Bidirectional Direct Sequencing of Noncanonical RNA by Two-Dimensional Analysis of Mass Chromatograms

Anders Björkbom*, Victor S. Lelyveld*, Shenglong Zhang*, Weicheng Zhang, Chun Pong Tam,
J. Craig Blain, and Jack W. Szostak

SUPPORTING INFORMATION

Table of Contents **Page**

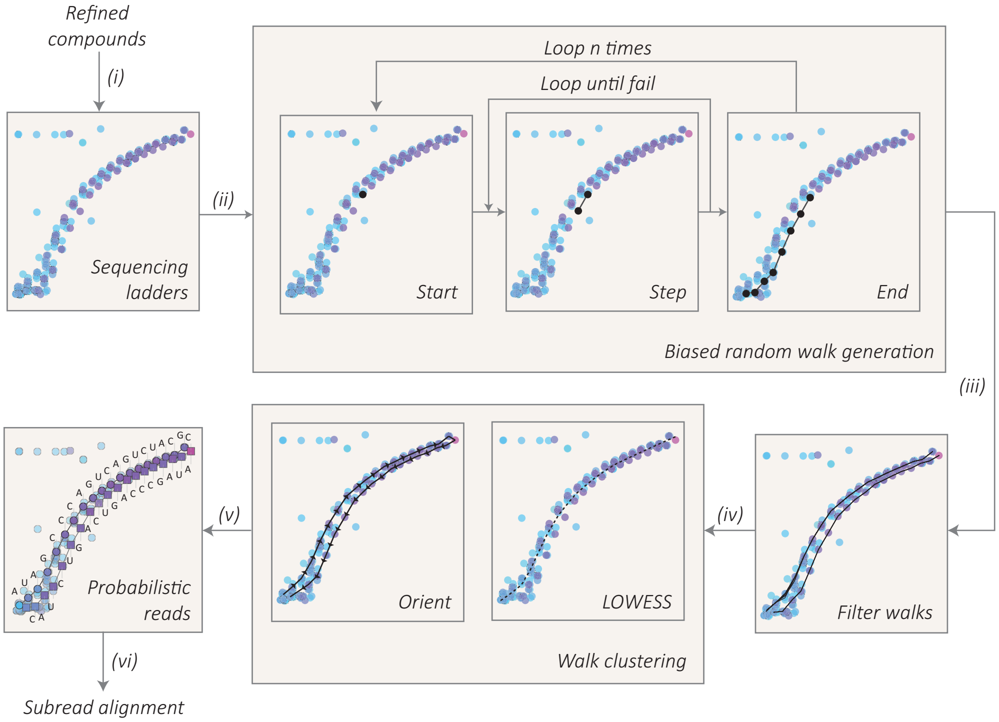
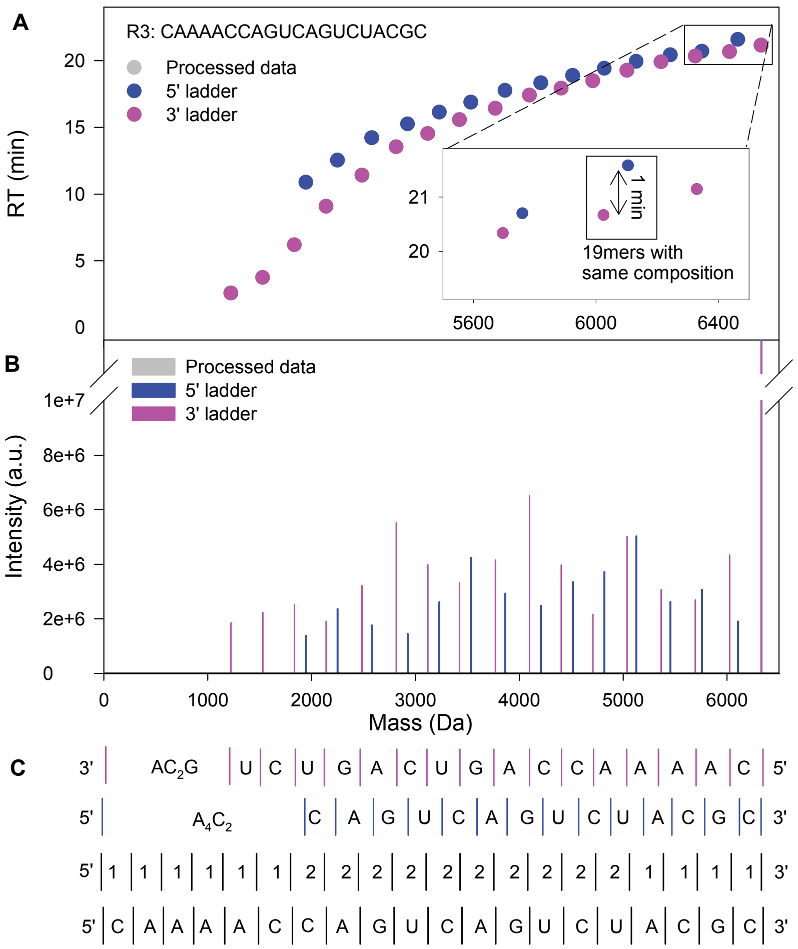


