on the other hand, did not present such problems for isolation. After centrifugation, RNA was extracted from the viral pellet by treatment with a chaotropic solution and phenol/chloroform extraction. Several methods were employed, and all were ~25% efficient (see Methods section for details).

The purity of RNA extracted from HIV was examined by agarose gel electrophoresis. As can be seen in [Figure 15b] above, there are two bands observed. The top band corresponds to an HIV dimer, and the bottom one to the monomer. Due to limitations in the ability of an agarose gel to resolve RNA or largely different sizes, such as the commonly co-packaged ~70 nt tRNA or ~300 nt 7SL RNA [103] and the ~9000 nt HIV genome, a number of techniques were explored to overcome this problem. First a time-course study was done on a traditional agarose gel, where the gel was infused with ethidium bromide and imaged at several points during the run. This failed to produce results, however, as the HIV genome migrates too slowly and the gel was left devoid of ethidium bromide before it resolved. Little more success was seen using an Agilent 2100 Bioanalyzer, where no heterologous RNA was seen in the sample, but the HIV genome itself could not be resolved. Indeed, the electropherogram showed only one large peak at ~1200 nt which could not be identified (data not shown).

Attempts were made to purify HIV genomic RNA from total extracted RNA by affinity capture on a solid support. Several
linker chemistries were explored (see Methods section) before a disulfide link between magnetic beads and capture oligos were decided upon.

The efficacy of the affinity capture was tested first by synthetic DNA oligonucleotides of ~25 and ~100 bases [Figure 16a]. The ability of the affinity capture method to pull down larger nucleic acids was investigated as well by over-expressing the HIV 5'-UTR in E. coli. Two plasmids were used, one which produced RNA of ~270 nt, and one that produced RNA of ~10000 nt. The affinity capture method was able to isolate RNA from either plasmid, as shown in Figure 16b. Attempts to isolate RNA from HIV extracts were less successful, with no RNA being recovered according to Bioanalyzer data (not shown), despite conflicting UV/Vis data.

Despite these problems, HIV RNA was digested and tested on the mass spectrometer for proof of principle. Full spectra look similar to those collected for other organisms [Figure 17], save for the high levels of DNA contamination, likely from plasmid rejected or released by cells in culture. Further analysis shows a number of covalent modifications present [Table 5]. At this point, however, it is not possible to attribute these

<table>
<thead>
<tr>
<th>Mass</th>
<th>Species Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>322.04402</td>
<td>A</td>
</tr>
<tr>
<td>323.02804</td>
<td>A</td>
</tr>
<tr>
<td>325.04369</td>
<td>A</td>
</tr>
<tr>
<td>336.0596</td>
<td>C</td>
</tr>
<tr>
<td>337.04369</td>
<td>C</td>
</tr>
<tr>
<td>338.0211</td>
<td>A</td>
</tr>
<tr>
<td>346.05525</td>
<td>A</td>
</tr>
<tr>
<td>347.03927</td>
<td>A</td>
</tr>
<tr>
<td>350.03894</td>
<td>A</td>
</tr>
<tr>
<td>351.05934</td>
<td>B</td>
</tr>
<tr>
<td>352.05459</td>
<td>A</td>
</tr>
<tr>
<td>353.02084</td>
<td>C</td>
</tr>
<tr>
<td>360.07090</td>
<td>C</td>
</tr>
<tr>
<td>361.0549</td>
<td>C</td>
</tr>
<tr>
<td>362.05017</td>
<td>A</td>
</tr>
<tr>
<td>364.05459</td>
<td>B</td>
</tr>
<tr>
<td>364.09097</td>
<td>A</td>
</tr>
<tr>
<td>365.08360</td>
<td>B</td>
</tr>
<tr>
<td>368.03174</td>
<td>A</td>
</tr>
<tr>
<td>374.08655</td>
<td>B</td>
</tr>
<tr>
<td>376.0658</td>
<td>C</td>
</tr>
<tr>
<td>378.07024</td>
<td>B</td>
</tr>
<tr>
<td>380.04950</td>
<td>A</td>
</tr>
<tr>
<td>381.03352</td>
<td>A</td>
</tr>
<tr>
<td>382.04739</td>
<td>A</td>
</tr>
<tr>
<td>386.05017</td>
<td>A</td>
</tr>
<tr>
<td>388.0658</td>
<td>A</td>
</tr>
<tr>
<td>390.08147</td>
<td>C</td>
</tr>
<tr>
<td>394.06515</td>
<td>B</td>
</tr>
<tr>
<td>395.04917</td>
<td>C</td>
</tr>
<tr>
<td>397.02843</td>
<td>A</td>
</tr>
<tr>
<td>400.0658</td>
<td>A</td>
</tr>
<tr>
<td>404.09712</td>
<td>B</td>
</tr>
<tr>
<td>406.05863</td>
<td>A</td>
</tr>
<tr>
<td>409.06482</td>
<td>B</td>
</tr>
<tr>
<td>414.11785</td>
<td>A</td>
</tr>
<tr>
<td>414.94456</td>
<td>A</td>
</tr>
<tr>
<td>424.0757</td>
<td>C</td>
</tr>
<tr>
<td>426.0372</td>
<td>A</td>
</tr>
<tr>
<td>428.09712</td>
<td>A</td>
</tr>
<tr>
<td>430.11277</td>
<td>A</td>
</tr>
<tr>
<td>438.06621</td>
<td>A</td>
</tr>
<tr>
<td>447.06655</td>
<td>A</td>
</tr>
<tr>
<td>450.13898</td>
<td>A</td>
</tr>
<tr>
<td>460.10558</td>
<td>A</td>
</tr>
<tr>
<td>488.11825</td>
<td>A</td>
</tr>
<tr>
<td>491.0927</td>
<td>A</td>
</tr>
<tr>
<td>537.08048</td>
<td>A</td>
</tr>
<tr>
<td>558.06384</td>
<td>B</td>
</tr>
<tr>
<td>574.05876</td>
<td>B</td>
</tr>
<tr>
<td>603.14519</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 5: Modifications observed in HIV-1 based on CRM data. The mass of the identified NMP is listed on the left. The right column indicates the species detected. An entry of “A” indicates the 2'-OH species, “B” indicates the 2'-methyl species, and “C” indicates both.
modifications to the HIV-1 genome or heterologous RNA packaged with the virus. Future work will be directed toward refining both capture and washing procedures to ensure that only the desired target is actually pulled down and isolated.

