analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 8500−8506

pubs.acs.org/ac

Oligonucleotide Sequence Mapping of Large Therapeutic mRNAs via Parallel Ribonuclease Digestions and LC-MS/MS

Tao Jiang,[†] Ningxi Yu,[‡] Jaeah Kim,[§] John-Ross Murgo,[†] Mildred Kissai,[†,⊥] Kanchana Ravichandran,[†] Edward J Miracco,[†] Vladimir Presnyak,[†] and Serenus Hua*[,†]

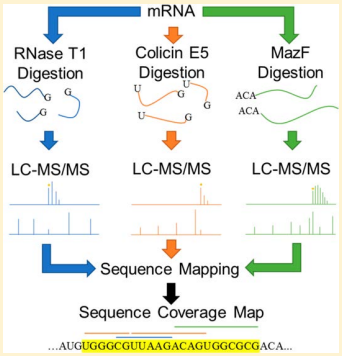
[†]Moderna Inc., 500 Technology Square, Cambridge, Massachusetts 02139, United States

[‡]Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States

[§]Department of Pharmaceutical and Biomedical Sciences, College of Pharmacy, University of Georgia, Athens, Georgia 30602, United States

● *Supporting Information*

ABSTRACT: Characterization of mRNA sequences is a critical aspect of mRNA drug development and regulatory filing. Herein, we developed a novel bottom-up oligonucleotide sequence mapping workflow combining multiple endonucleases that cleave mRNA at different frequencies. RNase T1, colicin E5, and mazF were applied in parallel to provide complementary sequence coverage for large mRNAs. Combined use of multiple endonucleases resulted in significantly improved sequence coverage: greater than 70% sequence coverage was achieved on mRNAs near 3000 nucleotides long. Oligonucleotide mapping simulations with large human RNA databases demonstrate that the proposed workflow can positively identify a single correct sequence from hundreds of similarly sized sequences. In addition, the workflow is sensitive and specific enough to detect minor sequence impurities such as single nucleotide polymorphisms (SNPs) with a sensitivity of less than 1%. LC-MS/MS-based oligonucleotide sequence mapping can serve as an orthogonal sequence characterization method to techniques such as Sanger sequencing or next-generation sequencing (NGS), providing high-throughput sequence identification and sensitive impurity detection.



As a novel drug modality, mRNA has significant therapeutic potential in multiple disease areas. Following insertion into a cell, mRNA drugs use endogenous cellular machinery to express a preprogrammed protein.[1] Such expressed proteins can fulfill a myriad of purposes, from promoting a specific immune response[2,3] to modulating or restoring a variety of metabolic processes.[4,5] Currently, clinical trials are ongoing or planned for a large variety of mRNA drugs designed to treat or prevent various cancers, cardiovascular diseases, and infectious diseases.

As with most biotherapeutics, sequence is a critical quality attribute (CQA) for mRNA drugs. Historically, sequencing methods such as Sanger sequencing and next-generation sequencing (NGS) have been used to determine the sequence identity and purity of long RNAs. In particular, Sanger sequencing has been applied to obtain sequence information from DNA and RNA for over 40 years,[6] and remains a key technology in the biopharmaceutical industry's toolkit due to its reliability, cost-effectiveness, and rapid turnaround time for sequence identification purposes.[7] NGS, on the other hand, is a newer technology in which nucleic acid chains are subjected to massively parallel sequencing, leading to large increases in throughput as well as sensitivity.[8]

In contrast, mass spectrometry, though infrequently used for nucleic acid sequencing, has served as a cornerstone analytical technology for biopharmaceutical protein characterization for several decades.[9] Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) can now identify thousands of proteins from biological samples,[10] or detect protein sequence variants that differ by a single amino acid residue at less than 1% relative abundance.[11] Sequence coverage is one of the key performance indicators of a bottom-up analysis, representing how much of a protein sequence is covered by detected peptides. A higher sequence coverage indicates more complete characterization and thus more confident identification. In sequence variant detection, only mutations at covered amino acid residues can be detected. The use of multiple enzymes of different specificities has been reported to enhance sequence coverage of proteomics analysis.[12]

For characterization of biologics, orthogonal methods are desired for a panoramic view. Corroborating information can be obtained on each characteristic, and analytical artifacts can be minimized. For analysis of nucleic acids, LC-MS is an orthogonal method to Sanger and NGS, providing unique advantages when compared to conventional sequencing technologies: direct analysis, improved sensitivity over Sanger sequencing, and faster turnaround times than NGS. Direct

Received: April 4, 2019
Accepted: May 26, 2019
Published: May 26, 2019

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

detection of oligonucleotides eliminates the need for polymerase processing steps such as reverse transcription and polymerase amplification, increasing fidelity and speed of analysis. As this study will show, LC-MS sensitivity is sufficient to directly detect even low-level sequence polymorphisms (<1% abundance) with minimal sample manipulation. Modified nucleotides, which can interfere with polymerase activity or fidelity, may also be directly detected, identified, and localized to specific modification sites.[13−15]