Figure S1. An algorithm for two-dimensional analysis of RNA sequencing ladders. An example set of points defined by mass (horizontal axis), retention time (vertical axis), and integrated intensity (color) are plotted for identified compounds. The algorithm proceeds as shown. (i) A sequencing ladder is refined by clustering related compound adducts and filtering on compound attributes. (ii) A biased rule-based random walk is performed to generate sequencing trajectories, and this walk is repeated *n* times along the two-dimensional space of compound points, where *n* should generally be an order-of-magnitude higher than the number of identified compounds. A compound point is chosen to start each walk. A step to a nearby point is then attempted repeatedly using a set of rules with a local intensity-weighted stochastic criterion (see Methods). A walk ends when no legal step can be taken. The resulting walk is a sequencing trajectory containing a candidate sub-read. (iii) The set of generated walks are filtered based on length. (iv) Related sequencing trajectories in the same orientation are clustered by bifurcating the dataset across a midline. The midline is generated by local regression using a LOWESS algorithm. Orientation is determined by observing whether the predicted full length compound is automatically contained within a cluster of trajectories. Given the sample preparation technique presented here, the full length compound is the one with maximum intensity. (v) Sequencing trajectories are converted to sub-reads and scored based on the compound intensity encountered along a trajectory. (vi) An alignment of sub-reads within each cluster is performed, and the consensus base call at each position is chosen by scoring. A bidirectional pair of reads is therefore generated and can be aligned with one another to generate a final consensus read.



3' | AC_2G | U | C | U | G | A | C | U | G | A | C | C | A | A | A | A | C | 5'

5' | A_4C_2 | C | A | G | U | C | A | G | U | C | U | A | C | G | C | 3'

5' | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 3'

5' | C | A | A | A | A | C | C | A | G | U | C | A | G | U | C | U | A | C | G | C | 3'

Figure S2. Effect of a 3′ phosphate group on the separation of sequence ladders. (a) The 19-nt fragments in the 3′ and 5′ ladders of RNA R3 (CAAAACCAGUCAGUCUACGC) have identical base composition due to loss of cytidine from either end. The chromatographic separation evident from the additional 3′-phosphate in the 5′ ladder leading to increased ion pairing retention is shown (inset). (b) Mass information for both sequence ladders. (c) Base calls based on mass differences for both sequence ladders. Immediately below the 3′ and 5′ ladders the confidence for each nucleotide position is shown and the bottom row displays the consensus sequence from aligning the 3′ and 5′ ladders. Manual data analysis was used for sequence determination.