**Isolation of Cross-Linked RNA:** In addition to eventually enabling the investigation of viral RNA modifications, we expect that the methods for RNA isolation will support the structural elucidation of the 5’-UTR by footprinting and cross-linking probes. The HIV-1 5’-UTR contains the dimerization initiation site and the packaging signal, and is posited to contain a riboswitch that regulates the function of the whole RNA between mRNA and genomic forms. As mentioned in the introduction, the 5’-UTR has proven difficult to examine with traditional high-resolution analytical techniques (such as NMR and X-ray crystallography), and thus MS3D is the primary option. Since HIV-1 samples were considered too scarce for method development, similar RNA from modified *E. coli* was used as a stand-in. *E. coli* does not secrete the RNA it produces, so the procedure is most similar to exploring the structure of HIV-1 RNA within the human cells, which is where it is most likely to exhibit both putative conformations in any case. Three different cross-linkers were explored: nitrogen mustard (HN2), formaldehyde, and *trans*-platinum.

All three species of cross-linker provided extensive cross-linking both *in vivo* and *in vitro*. RNA extracted from untreated cells displays a discrete banding
pattern on an agarose gel [Figure 18a, lane 2], while RNA extracted from cells treated with (for example) nitrogen mustard show a long smear [Figure 18a, lane 3].

The affinity capture procedure previously developed has shown its ability to purify not only synthetic DNA oligos [Figure 18b, lane 1], but also large RNA constructs from whole-cell extracts [Figure 18b, lane 2]. Cellular RNA cross-linked with nitrogen mustard in vitro could also be purified [Figure 18b, lane 3].

Application of the affinity capture procedure developed on in vitro samples encountered several difficulties. RNA-protein cross-links precluded the use of traditional liquid-phase extraction techniques [Figure 18c] for the isolation of RNA, as the RNA would not partition to the aqueous phase. Several alternative methods for lysing cells and purifying RNA were explored [Figure 18a] including direct lysis and analysis in a 2% SDS solution and purification of RNA-protein complexes by precipitation with ethanol and ammonium sulfate. As can be seen in Figure 18a, there is virtually no difference between the three methods when applied to cross-linked samples, though only the first method provides full extraction of RNA from blank samples. Ethanol precipitation was pursued from this point, as the affinity capture failed to work on DNA oligonucleotide standards in the presence of ammonium sulfate and SDS.
Unfortunately, affinity capture did not work for any samples cross-linked \textit{in vivo} (data not shown), regardless of how they were prepared.