To design a high-coverage, high-throughput oligonucleotide mapping method for long mRNAs, we designed a multi-endonuclease parallel digestion protocol that generates multiple sets of oligonucleotides. Use of multiple orthogonal enzymes helps achieve near-total sequence coverage in peptide mapping.[16,17] This strategy has also been applied to mapping of shorter RNAs, such as tRNA.[15,18,19] However, mRNAs are typically much longer than tRNAs, while also containing few or no base modifications. As a result, high-frequency cleavers such as RNase T1[20] generate far more isomeric or even identical oligonucleotides from mRNAs than from tRNA. To improve sequence coverage of long mRNA, endonucleases with a variety of cleavage frequencies are desired. RNase T1,[21] colicin E5,[22,23] and mazF[24] each have distinct RNA digestion specificities: RNase T1 cleaves at the 3′ end of G, colicin E5 cleaves between GU, and mazF cleaves at the 5′ end of ACA. The diversified specificities of these endonucleases make their sequence coverages complementary, maximizing overall sequence coverage of long mRNA.

EXPERIMENTAL SECTION

RNA Preparation. Human erythropoietin (Epo) mRNA, firefly luciferase (Luc) mRNA, and α-catenin mRNA were prepared by T7 RNA polymerase in vitro transcription using a DNA template containing the open reading frame flanked by the 5′ and 3′ untranslated regions (UTR) and a poly-A tail. Epo and Luc mRNAs with designed nucleotide point mutations (Table S1) were also prepared by the same process.

All oligonucleotides were prepared synthetically by Integrated DNA Technologies (Coralville, IA).

mRNA Digestion. For RNase T1 digestion, 40 μL mRNA or mRNA mixtures at 1 mg/mL concentration were mixed with 60 μL 8 M Urea (Sigma-Aldrich, St. Louis, MO), 12 μL 1 M pH = 7 Tris-HCl buffer (Invitrogen, Carlsbad, CA), and 0.8 μL 0.5 M EDTA (Invitrogen, Thermo Fisher). The samples were then denatured at 90 °C for 10 min. Samples were cooled to room temperature after denaturation. Twenty microliters of 1000 U/μL RNase T1 (Thermo Fisher Scientific, Waltham, MA) was added, followed by incubation at 37 °C for 15 min.

For colicin E5 digestion, 40 μL of mRNA or mRNA mixtures at 1 mg/mL concentration were mixed with 60 μL 8 M Urea (Sigma-Aldrich), 12 μL of 1 M pH = 8 Tris-HCl buffer (Invitrogen), and 0.8 μL of 0.5 M EDTA (Invitrogen). The samples were then denatured at 90 °C for 10 min. Samples were cooled to room temperature after denaturation. Ten microliters of 1.4 mg/mL colicin E5 was added, followed by incubation at 37 °C for 30 min. Detailed expression[25] and purification[26] methods for colicin E5 are described in Supporting Information. A plasmid map of colicin E5 is shown in Figure S1.

For mazF digestion, 20 μL of mRNA at 1.0 mg/mL, 20 μL of mazF at 20 U/ μL (Takara Bio, Kusatsu, Japan), 40 μL of 5× mazF buffer (200 mM sodium phosphate at pH 7.5 with 0.05% polysorbate 20), and 20 μL of LC-MS grade water were mixed and then incubated at 37 °C for 15 min.

Chromatographic Separation and MS Analysis. Digested mRNAs are chromatographically separated by reversed-phase ion pairing liquid chromatography (RPIP-LC) using an Agilent 1290 UPLC and then analyzed by an Agilent 6550 quadrupole time-of-flight mass spectrometer (Q-TOF). Mobile phase A is 1% 1,1,1,3,3,3-hexafluoro-2-propanol (Sigma-Aldrich) and 0.1% *N*,*N*-diisopropylethylamine (Sigma-Aldrich) in LC-MS grade water (Thermo Fisher Scientific). Mobile phase B is 0.075% 1,1,1,3,3,3-hexafluoro-2-propanol and 0.0375% *N*,*N*-diisopropylethylamine in 65% LC-MS grade acetonitrile (Thermo Fisher Scientific) and 35% LC-MS grade water. The UPLC column for separation is an Acquity UPLC Oligonucleotide BEH C18 Column, 130 Å, 1.7 μm, 2.1 mm × 100 mm (Waters, Milford, MA). LC gradients used for mapping are provided in Supporting Information. The 6550 QTOF is set to gather one spectrum per second in MS-only runs, and one to two spectra per second in MS/MS runs in extended dynamic range mode (2 GHz). MS/MS parameters are further described in Supporting Information.

Data Processing. All data processing steps are automated with an in-house C# program based on Agilent MassHunter Data Access Component. Details of the data processing algorithm are presented in Supporting Information. Output includes sequences, masses, mass errors, retention times, and abundances of oligonucleotide hits. Results can be reproduced and/or verified manually with Agilent MassHunter (version B.07).