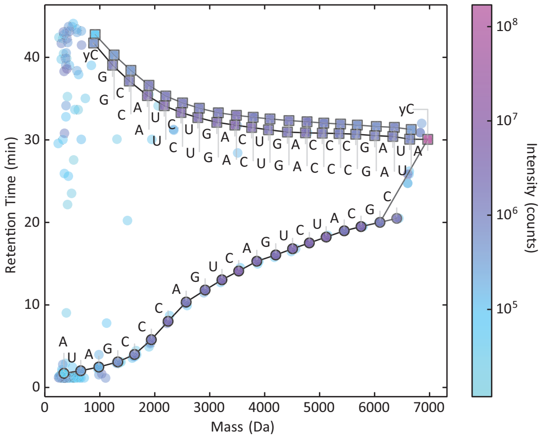


Figure S3. Formylation adduct on unhindered primary alcohol. Three major sequence trajectories were algorithmically generated from formic acid digestion of R4, which has a 3′ Cy3 label (**Table S1**). The uppermost 3′ ladder eluted 1 - 1.5 min after the authentic 3′ ladder and showed a +27.995 Da adduct, consistent with addition of CO by alcohol formylation. This neutral adduct was automatically clustered with parent compounds for the analysis of this dataset shown in Figure 3d. Formylation was particularly prevalent for the 3′ terminal propanol group in R4.