Time-course and concentration studies were performed for all three cross-linking compounds. None of the concentrations and timeframes investigated provided positive results for bacterial samples. However, all three cross-linkers were tested on HEK cells transfected with nL4-3 and VSV-G (i.e. cells that were expressing the pseudotyped virus). All three provided a modicum of cross-links, which were able to be extracted by traditional liquid-phase extraction procedures (data not shown). Formaldehyde provided the most extensive cross-linking, but may not be used in the future due to its inherent problems with mass spectrometry.

**Conclusions**

This study represents a first stab at establishing a comprehensive experimental workflow for the characterization of \textit{in vivo} RNA modifications regardless of their nature and provenance. Electing to employ MS detection without front-end separations poses numerous challenges. Indeed, we found that the high performance afforded by MS technologies cannot alone compensate for possible weaknesses of sample preparation and handling operations. HDMS and IM-MS have shown excellent potential for assessing the state of RNA modification in biological samples, regardless of source (bacteria, yeast, human, or virus). However, more effort will be clearly necessary to optimize the isolation of target RNA from different environments. Indeed, any isolation step must afford a sufficient level of specificity to enable unambiguous assignment of observed modifications to the respective class of RNA. In this direction, the application of MS approaches has raised the bar on the need to resolve RNA samples according to their cellular localization. Proper harmonization of
the isolation step will require retaining a MS-friendly character and preserving the ability to perform further analysis to identify the sequence position of observed modifications.

The investigation of biological samples poses challenges of its own. For example, it is very difficult to evaluate the reproducibility of an analytical protocol when the reproducibility of the biological standard itself is not firmly established. Even for the same organism, the significance of culture-to-culture or cell-to-cell deviations is not well established. This statement is particularly true in light if the lack of understanding of the functions performed by the vast majority of natural RNA modifications. Our preliminary metabolic studies indicate that RNA modification may not be affected by external changes in the growth environment, at least in *E. coli*. This finding will certainly require confirmation not just in this organism, but in many others. What is more important, however, it that we are now in the position of asking these types of questions. Our workflow will enable us to investigate whether the presence/absence of certain modifications, or the relative levels of different modifications, are affected in any ways by specific external stimuli (e.g., drugs and environmental xenobiotics), organism health (e.g., viral infection or cancer), or general epigenetic state. The answers to these questions will determine whether whole transcriptome analysis of RNA modifications will develop into a valid phylogenetic and diagnostic tool.
Materials and Methods

Unless otherwise noted, all reagents were purchased from Sigma Aldrich (St. Lois, MO) and were reagent-grade or better.

Analysis of Nucleotide Standards: Nucleotide standards were purchased from Sigma (dCMP, dTMP, dAMP, dGMP, CMP, UMP, AMP, GMP) or Dharmacon (Lafayette, CO) (ΨMP) and used without further purification. These analytes correspond to the expected products from enzymatic digestion of whole-cell RNA extracts [30, 139]. Stock solution concentrations were adjusted to 1 mM as determined by UV absorbance at 260 nm using extinction coefficients from the literature [140]. Prior to analysis, stock solutions were combined in equimolar amounts and diluted to 9.0 uM total NMPs in 100 mM ammonium acetate and 10% isopropanol.

Samples were analyzed by direct infusion on either a Thermo Finnegan Scientific (West Palm Beach, CA) LTQ-Orbitrap Velos mass spectrometer, or a Waters (Milford, MA) Synapt G2 HDMS IMS mass spectrometer. All analyses were performed on the nano-scale using pulled quartz emitters produced in house by a Sutter Instruments Co. (Novato, CA) P2000 laser puller. Up to 4 µL samples were loaded into each needle with a gel-loader pipette tip. A stainless steel wire was inserted in the back-end of the emitter to apply an electrospray voltage that between 0.6 and 1.4 kV. Source temperature and desolvation conditions were adjusted by closely monitoring the presence of ammonium adducts and water clusters [141].

For high-resolution analysis, the LTQ-Orbitrap instrument was calibrated using the Thermo-recommended anionic calibration solution, which included SDS, sodium taurocholate, and Ultramark 1621. The standard solution allowed a calibration from 50-2000 m/z with an average mass accuracy of ~386 ppb. Tandem mass spectrometry was accomplished by isolating the ion of interest in the LTQ element and activating by reaction with N₂ gas. The resulting fragments were
analyzed either in the LTQ or Orbitrap element. In all cases data were processed using the Thermo XCalibur 2.1 software package supplied with the instrument. Exact mass calculations and elemental composition calculations were performed with the molecular weight calculator made available by Pacific Northwest National Laboratories [113] instead of that supplied with the XCalibur software.