RESULTS AND DISCUSSION

Evaluation of Endonuclease Activity. Specificity is crucial for ensuring the fidelity of oligonucleotide databases generated by in silico digestion. To confirm that denaturing conditions do not affect digestion specificity, oligonucleotides with designed cut sites were synthesized to test each endonuclease. Digestion specificity and efficiency were examined by LC-MS/MS analysis of product oligonucleotides. Injection amounts were set between 5 and 20 μg. Such excessive injection amounts caused LC peak shouldering but ensured the detection of low-level minor digestion products. LC-UV profiles for each digested oligonucleotide are shown in Figure S2. For RNase T1, cleavages 3′ to G are the only type of cleavage observed. For colicin E5, in addition to the major cleavage at GU sites observed in previous literature,[22] minor cleavage activity is observed at CU sites at <0.5% signal abundance by UV relative to total product oligonucleotides. For mazF, in addition to the major cleavage at ACA sites, minor cleavage activities were also observed at AUA and ACU sites with <1% relative signal abundance by UV per cleavage site. We also observed that the buffer concentration used in mazF affects both endonuclease activity and specificity: lower buffer concentrations generally promoted endonuclease activity but also incurred a relatively higher rate of AUA and ACU cleavages.

Of the three endonucleases tested, RNase T1 demonstrated the best specificity and digestion efficiency. This endonuclease is also commercially available and inexpensive. In cases where only sequence identity confirmation is needed, LC-MS analysis of an RNase T1 digestion alone will typically suffice. However, due to the single nucleotide specificity of RNase T1, large numbers of repetitive oligonucleotides are generated during digestion, significantly reducing sequence coverages for large

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

mRNAs. For common protein drug modalities such as monoclonal antibodies, trypsin digestion followed by LC-MS/MS analysis commonly results in >80% sequence coverage, whereas for large mRNAs (>3000 nt) or for coformulated mixtures of mRNAs, RNase T1 digestion alone followed by LC-MS/MS analysis commonly results in sequence coverages below 30%. In such cases, the high LC-MS sequence coverages (>70%) necessary for thorough characterization are only available through the use of parallel endonuclease digestions.

Unlike RNase T1, colicin E5 and mazF both exhibited condition-dependent minor digestion products. Salt concentration and pH were both found to affect abundance of minor digestion products. At optimal digestion conditions, minor digestion products are less than 0.5% of total product for colicin E5 and less than 2% for mazF, which is comparable to some common proteases used for protein characterization, such as Asp-N.[27] Wild-type colicin E5 and mazF tend to generate large oligonucleotides, so minor digestion products rarely cause interference. However, in cases of low-level impurity analyses such as SNP detection, interference from minor digestion products must be ruled out. Given the importance of alternate cleavage specificities to mRNA sequence characterization, manipulation of endonucleases for better specificity and efficiency is worthwhile. Sequence engineering and directed mutation of these endonucleases may be important possibilities to consider in the quest for improved digestion efficiency and specificity. Such methods have been applied to engineering of restriction enzymes[28] and endonucleases.[29]

Even with highly specific and efficient endonucleases, endonuclease digestion specificity brings up an informatic dilemma in the digestion of a long strand mRNA: endonucleases with single nucleotide specificity result in frequent cuts, producing identical or isomeric oligonucleotides from multiple regions of the mRNA, while endonucleases that have dimer or trimer specificity may produce oligonucleotides that are too large for isotopic resolution and accurate mass determination. The solution demonstrated in this study is to use multiple endonucleases with distinct specificity and different recognition site lengths such that sequence coverages are complementary.

LC-MS Profile of mRNA Digests. LC-MS total ion chromatogram (TIC) profiles of Epo mRNA digested by RNase T1, colicin E5, and mazF are shown in Figure 1. The polyA tail oligonucleotide of the mRNA elutes around 22.6 min for all digests and serves as a rapid visual marker of reaction progress.

In general, longer oligonucleotides elute later than shorter ones. Colicin E5 and mazF are observed to produce longer oligonucleotides than RNase T1, as expected based on their respective specificities. Due to digest complexity, multiple oligonucleotides may coelute; therefore, the number of chromatographic peaks observed is less than the number of oligonucleotides. The LC profile is directly correlated to the mRNA sequence and therefore may serve as a fingerprint to identify an mRNA against a known standard. However, precise sequence identification, as well as low-level impurity detection, require additional MS analyses.

MS/MS Differentiation of Isomeric Oligonucleotides. As RNase T1 tends to produce shorter oligonucleotides than colicin E5 and mazF, coeluting isomeric oligonucleotides are often observed after digestion with RNase T1. In such cases, MS/MS is highly effective for identifying which isomer, or isomers, are present. Figure S3 shows two examples in which chromatographically unresolved isomers are isolated together and subjected to CAD. The resulting MS/MS spectra contain a multitude of high-abundant, isomer-specific fragment ions that confirm the copresence of all isomeric oligonucleotides predicted by in silico digestion.

