RoboOligo: Software for Mass Spectrometry Data to Support Manual and *De Novo* Sequencing of Post-transcriptionally Modified Nucleic Acids

Paul J. Sample[1]*, Kirk W. Gaston[2], Juan D. Alfonzo[1,3], and Patrick A. Limbach[2]*

[1] Department of Microbiology and The Center for RNA Biology, The Ohio State University, Columbus, OH 43210
[2] Rieveschl Laboratories for Mass Spectrometry, Department of Chemistry, PO Box 210172, University of Cincinnati, Cincinnati, OH 45221-0172
[3] Ohio State Biochemistry Program, The Ohio State University, Columbus, OH 43210

* To whom correspondence should be addressed. Tel: 1-513-556-1871; Fax: 1-513-556-9239; Email: Pat.Limbach@uc.edu. Correspondence may also be addressed to Paul J. Sample Email: pjsample@gmail.com.

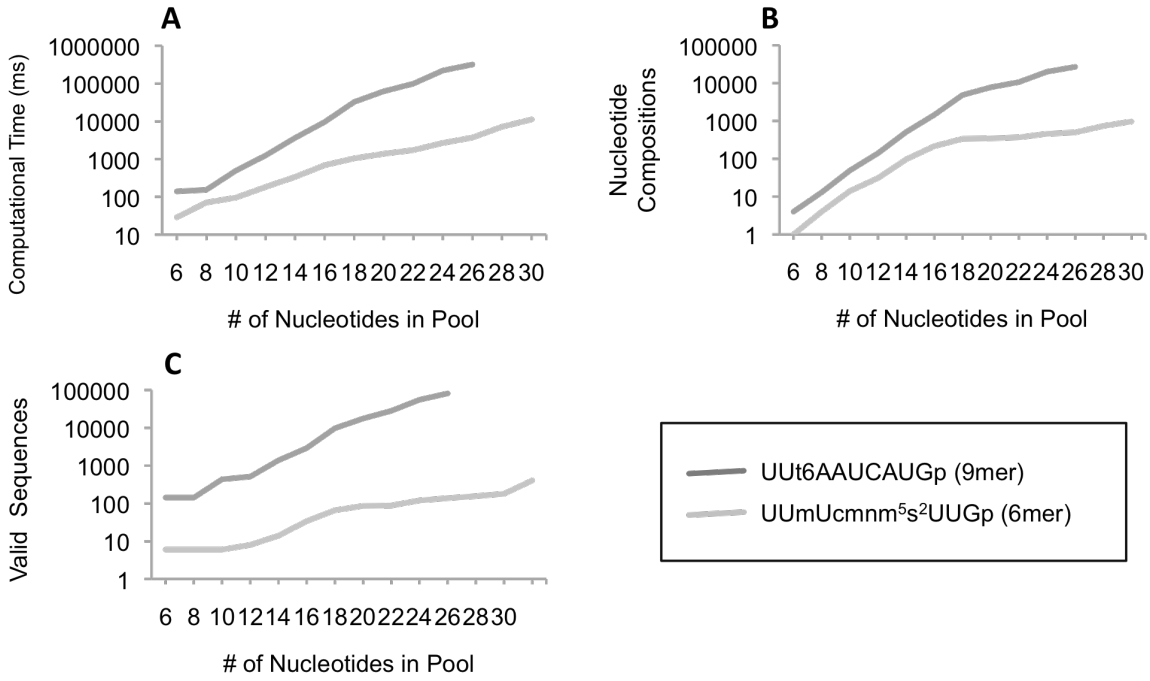
ABSTRACT

Ribosomal ribonucleic acid (RNA), transfer RNA, and other biological or synthetic RNA polymers can contain nucleotides that have been modified by the addition of chemical groups. Traditional Sanger sequencing methods cannot establish the chemical nature and sequence of these modified-nucleotide containing oligomers. Mass spectrometry (MS) has become the conventional approach for determining the nucleotide composition, modification status, and sequence of modified RNAs. Modified RNAs are analyzed by MS using collision-induced dissociation tandem mass spectrometry (CID MS/MS), which produces a complex data set of oligomeric fragments that must be interpreted to identify and place modified nucleosides within the RNA sequence. Here we report the development of RoboOligo, an interactive software program for the robust analysis of data generated by CID MS/MS of RNA oligomers. There are three main functions of RoboOligo: 1. Automated *de novo* sequencing via the local search paradigm. 2. Manual sequencing with real-time spectrum labeling and cumulative intensity scoring. 3. A hybrid approach, coined ‘variable sequencing’, which combines the user intuition of manual sequencing with the high-throughput sampling of automated *de novo* sequencing.