For drift-time analysis, the Synapt G2 was calibrated from 200-1200 m/z with a CsI mixture (2 mg/mL in 50/50 water/methanol) which afforded ~9 ppm mass accuracy. Drift time (tD) was determined by allowing ions to travel through the traveling-wave (Tri-WAVE) element of the IM-MS instrument, followed by mass analysis in single-reflectron (“Sensitivity”) mode. The Tri-WAVE region was held at a pressure of approximately 4.40 mbar (uncalibrated gauge reading) by a 90 mL/min flow of N2 and 180 mL/min of He. It was operated with an approximately 650 m/s IMS wave velocity, a 40 V wave height, a 109 m/s transfer wave velocity, and a 2.0 V transfer wave height. Analysis of mass-selected isobars was performed by raising the cell pressure to ~4.60 mbar (uncalibrated gauge reading) with a flow of 140 mL/min N2 and 180 mL/min He. At the same time, wave velocity was raised to ~700 m/s, transfer wave velocity to ~600 m/s, and transfer wave height to 4.0 V. Observed mobility profiles were compared to theoretical Gaussian distributions by a curve-fitting algorithm from the PeakFit 4.1 software package.

Time-alignment parallel (TAP) fragmentation of all ions travelling through the Tri-Wave element with different drift times was performed by CID in the transfer region of the instrument, between the IMS cell and the time-of-flight (TOF) analyzer [126]. Fragmentation energies between 10-25 V were explored. Data were collected under the same conditions listed above. Heat-map representations of data obtained from full-range MS and TAP experiments were produced in the form of mass-over-charge ratio (m/z) vs. tD vs. relative intensity plots by the DriftScope v2.1.
software provided with the Waters’ MassLynx 4.1 package. Diagnostic fragments were assigned to their mobility-separated parent ions by investigating $t_0$ transects of the data.

**Analysis of Modified NMPs:** The applicability of the technique to modified nucleotides was investigated using modified NMP standards and cellular extracts from *E. coli*.

Modified NMP standards were produced by the hydrolysis of a selection of methylated CTP derivatives ($m^4$CTP, $m^5$CTP, CmTP) purchased from Trilink (San Diego, CA). Dephosphorylation was carried out by hydrolysis by incubating the NTP mix in 100 mM HCl at 37 C for 5 minutes [142, 143], and was >75% effective. Methylated cytidine nucleosides were produced by incubating mCTPs with bacterial alkaline phosphatase (BAP) (0.5 units, 1/10 volume 1 M ammonium bicarbonate) or cow-intestine phosphatase (CIP) (1 unit, 1x NEB Buffer 3) for 1 hour at 37 C. Cow-intestine phosphatase was purchased from New England Biolabs (Ipswitch, MA). This procedure was <20% effective at removing all phosphates, but >75% effective at creating NMPs.

Cellular extracts were produced from MG1655 *E. coli* supplied by Robert Osuna. RNA isolation was accomplished by many methods over the years, including TRIzol (Invitrogen [Carlsbad, CA]), Tri-Reagent (Sigma), ToTALLY RNA Kit (Ambion [Carlsbad, CA]), and the Gough method [57]. Each method was derived from an established extraction procedure [56, 57], all of similar origin. Briefly, cells were lysed in a high-salt buffer of near-neutral pH. This buffer contained a chaotrope (guanidinium, urea), a detergent (sarcosine, SDS), and often a chelating agent (EDTA) or reducing agent (BME) to disrupt cell walls and denature RNase and other proteins. The mixture was then subjected to multiple rounds of liquid-phase extraction using acid phenol/chloroform. RNA was concentrated by precipitation with
a salt (sodium or ammonium acetate) and an organic (ethanol or isopropanol). All methods used produced similar RNA yields, within a factor of two (2). RNA was purified from DNA by digestion with DNase I (Fermentas [Vilnius, Lithuania]) at 37 C for one hour, followed by a second round of precipitation. Purified RNA was re-suspended in diethylpyrocarbonate (DEPC)-treated water or an appropriate buffer. The final samples collected were prepared almost entirely by using the Ambion ToTALLY RNA extraction kit.

For analysis of RNA modifications, digestion of whole-cell RNA was accomplished by exonuclease digestion based on an established procedure [31]. After re-suspension of RNA, 1/10 volume of 100 mM ammonium acetate was added, and 2 units of nuclease P1 were added. The sample was incubated for two hours at 37 C. It was then removed and 1/10 volume of 1 M ammonium bicarbonate added. Phosphodiesterase was added (0.002 units), and the sample was placed back at 37 C for two hours. At this point the samples are ready analysis (pending dilution).

Mass spectrometric analysis was carried out as for the standards.