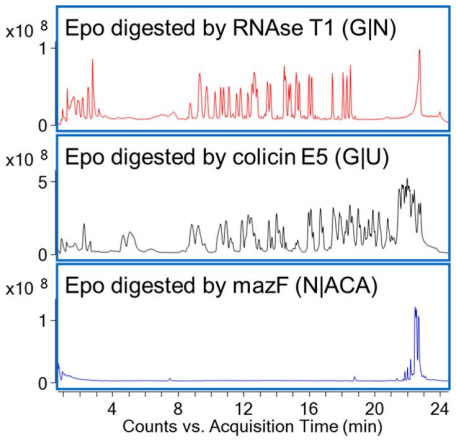


Figure 1. LC-MS total ion chromatograms of Epo mRNA digested by RNase T1 (top, red), colicin E5 (middle, black), and mazF (bottom, blue). Note that colicin E5 and mazF tend to produce larger, later-eluting oligonucleotides.

As oligonucleotides increase in length, the complexity of their MS/MS spectra increases drastically due to fragment ions of various charge states. Happily, longer oligonucleotides are also statistically less likely to have isomeric matches within the same sequence, as seen in Table S2. In most cases, accurate mass MS is sufficient to identify the majority of unique oligonucleotides within an endonuclease digest.

Database Search and Sequence Mapping. Oligonucleotides identified by LC-MS/MS may be mapped out in the context of the entire mRNA sequence. Examples are shown in Figure 2, which visualizes the sequence coverages obtained from individual digestions of Epo (745 nt), Luc (1816 nt), and α-catenin (2884 nt) mRNA by RNase T1, colicin E5, and mazF. Near-complete digestions can be obtained using each of the three endonucleases. For RNase T1, all predicted unique oligonucleotides under 12 kDa are observed. For both the colicin E5 and mazF digestions, over 90% of predicted unique oligonucleotides under 12 kDa are observed, as well as some oligonucleotides with missed cleavages. Predicted oligonucleotides in the colicin E5 and mazF digests are seen to differ by several-fold in intensity, suggesting some sensitivity to mRNA local structure for these endonucleases. The numbers of uncoverable nucleotides due to repetitive oligonucleotides from digestion of each enzyme are shown in Table S3, underscoring the disproportionately large effect of length on sequence coverage for frequent cleavers such as RNase T1.

For our current data analysis and simulation workflow, which is based on resolved-isotope deconvolution of mass spectra, mazF has the lowest sequence coverage among all endonucleases tested because the digestion products are often greater than 12 kDa and therefore isotopically unresolved by the Q-TOF mass spectrometers used in this study. However, with the use of higher-resolution mass spectrometers, a higher range of masses may be reached, greatly enhancing the level of data that might be extracted from a mazF digest.

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

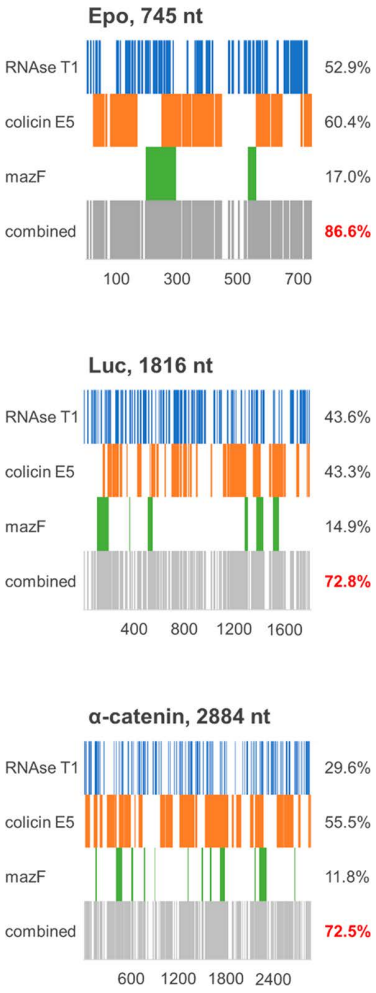


Figure 2. Sequence coverage maps obtained from individual digestions of Epo, Luc, and α-catenin mRNA by RNase T1, colicin E5, and mazF.

Sequence Identification by Relative Sequence Coverage. One major application of oligonucleotide sequence mapping is sequence identification. In Figure 3, experimental data sets generated from Epo mRNA are searched against in silico digests of multiple other mRNA sequences. Database searches against the correct Epo sequence result in significantly higher sequence coverages than searches against incorrect sequences of identical length, such as sequences with alternative codons or random sequence isomers. However, for sequences with disparate lengths, the likelihood of false discovery (e.g., random mass match with a totally different predicted oligonucleotide from a totally different mRNA sequence) increases with the length of the searched mRNA sequence(s). For example, Epo (745 nt) experimental data searched against a much longer, incorrect sequence (α-catenin, 2884 nt) results in an absolute sequence coverage of 699 nt, whereas the same data searched against the correct Epo sequence results in an absolute sequence coverage of only 645 nt. In contrast, the relative sequence coverage tells a more accurate story. Epo experimental data searched against α-catenin results in a relative sequence coverage of just 24.2%, whereas the same data searched against the correct Epo sequence results in a relative sequence coverage of 86.6%.