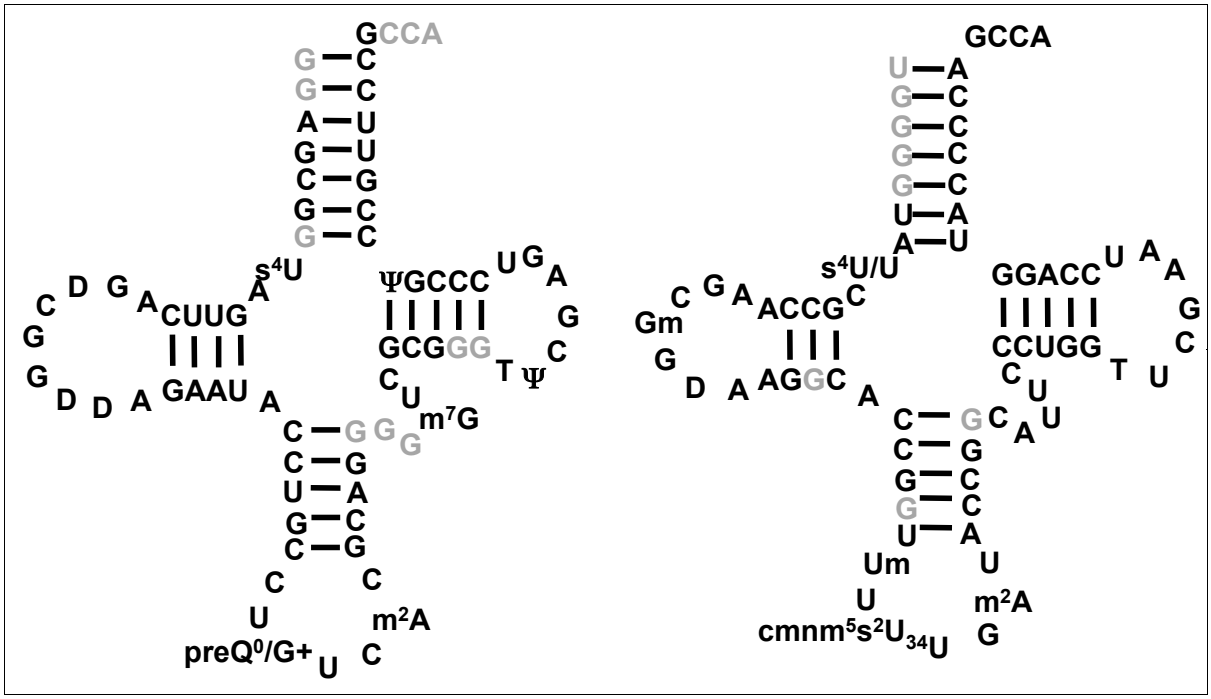
SUPPLEMENTARY DATA

Supplementary Figures S1-S5

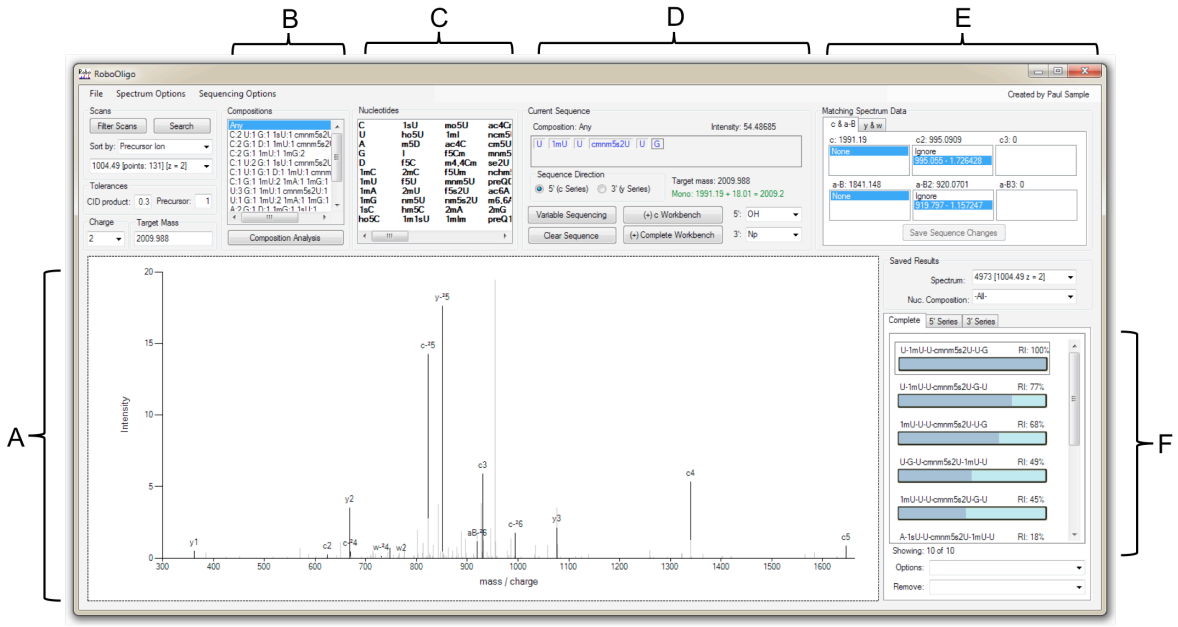
Supplementary Table S1



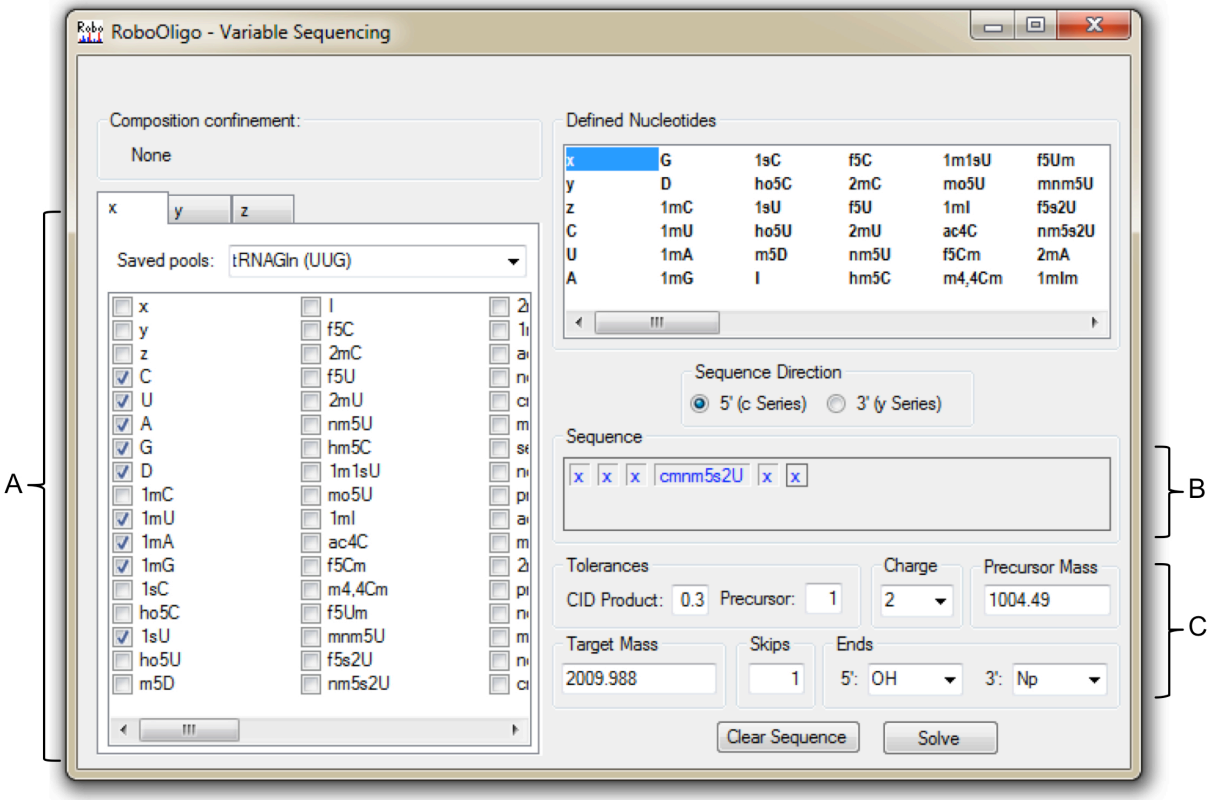
Supplementary Figure S1. Effects of increasing nucleotide pool size on automated *de novo* sequencing efficiency of the 6-mer U[Um]U[cmnm^{5}s^{2}U]UGp and the 9-mer UU[t^{6}A]AUCAUGp. **A** The computational time of each analysis. For the 9-mer, a pool size 26 nucleotides was the maximum condition before encountering issues with system memory usage. **B** The number of nucleotide compositions that fit within the target mass range and pass the restriction digest confinement filter. **C** The number sequences returned that are supported by the given data. The correct sequence of the 9-mer was maintained with all nucleotide pools tested while the 6-mer sequence was tied with UCCU[1mU][I]p after the addition of inosine (I) to the nucleotide pool.