Assessment of Reproducibility: The reproducibility of these techniques for biological samples was investigated using E. coli strain MG1655. E. coli were grown to a consistent optical density in lysogeny broth [144] at 37 C with vigorous shaking (150-200 RPM), RNA extracted, and digested according to established procedures (see above). An optical density between 1.2 and 1.5 provided the most reproducible results.

Samples were diluted in 100 mM ammonium acetate with 10% IPA to a final concentration of ~10 uM total NMPs. Direct infusion from quartz needles onto the Orbitrap provided a strong signal for the canonical nucleotides, as well as a number of smaller peaks for investigation. Instrumental conditions were similar to Part 1, except as noted below.
Spectra were collected in the FT-Orbitrap for a total of 30 scans per sample. The maximum inject time was set to 500 ms, which was high enough that all samples reached the AGC limit of the instrument \(10^6\) before the inject time had elapsed. Five microscans were averaged per analytical scan, resulting in a scan time of \(\sim10\) seconds. TICs from the acquisition were averaged using the Thermo Xcalibur 2.1 software. Full spectrum lists were exported with the following data for each point: exact mass to 5 decimal places, intensity (counts) relative intensity (percent), noise (counts) and baseline (counts). The resulting data were stored in a comma-separated variable (CSV) file. Masses not in the charge state of \(Z=-1\) were discarded at this stage, and the remaining masses were compared to a reference file of known modifications compiled from the SUNY Albany database [15] and the University of Warsaw Modomics database [145-147]. Comparisons were done using a Microsoft Excel 2010 macro called DataSnack, developed in-house by Dr. Daniele Fabris. The accepted limit of mass error was determined by making a standard curve for the four canonical nucleotides (CMP, UMP, AMP, and GMP) and applying an adjustment factor to all returned masses. Putative matches were organized into a new data file with their respective total and relative intensities. Potential matches at this point were accepted if their signal-to-noise ratio exceeded two (2). The resulting mass list was compiled for all technical and biological replicates.

Since some biological samples returned AMP as the base peak in MS analysis and some returned GMP, the reported intensities were standardized by redefining the relative intensities of each peak in terms of the combined number of counts for all four canonical NMPs (CMP, UMP, AMP, and GMP). Technical replicates were combined by determining the relative standard deviation (RSD) between modified relative intensities for all five replicates. Species with an RSD greater than 50%
(signal-to-noise ratio [SNR] less than 2, see below) were discarded. This final list was used as the peak list for each biological replicate.

**Analysis of Bacterial and Yeast Samples:**

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>Source</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candida Albicans</td>
<td>ATCC 10231</td>
<td>ATCC</td>
<td></td>
</tr>
<tr>
<td>Saccharomyces Cerevisiae</td>
<td>W303-1B, ATCC 201238</td>
<td>ATCC</td>
<td></td>
</tr>
<tr>
<td>Saccharomyces Cerevisiae</td>
<td>K12, ATCC 28683</td>
<td>ATCC</td>
<td>Expresses K1 phenotype</td>
</tr>
<tr>
<td>Escherichia Coli</td>
<td>MG1655</td>
<td>Robert Osuna</td>
<td>Wild Type</td>
</tr>
<tr>
<td>Escherichia Coli</td>
<td>XL18</td>
<td>Alexander Shekhtman</td>
<td>Competent</td>
</tr>
<tr>
<td>Escherichia Vulneris</td>
<td>ATCC 33821</td>
<td>ATCC</td>
<td>Type Strain Derivative</td>
</tr>
<tr>
<td>Serratia Marcescens</td>
<td>RO429</td>
<td>Robert Osuna</td>
<td></td>
</tr>
<tr>
<td>Salmonella Typhimurium</td>
<td>RJ1826</td>
<td>Robert Osuna</td>
<td></td>
</tr>
<tr>
<td>Klebsiella Aerogenes</td>
<td>RO284</td>
<td>Robert Osuna</td>
<td></td>
</tr>
<tr>
<td>Proteus Hauseri</td>
<td>RO528</td>
<td>Robert Osuna</td>
<td>Type Strain</td>
</tr>
<tr>
<td>Bacillus Subtilis</td>
<td>PY79, ATCC 6051</td>
<td>Mariene Belfort</td>
<td>Type Strain Derivative</td>
</tr>
</tbody>
</table>

**Table 6:** Bacterial and yeast genus, species, and strains used in this investigation.

A variety of prokaryotic and eukaryotic cells were examined in this investigation. Several strains of bacteria and yeast were used, as shown above [Table 6]. Bacterial species were chosen based on their relatedness, as shown below [Figure 19] [148].