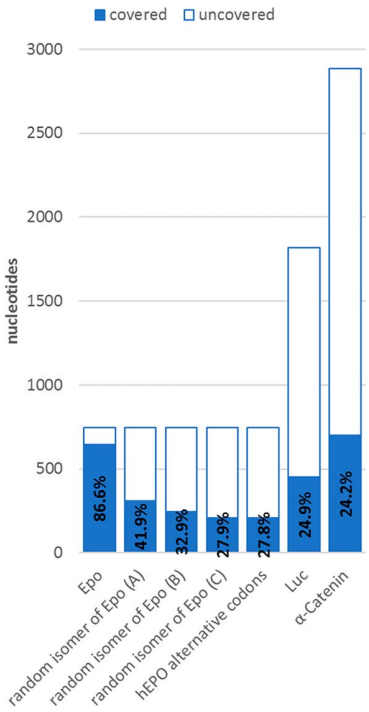


Figure 3. Absolute sequence coverages (in nucleotides) and relative sequence coverages (in percentage) calculated from experimental data on three parallel Epo digestions, searched against in silico digest databases for the true sequence of Epo as well as some other, incorrect mRNA sequences.

Sequence length is also crucial for accurate sequence identification. Consider, for example, a scenario where digests of a long mRNA are searched against both the true long mRNA sequence and a significantly shorter mRNA sequence. The increased likelihood of isomeric or repetitive oligonucleotides in the true, long mRNA sequence would depress both the absolute and relative sequence coverages of the true sequence. In contrast, an incorrect short sequence with similar digestion products (such as a truncate) would not experience the same likelihood of isomers/repetitions and thus might score a higher sequence coverage than the true sequence. To counter this effect, mRNA database matching should only be attempted against sequences of similar length. Approximate mRNA sequence length may be obtained through complementary analytical technologies such as capillary electrophoresis[30] or liquid chromatography.[31]

The use of relative sequence coverage and sequence length for sequence identity is modeled on a grander scale in Figure 4. A library of 2000 mRNA coding regions was randomly selected from the human transcriptome. The sequences were binned into 10 length groups with 200 sequences in each group (600–3600 nucleotides, with 300-nucleotide steps) to simulate a practical use case in which similarly sized constructs that cannot be distinguished by LC or electrophoresis are investigated for sequence identity. In silico digestion was applied to each of the 2000 RNA sequences using RNase T1, colicin E5, and mazF, allowing zero missed cleavages. On the basis of method performance with real data sets, we assumed for simulation purposes that all oligonucleotides smaller than 12 kDa could be detected. All detectable digestion product oligonucleotides from each RNA were searched against all 200 sequences in the same length group. Predicted isomeric or

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

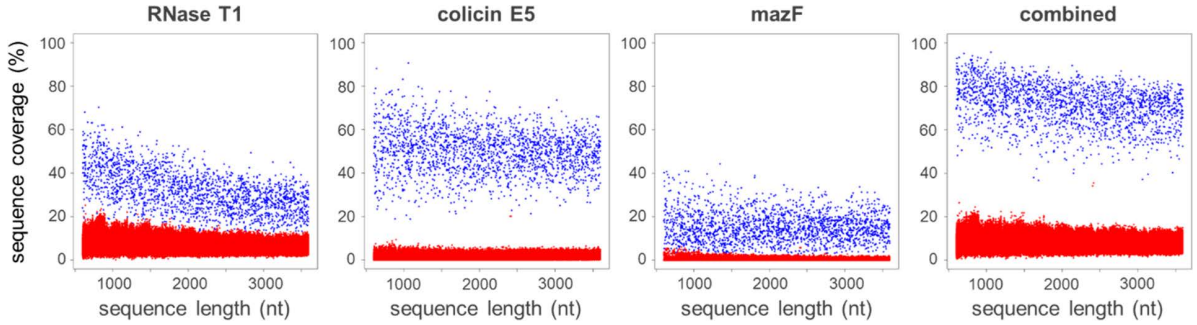


Figure 4. Predicted sequencing performance of four different mRNA digestion schemes for a randomly selected group of 2000 human mRNA coding regions. Each plot shows the sequence coverages obtained by searching every individual simulated oligonucleotide data set against either the correct sequence (blue, $n = 2000$) or an incorrect sequence of similar length (red, $n = 199 \times 2000$). In total, 1.6 million simulations were performed across all digestion schemes (RNase T1, colicin E5, mazF, or all enzymes in parallel) and all mRNA lengths (600 to 3600 nucleotides, 300-nucleotide bins).