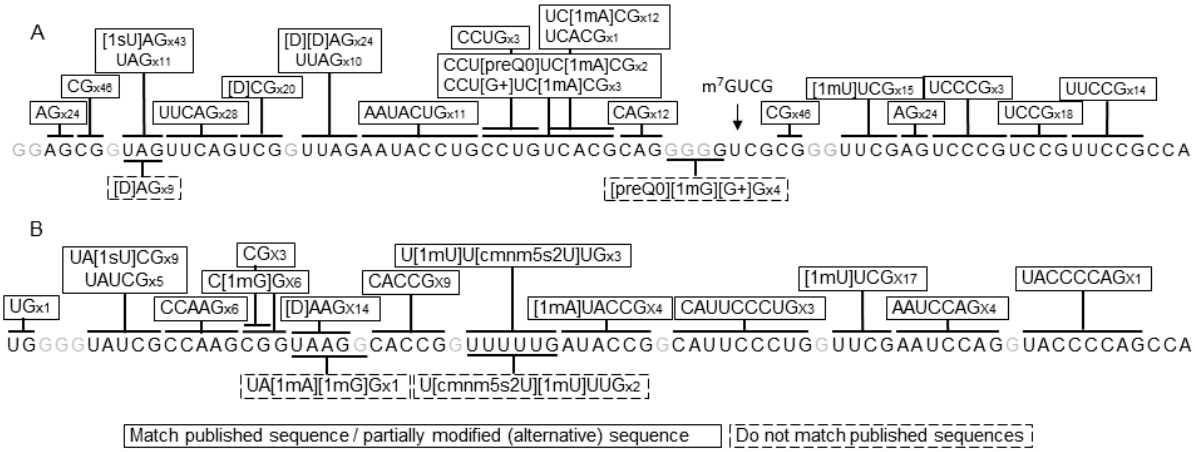
Supplementary Figure S2. Analysis of independently verified *E. coli* ΔqueC ΔqueF pGAT-queC $tRNA^{Asp}$ and *E coli* $tRNA^{Gln}$(UUG) using the automated *de novo* sequencing function of RoboOligo. Black letters indicate verified sequences that were corroborated as top scorers by RoboOligo automate**d de novo sequencing. A Sequence coverage of *E. coli* ΔqueC ΔqueF pGAT-queC $tRNA^{Asp}$. Nucleotide** Pool: C, U, A, G, 1sU, 1mG, 1mA, 1mU, D, $cmnm^5s^2U$ **B** Sequence coverage of *E coli* $tRNA^{Gln}$(UUG). Nucleotide Pool: C, U, A, G, 1mA, 1mG, 1sU, $preQ_0$, D, G+