Bacterial samples were grown overnight for general analysis or to early stationary phase ($OD_{600}$ 1.2-1.5) for reproducibility studies. Except for metabolic

![Phylogeny of bacterial and yeast species examined. Since E. coli was used as the standard, other samples were chosen to build a phylogenetic tree accordingly [148].](image)
studies, all bacteria were grown in lysogeny broth (1% tryptone, 1% NaCl, 0.5% yeast extract in distilled water) at 37 C with vigorous shaking (150-200 RPM). RNA was extracted from all bacterial types by either the Gough procedure, or with the Ambion ToTALLY RNA extraction kit (see above).

Yeast samples were grown to saturation in YEPD medium (2% peptone, 2% dextrose, 1% yeast extract in distilled water) at 37 C with moderate (~100 RPM) shaking. RNA extraction from yeast was begun by partial digestion of the yeast by lyticase (40 units/mL of overnight culture in TE) for 30 minutes at room temperature. The resulting sphereoplasts were collected by centrifugation (800 g, 20 minutes) and re-suspended in 5 mL of extraction solution A from the Gough procedure (10 mM Tris-HCl [pH 7.5], 150 mM NaCl, 1.5 mM MgCl₂, 0.65% NP-9) [57]. The cells were then lysed by two passages through a French pressure cell and extraction was finished according to an established procedure [57].

Metabolic studies on E. coli were carried out using two types of media to produce different metabolic environments. Davis minimal media [149] was used to induce the stringent response, and terrific broth (1.2% tryptone, 2.4% yeast extract, 0.4% glycerol with 17 mM KH₂PO₄ and 72 mM K₂HPO₄) was used as a nutritionally rich medium. The control group was grown in lysogeny broth, which is a complete medium but is nutritionally poor due to its original purpose [144].

Subcellular Fractionation: Wild-type Saccharomyces cerivisiae were grown to saturation in 1 L YEPD at 37 C. The cells were collected by centrifugation at 6000 g for 20 min in a Beckman Coulter (Brea, CA) J-series centrifuge with a JA-10 rotor. The cells were re-suspended in 150 mM NaCl and lysed by two passages through a French pressure cell. The lysate was cleared by centrifugation at 9600 g for 10 minutes. The supernatant was collected for fractionation.
Subcellular fractionation was accomplished by ultracentrifugation, which was approached from two different avenues.

Sucrose density gradient centrifugation was performed both with continuous and step gradients. In each case a 1000 uL micropipette was used to layer a series of sterile sucrose solutions from 15-65% w/v (in 5% increments) on top of one another. For continuous gradients, the layered solution was allowed to sit overnight at 4°C. Continuous gradients were also poured directly with an HPLC pumping two solvents, sterile water and 65% sucrose, into a mixer and then into the ultracentrifuge tube. This method was eventually abandoned due to the inability of the HPLC to effectively pump the viscous sucrose solution.

Samples were layered on top of the gradient with a micropipette and spun in a Beckman Coulter S- or R-series ultracentrifuge in an SW-28 or SW-40 swinging-bucket rotor at 120000 g for 16-20 hours. Fractions were collected with a micropipette and stored in 0.5-1.0 mL portions in a 96-well plate. Density of the solution was determined by refractive index.

Percoll gradient centrifugation was performed by mixing 9 volumes of Percoll with one volume of 1.5 M NaCl to make the solution isosmotic to the samples. Cleared cell lysates were mixed with the Percoll solution and centrifuged at 30000 g for 15 minutes in a Beckman S- or R-series ultracentrifuge with a 40ti fixed-angle rotor. Fractions were collected and analyzed as above. Different concentrations of Percoll solution were tested by diluting isosmotic Percoll with 150 mM NaCl.

Fractions from the ultracentrifugation separation were analyzed for the presence of organelles by a variety of procedures. Mitochondria were detected by assaying for succinate dehydrogenase by an established procedure [150]. Lysosomes were identified by the presence of acid phosphatase, which was assayed for by an established procedure [150]. Peroxysomes were detected by a qualitative
hydrogen peroxide test to assay for catalase and peroxidase. The appearance of bubbles was taken to indicate the existence of peroxisomes in that fraction. Other organelles were not assayed for.

**RNA Fractionation by Class:** RNA was fractionated into the three main classes by established procedures. Transfer RNA was collected from total RNA by differential precipitation as previously described [87]. Ribosomal RNA was collected from pelleted ribosomes [86] by SDS-phenol/chloroform extraction [57]. Messenger RNA was purified from total eukaryotic RNA by poly dT affinity capture [84].