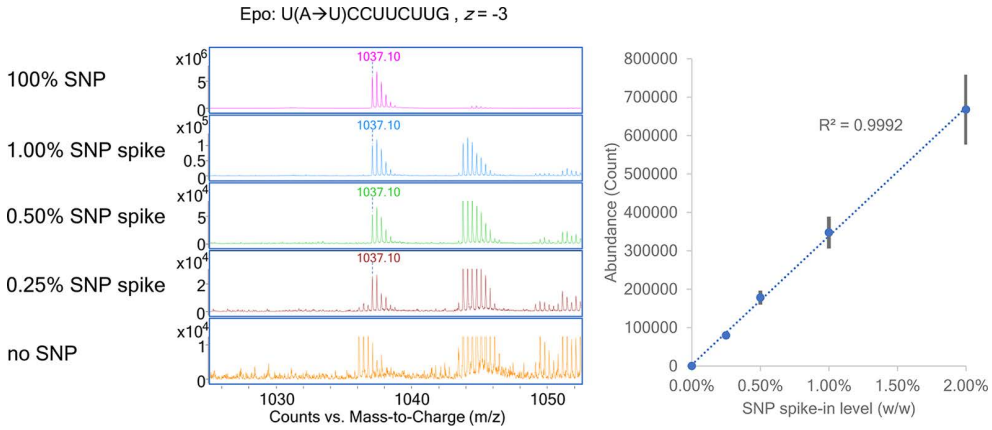


Figure 5. Detection and quantitation of low-level Epo SNPs by LC-MS. SNP mRNAs were mixed with "normal" Epo mRNA at various levels and then digested with RNase T1. The SNP oligonucleotide U(A→U)CCUUCUUG ($z = -3$, monoisotopic $m/z = 1037.10$) is clearly differentiated from other coeluting compounds by MS^1 alone (left). SNP spike percentage and resulting oligonucleotide ion abundance have high linear correlation. Error bars represent the 95% confidence interval of digestion triplicates.

isobaric oligonucleotides were all considered undetectable to simulate the worst-case scenario of no MS/MS isomeric differentiation (or no MS/MS data gathered). Sequence coverages generated from such simulated searching can be defined as the maximum sequence coverage possible when all predicted informative oligonucleotides are detected with LC-MS.

The plots in Figure 4 show the relative sequence coverage for each individual simulated data set when searched against the correct sequence (in blue) as well against the other 199 sequences in the same length group (in red). For every simulation, searching against the correct sequence always results in the highest sequence coverage. However, sequence coverage tends to be lower for longer RNAs due to the increased likelihood of repetitive and isomeric digestion products. When simulation results from different endonucleases are compared against each other, it is evident that the sequence coverage obtained from multiple endonucleases is higher than what can be obtained from any single endonuclease. In contrast to RNase T1, the sequence coverages provided by colicin E5 and mazF (both of which tend to generate long oligonucleotides) are less affected by RNA length yet also have higher variances from construct to construct. As seen also with experimental data, MazF sequence coverage is significantly lower than that of other enzymes, primarily because the majority of oligonucleotides produced by digestion with mazF are larger than 12 kDa and therefore considered undetectable in this simulation; this may be significantly addressed in future studies through the use of higher-resolution instruments. Plotting the sequence coverage percentage of each correct sequence against the maximum sequence coverage percentage of the incorrect sequences, receiver operating characteristic (ROC) curves can be calculated for each enzyme (Figure S4). For RNase T1, colicin E5, and mazF, the areas under the ROC curves are 0.9879, 1.0000, and 0.9904, respectively, suggesting that sequence coverage percentages may be used to identify the correct sequence from hundreds of possibilities with exceedingly high sensitivity and specificity.

Mapping of mRNA Mixtures. In some cases, concurrent expression of multiple mRNA sequences is desired; for example, where two proteins work synergistically to achieve an effect or when a single protein is made up of multiple subunits[32] In such cases, an mRNA drug product might consist of multiple mRNAs with distinct sequences mixed in a specific ratio. To evaluate oligonucleotide mapping of a multivalent mRNA drug, a cocktail of mRNA Epo, Luc, and α-catenin was analyzed. The concentration of each mRNA in the cocktail was 0.33 mg/mL. Data were searched against all three mRNA sequences combined. When the cocktail mapping result

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

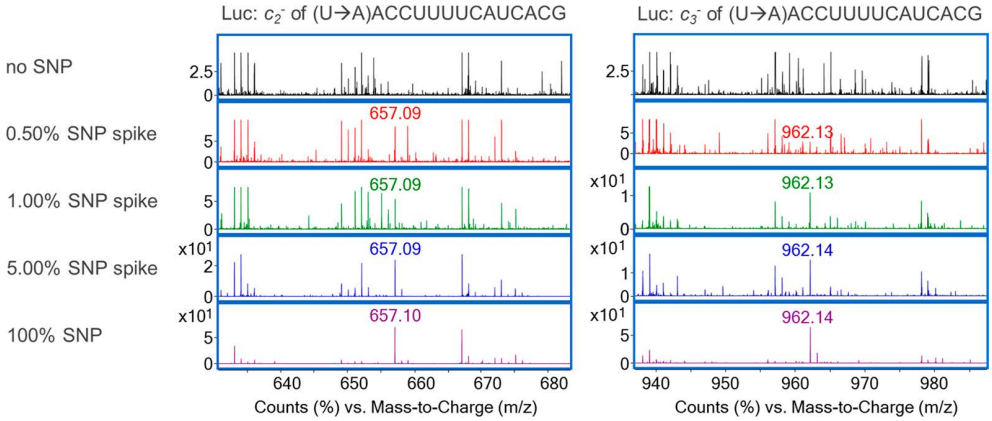


Figure 6. Detection and quantitation of low-level Luc SNPs by LC-MS/MS. SNP mRNAs were mixed with "normal" Luc mRNA at various levels and then digested with RNase T1. The c_2 (left) and c_3 (right) fragments of SNP oligonucleotide (U→A)ACCUUUUCAUCACG ($z = -3$, monoisotopic $m/z = 1577.20$) are clearly identified despite the presence of persistent background signals arising from isobaric, coisolated precursor ions.