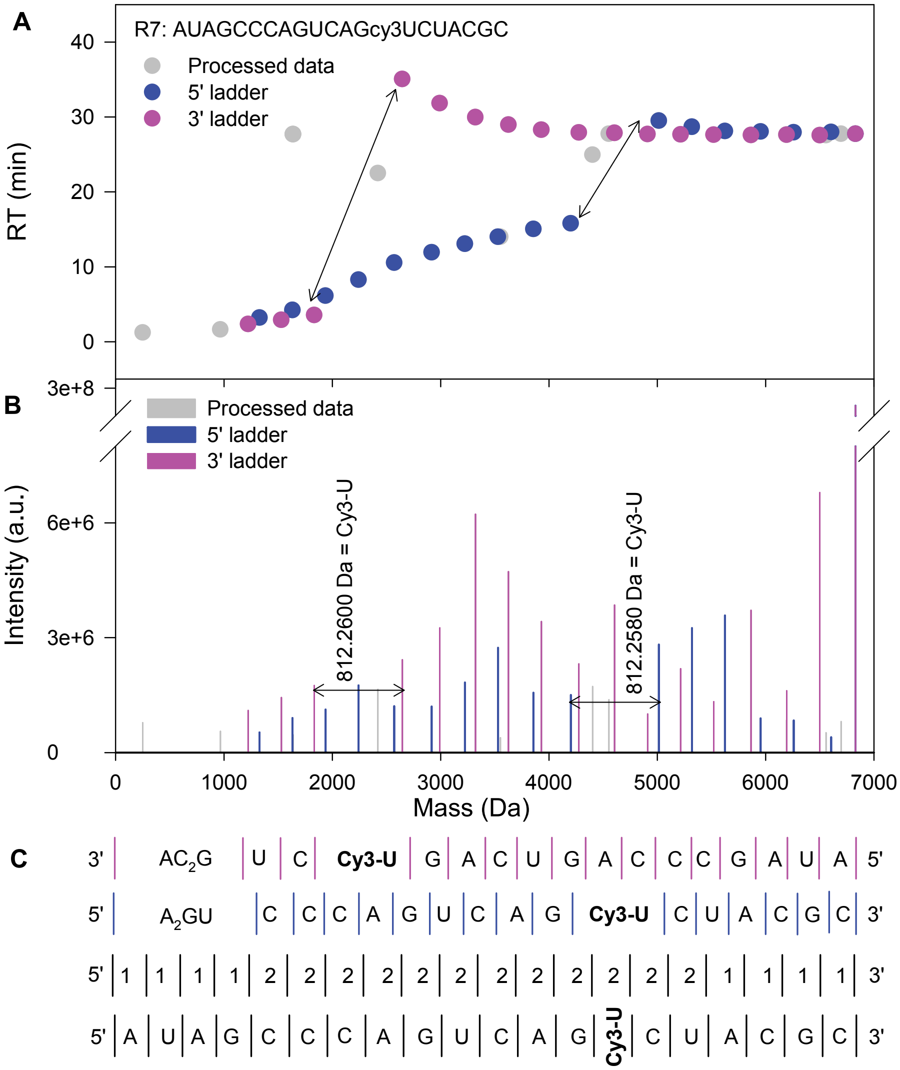


Figure S4. Effects of internal hydrophobic modifications on sequence ladder separation. (a) Chromatographic separation of sequence ladders for the RNA R7 (AUAGCCCAGUCAG-Cy3-UCUACGC). (b) Mass data for both sequence ladders. (c) Base calls based on mass differences for both sequence ladders. Immediately below the 3′ and 5′ ladders, the confidence for each nucleotide position is shown, and the bottom row displays the consensus sequence from alignment of the 3′ and 5′ ladders. Manual data analysis was used for sequence determination.

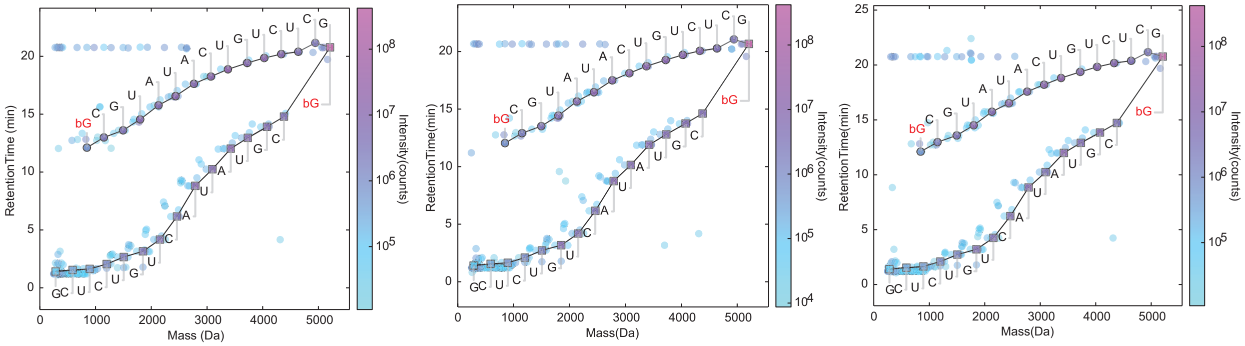


Figure S5. Repeatability of LC-MS sequencing of RNA. Three panels show equivalent sequencing results from three independent replicates of the sequencing workflow applied to a synthetic 5′-biotinylated 15-nt RNA (R6: Biotin-GCGUAUACUGUCUCG).

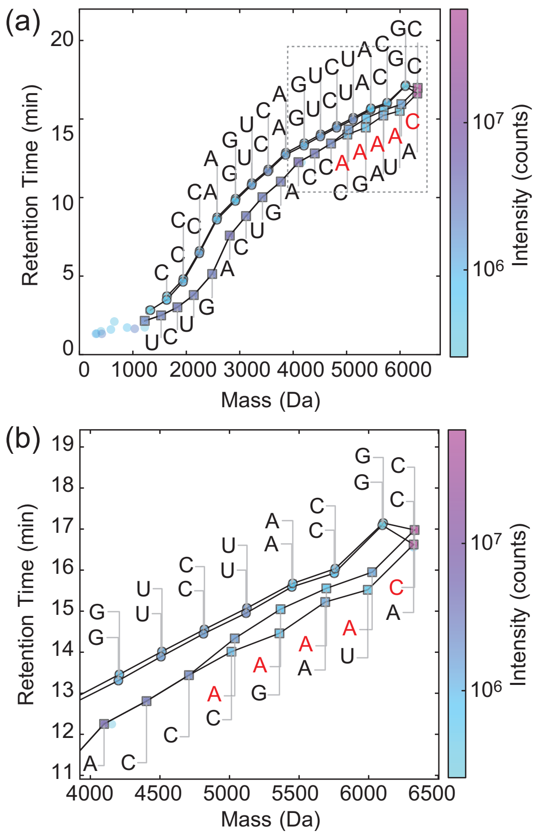
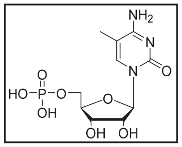