Supplementary Figure S3.The RoboOligo primary interface. A. The MS/MS data analyzed in this figure was generated from a mutant *E.* coli tRNA^{Gln}(UUG) and the returned sequences are the product of RoboOligo's 'variable sequencing' function when given the seed sequence of 'x-x-x-cmnm^{5}s^{2}U-x-x'. The top score, U[1mU]U[cmnm^{5}s^{2}U]UGp is the correct sequence. **B.** Nucleotide composition analysis results that fall within the target mass range (2009.988 -/+ 1) and that obey the RNase T1 'strict' digestion confinement, which limits sequences to one G at the 3' end. Nucleotides used in the analysis are those found in WT *E. coli* tRNA^{Gln}(UUG): C, U, A, G, D, 1mG, 1mU, 1mA, 1sU, and cmnm^{5}s^{2}U. **C.** The nucleotide pool contains 108 unique masses of normal and modified nucleotides. Clicking on a nucleotide will add it to the 'Current Sequence' and attempt to find the corresponding CID products. **D.** The 'Current Sequence' contains the cumulative abundance of all detected CID products, sequence orientation selection, 5' and 3' end selections, and buttons to add sequences to the appropriate workbench. **E.** Clicking on a nucleotide in the 'Current Sequence' will display the *m/z* data points that fall within the range of that nucleotide's theoretical CID products (-/+ tolerance). **F.** The three workbenches store sequences that result from automated *de novo* sequencing, variable sequencing, and manual sequencing; and are sorted from high to low cumulative intensity. The relative intensity (RI) is the ratio of a sequence's cumulative intensity compared to the sequence with the highest cumulative intensity in the workbench.



Supplementary Figure S4. Variable sequencing of an MS/MS spectrum with a precursor *m/z* of 1004.49. In this scenario, the fourth nucleotide (cmnm^{5}s^{2}U) is defined, while the three preceding and two following user-defined 'x' variable nucleotides will be determined using a local search method that is similar to the automated de novo sequencing algorithm. **A.** The variable nucleotides x, y, and z can be individually assigned different nucleotide pools. **B.** The 'seed' sequence can contain any number and combination of defined and variable nucleotides. **C.** User-defined CID product *m/z* tolerance, precursor *m/z* tolerance, skips, 3' and 5' ends.



Supplementary Figure S5. Alignment of the results generated by the automated *de novo* sequencing batch analysis to the *E. coli* and **(A)** Asp-tRNA (GUC) and **(B)** Gln-tRNA (UUG) primary sequences. Solid boxes on top of the primary sequences match the published, fully-modified sequences. Overlapping regions represent alternative sequences that result from partially-modified tRNAs. Dashed boxes below the primary sequence are results that do not match the published sequences.The number of spectra with an identical top ranked sequence are indicated next to each sequence. For example, UA[1sU]CG was ranked as the highest scoring sequence for 9 spectra when all spectra of Gln-tRNA (UUG) were analyzed.

Supplementary Table S1. Composite results of all automated sequencing attempts using independently sequenced oligomer data listed in order of oligomer length in nucleotides (nts). The number of tested indicates how many oligomers of that length, but from different data sets, were analyzed. The '**Correct**' column refers to the number of tested data sets in which RoboOligo automated *de novo* sequencing returned the correct sequence as the highest scoring (in terms of cumulative abundance). '**Avg Complete Sequences**' - the average number of valid sequences that were returned for all tested oligomers of the same length. '**Avg Nucleotide Pool**' – the average number of nucleotides within the nucleotide pools used during automated sequencing. '**Avg Compositions**' – the average number of compositions that fit the calculated total mass range with a given nucleotide pool. '**Avg Time**' – the average time for the automated sequencing function to determine compositions, construct sequences, score, and display sequence results in milliseconds.

Length (nts)	Tested	Correct	% Correct:	Avg Complete Sequences	Avg Nucleotide Pool	Avg Compositions:	Avg time (ms):
2	8	8	100%	1.5	8.9	1.5	38.8
3	14	12	86%	4.4	9.3	2.1	74.9
4	20	19	95%	6.2	9.0	2.4	76.5
5	11	11	100%	22.8	9.5	3.3	95
6	2	2	100%	66.0	10.0	18.0	173.0
7	6	6	100%	1038.3	13.8	97.2	1309.5
8	4	4	100%	618.0	12.5	86.5	1193.8
9	7	7	100%	6216.6	10.0	273.6	6358.4
10	2	2	100%	2325.5	8.0	18.0	1543.5
11	2	1	50%	7221.0	8.0	11.5	6219.5
12	1	1	100%	65030.0	8.0	91.0	99222.0
All	**77**	**73**	**94%**				