(Figure S5) is compared against the mapping results of each individual mRNA, a significant drop in sequence coverage is observed. This sequence coverage loss is due to the presence of identical digestion product oligonucleotides generated both coincidentally from the open reading frame (ORF) regions of multiple mRNAs, as well as by design due to the common 5′ and 3′ untranslated regions (UTRs) used by all three manufactured mRNAs.

Application in SNP Detection. In addition to identification of RNA sequence, oligonucleotide maps with high sequence coverage can be applied to detecting low-level impurities such as SNPs. An SNP in an mRNA drug substance can be detected so long as the oligonucleotide containing the SNP is not identical to any other oligonucleotides produced in the same digestion. Such detection can be achieved by MS (when no/little interference is present in MS^1) or by MS/MS (when MS^1 interference is present). To simulate the presence of SNPs in an mRNA drug substance, altered Epo and Luc mRNAs were created with designed point mutations and then spiked into "normal" Epo and Luc mRNA at various levels. The resulting mixtures were then subjected to endonuclease digests followed by LC-MS and LC-MS/MS analyses.

An example of SNP detection by LC-MS is shown in Figure 5. Despite the presence of other, overlapping isotopic clusters, the isotopic cluster of the SNP-containing oligonucleotide U(A→U)CCUUCUUG (monoisotopic m/z 1037.10, $z = 3$) is quite clear at 0.5% w/w. Absolute signal abundance from the single isotope at m/z 1037.10 shows high linear correlation with the amount of SNP added. The same signal may also be normalized against the total (SNP-containing plus original oligonucleotide) signal abundance, with similarly high linear correlation (Figure S6). Together, these results demonstrate the feasibility of SNP quantitation by LC/MS.

In contrast to the previous scenario, which involves overlapping but partially mass-resolved isotopic clusters, Figure 6 portrays a slightly more complex situation in which the isotopic cluster of the SNP oligonucleotide (U→A)-ACCUUUUCAUCACG (monoisotopic m/z 1577.20, $z = -3$) is wholly eclipsed by that of another oligonucleotide. Despite the lack of usable information at the MS level, MS/MS may still be applied to reveal a number of unique, isomer- or isobar-specific diagnostic ions: in this case, the c_2^- and c_3^- fragments of the SNP-containing oligonucleotide.

SNP detection was achieved at 1.0% or lower in all spike-in experiments, across all endonucleases and all mRNA sequences tested (Table S1). In general, the limit of detection of the SNP is determined by the limit of detection of the digested oligonucleotide containing the SNP. As digested oligonucleotides are typically detected at varied intensities (depending on ionization and/or digestion efficiency), some sequence-specific variance in the actual limit of detection is to be expected.

Detection of sequence impurities by oligonucleotide mapping is achieved by detection of the signature oligonucleotides produced by such impurities. In addition to the SNPs demonstrated above, oligonucleotide mapping can detect other types of mRNA sequence impurities such as insertions, deletions, duplications, and frameshifts, so long as the corresponding signature oligonucleotides are detectable within the digestion products of a given endonuclease. When another oligonucleotide interferes with detection of the impurity oligonucleotide (such as by coincidentally having the same mass and retention time), switching endonucleases and/or altering LC gradients can help avert such interference. Orthogonal endonucleases such as cusativin,[33,34] RNase U2,[35] or MC1[36,37] might be integrated into the workflow for such cases. Alternatively, if a platform approach is not required, custom oligonucleotide guides could be used in conjunction with RNase H to produce a site-specific cut for high sequence coverage.[38]

CONCLUSION

Parallel digestions of long mRNA by multiple endonucleases of varying cleavage frequency enable mRNA oligonucleotide sequence mapping with significantly higher sequence coverage than digestion by any single endonuclease. High-resolution, accurate mass MS data can be supplemented by MS/MS fragmentation to identify isomeric oligonucleotides. Oligonucleotide mapping provides high-confidence mRNA sequence identification and verification, enabling identification of a single correct sequence from hundreds of sequences of similar length. Low-level sequence impurities such as SNPs can also be detected and quantified at sub-1% levels.

Currently, development of mRNA medicines continues to accelerate day by day, underscoring the need for analytical technologies capable of thoroughly characterizing this new drug modality. Novel methods such as parallel ribonuclease

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506

digestion followed by LC-MS sequence mapping provide a wealth of useful information and may be directly applied to analysis of biopharmaceutical mRNA.