**Collection of Viral Samples:** The primary stand-in for HIV genomic RNA was artificial 5′-UTR extracted from *E. coli* transformed with a plasmid containing the HIV-1 5′-UTR sequence under the control of the T7 bacteriophage promoter and the LAC operon. The bacteria were grown to saturation and induced with isopropyl β-D-1-thiogalactopyranoside (IPTG), causing them to produce 5′-UTR which was extracted by an established procedure [57]. Since the plasmid used to produce 5′-UTR did not contain a stop sequence for T7 RNA polymerase (RNAP), one was added by the Invitrogen GeneArt service. The mutated plasmid produced RNA of ~270 nt, while the original plasmid produced RNA of ~10 Kb. In the latter case the RNAP traveled around the plasmid multiple times before terminating transcription.

Isolation of the L-A virus from yeast was accomplished by an established procedure [60], with the following modifications. Digestion with Zymoylase (or lyticase) was skipped in favor of a second passage through the French pressure cell. All dithiothreitol (DTT) concentrations were doubled due to problems with irreversible aggregation of the virus. RNA was extracted from viral particles with Tri-Reagent, ToTALLY Kit, or Gough procedure, all with similar results.

HIV-1 samples were collected from transfected human embryonic kidney (HEK) cells (293T line) supplied by Mario Canki.
The 293T cells were plated at 1.5 million cells per 100 mm plate and grown overnight in 10 mL Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal bovine serum, glutamine, and antibiotics (“full DMEM”) (all from GIBCO [Carlsbad, CA]). The following day they were transfected with 10 ug nL4-3 and 15 ug VSV-G for six hours in 5 mL full DMEM using the Promega (Fitchburg, WI) Profection Mammalian Transfection Kit. Some transfections were carried out without VSV-G to produce less dangerous viral particles, albeit in lesser amounts.

After transfection, cells were grown in 10 mL full DMEM on 100 mm plates, 37°C with 5% CO2. At the end of 36-40 hours, media containing viral particles was removed and purified of cells and debris by vacuum filtration through a Millipore (Billerica, MA) Steriflip system and frozen at -80°C until used. To collect virions, viral samples were thawed, and the virus pelleted by centrifugation. Four hours is the standard reported in the literature [151], but it was determined that six hours provided a larger and more tightly packed viral pellet and was adopted as the standard. DMEM was decanted and the viral pellet was re-suspended in 1 mL TRIzol (Invitrogen), 400 uL extraction solution B [57], or 400 uL denaturation solution (Ambion ToTALLY Kit) and the RNA purified by phenol/chloroform extraction and ethanol/isopropanol precipitation with ammonium or sodium acetate. Choice of extraction procedure and the organic/salt for precipitation did not have a noticeable effect on the recovery of viral RNA, which was ~25% efficient in all cases.

Transfection with both the nL4-3 and VSV-G plasmids proved stressful for 293T cells, and the standard transfection time of 12 hours killed many cells outright, or caused them to die during viral propagation. Likewise, viral outgrowth for 48 hours proved too stressful for the cells, as they would detach and become problematic during filtration. Thus cells were transfected for 6 hours and grown for 40 before harvesting the virus. Cells had a tendency to overgrow during viral
propagation when plated at a starting density of greater than 1.5^6 per plate 24 hours before transfection to reach 30-40% confluence by the next day. Do to the hyper-infectious nature of the N/V virus, it was eventually decided to abandon pseudotyping and transfec... cells, as seen by cell death rates 36 hours post-transfection.

**Affinity Capture:** Purification of HIV-1 genomic RNA from heterologous packaged RNA was accomplished by affinity capture. Affinity capture oligonucleotides were immobilized on superparamagnetic beads by a variety of chemistries. The original procedure used carboxyl-coated beads from Qiagen (Venlo, Netherlands) with affinity capture oligos modified with an amine group on the 5’ end. The carboxyl groups were activated with 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDAC) and reacted with the oligos, which resulted in a low yield and damage to the capture oligos, as EDAC-activated carboxylic acids will react with amines in nucleobases, albeit at a lower rate [152].

The most successful linker chemistry used was a disulfide bond, based on a number of established procedures [153, 154]. BioMag amine-derivatized superparamagnetic beads (Polysciences [Warrington, PA]) were derivatized to thiol moieties by treatment with 10 bead-volumes of 2-iminothiolane (Traut’s Reagent) in phosphate-buffered saline with 10 mM EDTA (PBS+EDTA) for 1 hour at room temperature [Figure 20a and b]. For small amounts of beads, the storage solution was not removed prior to derivatization due to loss of bead material. For volumes larger than 5 uL, the beads were separated from the solvent with a magnet and the storage solution was pipetted out.