Figure S6. Effect of RNA sequence heterogeneity. (a) Sequencing of a mixture of two 20-nt RNAs varying in the first five bases (R1 and R3: CAAAACCAGUCAGUCUACGC). (b) Enlargement of part of sequence ladders in panel a to highlight separation based on sequence composition. The first five bases in the 3′ ladders were separated by ~30 s and at position C6 the ladders merged, since their sequences were identical. The 5′ ladders for R1 and R3 are unique over the whole sequence leading to a ~10 s separation of the 5′ ladders.

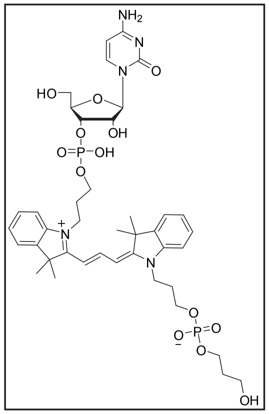
Table S1 Sequence information for all RNAs analyzed and modified base structures.

Name	Sequence 5'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	3'
R1	HO-	A	U	A	G	C	C	C	A	G	U	C	A	G	U	C	U	A	C	G	C	-OH
R2	HO-	A	U	A	G	C	C	C	A	G	U	m^{5}C	A	G	U	m^{5}C	U	A	C	G	C	-OH
R3	HO-	C	A	A	A	A	C	C	A	G	U	C	A	G	U	C	U	A	C	G	C	-OH
R4	HO-	A	U	A	G	C	C	C	A	G	U	C	A	G	U	C	U	A	C	G	C	-Cy3
R5	HO-	A	U	A	G	C	C	C	A	G	U	Cm	A	G	U	C	U	A	C	G	C	-OH
R6	Biotin-	G	C	G	U	A	U	A	C	U	G	U	C	U	C	G	-OH					
R7	HO-	A	U	A	G	C	C	C	A	G	U	C	A	G	Cy3-U	C	U	A	C	G	C	-OH
R8	HO-	A	s^{2}U	C	s^{2}U	C	C	C	A	G	U	C	A	G	U	C	U	A	C	G	C	-OH

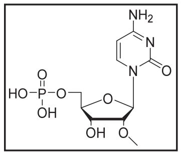
m^{5}C



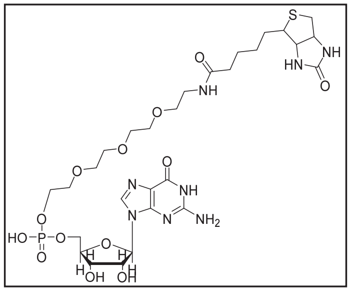
C-Cy3



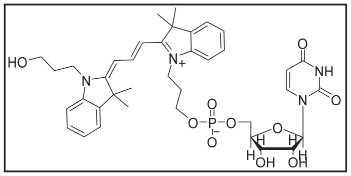
Cm



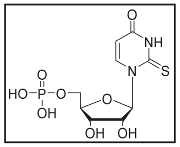
Biotin-G



Cy3-U



s^{2}U



SUPPLEMENTARY MATERIALS AND METHODS

Materials. All chemicals and LC-MS-grade solvents were purchased from Sigma-Aldrich and Fisher Scientific unless otherwise specified. All RNA was HPLC purified and purchased from Integrated DNA Technologies or ChemGenes or synthesized in-house. In-house synthesized RNA was further purified by triethylammonium bicarbonate ion pairing (IP) reverse phase (RP) chromatography using an Agilent 1100 HPLC and a 250 × 4.6 mm Zorbax Eclipse Plus C18 (5 μm particle size) column (Agilent Technologies) at 25 °C.

Sample preparation and acidic hydrolysis of RNA. A 0.2 mL PCR tube was charged with a 5 - 10 μL sample of RNA (50 - 100 μM) in water and an equal volume of room-temperature concentrated LC-MS Ultra-grade formic acid (Sigma-Aldrich/Fluka). The vortexed sample was then digested in a Bio-Rad T100 Thermo Cycler at 40 °C for 5 min unless otherwise indicated. The reaction mixture was immediately frozen on dry ice followed by lyophilization to dryness, which was typically completed within 1 h. The dried samples were immediately suspended in 20 - 40 μL LC-MS grade water for subsequent LC-MS analysis or stored at −30 °C.

Compound identification by molecular feature extraction. To extract relevant spectral and chromatographic information from the LC-MS experiments, we used the Molecular Feature Extraction workflow in MassHunter Qualitative Analysis (Agilent Technologies). This proprietary molecular feature extractor algorithm performs untargeted feature finding in the mass and retention time dimensions. In principal any software capable of compound identification could be used. The software settings were varied depending on the amount of RNA used in the experiment. In general, we wanted to include as many identified compounds as possible, up to a maximum of 1000. For samples with low concentrations, profile spectral peaks were filtered using a signal-to-noise ratio (SNR) threshold of 5 and, for more concentrated samples, an SNR threshold of up to 20. The other algorithm settings were as follows: “Small Molecules (chromatographic)” extraction algorithm, charge states from −1 to −15, only loss of hydrogen (−H) ions, “Common Organic Molecules” isotope model, minimum quality score 70 (range 0 - 100), and minimum ion count 500.

Manual sequence analysis and confirmation. Data analysis was performed manually to confirm the algorithmic results, as well as in cases where starting material contained more than one RNA (**Figure S6**) and for samples in which the sequence ladders cross each other (**Figure S4**). As with the automated analysis, the manual analysis was based on filtering of the compounds extracted by the Molecular Feature Extraction protocol. Here, we manually assigned compounds on the sequence ladders. Also, mass adducts were not clustered and were instead discarded. The compounds were filtered according to the following protocol:

1) Full length assignment: The full length RNA was the compound with highest intensity in our sample preparation. Compounds with a higher mass than the most intense mass were neglected.
2) Window by mass/RT ratio: By specifying a low and high boundary for the neutral mass/RT ratio, compounds that lie far outside the sequence ladders in both mass and RT space were neglected.
3) Quality score: A quality score was assigned to each compound by the Molecular Feature Extraction algorithm. A high threshold was used to eliminate many undesired compounds (usually 85 on a range from 0 - 100)
4) Mass adducts: Mass adducts such as +Na−H, +K−H, −H_2O, +formate, +formyl, and combinations of these adducts were discarded if within a specific mass error and RT window (usually 3 × average peak width and 10 ppm).
5) Intensity cut-off: By setting a high intensity cut-off only the most intense peaks were used for sequencing.

We manually performed peak assignment by calling the mass difference between visibly adjacent sequence fragments against a database of modified nucleotide masses within a set mass error. To facilitate assignment, compounds were filtered with highly stringent criteria such that the resulting set of compounds was of a tractable size for manual analysis — on the order of the number of expected fragments. (Because of less restrictive filtering, the automated analysis algorithm was able to identify some ladder fragments not considered during the manual analysis.) Once a complete sequence was predicted, the mass error between the measured sequence fragments and the predicted sequence fragments was calculated. In some cases, manual assignment was possible only after the settings used to filter the initial compound list were varied for different mass ranges. This windowed analysis approach made it possible to identify more compounds on the sequence ladders as compared to using a static threshold.