Thiols were reduced by incubation with 50 mM DTT for one hour at room temperature. After washing with water and dimethyl formamide (DMF), the thiols are activated by treatment with 10 mM Ellman’s reagent (in 100 mM sodium...
phosphate, pH 8.0 with 1 mM EDTA) for one hour at room temperature [Figure 20c]. If the reaction has succeeded, the solution turns yellow at this time [155].

Final derivatization with affinity capture oligodeoxyribonucleotides is accomplished by treating the beads with thiol-modified oligos from IDT (Coralville, IA) (100 uM in water) [Figure 20d]. The solution will again turn yellow if this step is successful. The final incubation may be left overnight at room temperature. A final incubation of the beads with a “junk” oligo to neutralize any sulfur radicals or nucleophiles left on the beads, which would otherwise degrade the target RNA, is optional.

For affinity capture, target RNA (in water) is mixed with capture beads and an equal volume of RNase-free 20 mM NaCl. The sample is heated to 70 C and slowly cooled to 50 C over about 9 minutes (ramp rate of 1% on an Applied Biosystems [Carlsbad, CA] Veriti 96-Well Thermocycler, ~0.039 C/second change). This cycle is repeated seven times. After capture, the beads are washed with 10 mM NaCl, and then re-suspended in 95% formamide with 10 mM EDTA and heated to 95 C for 5 minutes to release bound RNA. 7 M urea was also used to elute RNA, with a slightly better effect.

Figure 20: The process of creating affinity capture beads. Amines are derivatized to thiols (a and b). Thiols are activated with Ellman’s Reagent (c) and finally attached to affinity capture oligonucleotides (d).
The identity and purity of isolated HIV RNA was assessed by a number of methods. First among them was UV/Vis with a Thermo Scientific Nanodrop 2000-C, which was used to measure concentrations and assess purity by the A_{260/280} ratio. Agarose gel electrophoresis (1% agarose in TAE, 150 V) was performed to visualize genomic RNA and assure its integrity. An Agilent Technologies 2100 Bioanalyzer (Santa Clara, CA) was used to test the RNA samples for purity relative to heterologous RNAs due to its high dynamic range and resolution.

**Development of Cross-Linking:** Method development for the cross-linking of HIV-1 5' UTR was investigated by many means. *E. coli* strain T7 Express from NEB was transformed with a plasmid containing the HIV-1 5’-UTR under the control of the T7 bacteriophage promoter and the *Lac* operon. *E. coli* were grown in LB broth with ampicillin (100 ug/mL) at 37 C to saturation and induced for 2 hours with 1 mM IPTG. If *in vitro* cross-linking was desired, RNA was extracted at this point by any of the procedures listed above. Otherwise cross-linking was begun.

Three different cross-linkers were explored: nitrogen mustard (HN2), trans-platinum, and formaldehyde. Cross-linking with HN2 was accomplished by mixing a stock solution (1 M in DMF) with the bacterial culture directly. A final concentration between 2 uM and 8 mM was used depending on the goal. HN2 has a half-life of 15 minutes in water, so cross-linking times between 5 minutes and one half hour were explored. *Trans*-platinum was added to cell cultures to a final volume of 1-16 uM. Since *trans*-platinum is activated in water, stock solutions were made in DMSO. Cross-linking proceeded for 0.5-4 hours. In some cases *trans*-platinum was activated in water before mixing with cell culture. In these instances activation lasted one hour and cross-linking lasted 0.5-2 hours. Formaldehyde was used in concentrations of 1, 2, or 4%. Cross-linking was performed for between 15 and 120 minutes.
For successful cross-linking of transfected HEK cells, after 36 hours of growth, media was collected and the cells were covered with 5 mL warm PBS. For formaldehyde cross-linking, concentrated formaldehyde was added to a final concentration of 1%. Cultures were incubated for 10 minutes at room temperature, then PBS was decanted. Cells were washed with warm PBS and lysed in 1 mL Tri-Reagent solution. RNA was extracted as per the Tri-Reagent procedure [156]. For HN2 cross-linking, 10 mM HN2 in 100 mM HCl was added to a final concentration of 2 uM. Cross-linking was allowed to proceed for 30 minutes, and then PBS was decanted. Cells were washed with warm PBS and lysed in 1 mL Tri-Reagent solution. RNA was extracted as per the Tri-Reagent procedure [156].
References


94. Laughrea, M., Jetté, L.: A 19-nucleotide sequence upstream of the 5’ major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. Biochemistry 33, 13464–13474 (1994).
120. Schwartz, J. C., Syka, J. E. P., Quarmby, S. T. The 53rd ASMS Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, June 5-9, 2005