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b01664.

Additional experimental details (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: serenus.hua@modernatx.com

ORCID

Tao Jiang: 0000-0002-0449-7136

Kanchana Ravichandran: 0000-0001-5050-9719

Present Address

⊥Mildred Kissai, Department of Chemistry, The Scripps Research Institute 10550 North Torrey Pines Rd La Jolla, CA 92037.

Notes

The authors declare the following competing financial interest(s): All authors are current or former employees of Moderna and receive (or have previously received) compensation such as salary, stocks, or stock options.

■ REFERENCES

(1) Sahin, U.; Karikó, K.; Türeci, Ö. *Nat. Rev. Drug Discovery* **2014**, *13*, 759.

(2) Bahl, K.; Senn, J. J.; Yuzhakov, O.; Bulychev, A.; Brito, L. A.; Hassett, K. J.; Laska, M. E.; Smith, M.; Almarsson, Ö.; Thompson, J.; et al. *Mol. Ther.* **2017**, *25* (6), 1316−1327.

(3) Pardi, N.; Hogan, M. J.; Pelc, R. S.; Muramatsu, H.; Andersen, H.; DeMaso, C. R.; Dowd, K. A.; Sutherland, L. L.; Scearce, R. M.; Parks, R.; et al. *Nature* **2017**, *543*, 248.

(4) Kormann, M. S.; Hasenpusch, G.; Aneja, M. K.; Nica, G.; Flemmer, A. W.; Herber-Jonat, S.; Huppmann, M.; Mays, L. E.; Illenyi, M.; Schams, A.; et al. *Nat. Biotechnol.* **2011**, *29*, 154.

(5) Thess, A.; Grund, S.; Mui, B. L.; Hope, M. J.; Baumhof, P.; Fotin-Mleczek, M.; Schlake, T. *Mol. Ther.* **2015**, *23* (9), 1456−1464.

(6) Sanger, F.; Nicklen, S.; Coulson, A. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74* (12), 5463−5467.

(7) Arsenic, R.; Treue, D.; Lehmann, A.; Hummel, M.; Dietel, M.; Denkert, C.; Budczies, J. *BMC Clin. Pathol.* **2015**, *15* (1), 1−9.

(8) Kohlmann, A.; Klein, H.-U.; Weissmann, S.; Bresolin, S.; Chaplin, T.; Cuppens, H.; Haschke-Becher, E.; Garicochea, B.; Grossmann, V.; Hanczaruk, B.; et al. *Leukemia* **2011**, *25* (12), 1840.

(9) Srebalus Barnes, C. A.; Lim, A. *Mass Spectrom. Rev.* **2007**, *26* (3), 370−388.

(10) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367.

(11) Zhang, Z.; Shah, B.; Bondarenko, P. V. *Biochemistry* **2013**, *52* (45), 8165−8176.

(12) Choudhary, G.; Wu, S.-L.; Shieh, P.; Hancock, W. S. *J. Proteome Res.* **2003**, *2* (1), 59−67.

(13) Kowalak, J. A.; Pomerantz, S. C.; Crain, P. F.; McCloskey, J. A. *Nucleic Acids Res.* **1993**, *21* (19), 4577−4585.

(14) Ehrich, M.; Nelson, M. R.; Stanssens, P.; Zabeau, M.; Liloglou, T.; Xinarianos, G.; Cantor, C. R.; Field, J. K.; van den Boom, D. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (44), 15785−15790.

(15) Ross, R.; Cao, X.; Yu, N.; Limbach, P. A. *Methods* **2016**, *107*, 73−78.

(16) Reynolds, K. J.; Yao, X.; Fenselau, C. *J. Proteome Res.* **2002**, *1* (1), 27−33.

(17) Ingrosso, D.; Fowler, A. V.; Bleibaum, J.; Clarke, S. *Biochem. Biophys. Res. Commun.* **1989**, *162* (3), 1528−1534.

(18) Hossain, M.; Limbach, P. A. *Anal. Bioanal. Chem.* **2009**, *394* (4), 1125.

(19) Gupta, R. C.; Randerath, K. *Nucleic Acids Res.* **1977**, *4* (10), 3441−3454.

(20) Kirpekar, F.; Douthwaite, S.; Roepstorff, P. *RNA* **2000**, *6* (2), 296−306.

(21) Sato, K.; Egami, F. Studies on Ribonucleases in Takadiastase. *J. Biochem.* **1957**, *44* (11), 753−767.

(22) Ogawa, T.; Inoue, S.; Yajima, S.; Hidaka, M.; Masaki, H. *Nucleic Acids Res.* **2006**, *34* (21), 6065−6073.

(23) Yajima, S.; Inoue, S.; Ogawa, T.; Nonaka, T.; Ohsawa, K.; Masaki, H. *Nucleic Acids Res.* **2006**, *34* (21), 6074−6082.

(24) Muñoz-Gómez, A. J.; et al. *FEBS Lett.* **2004**, *567* (2−3), 316−320.

(25) Luna-Chávez, C.; Lin, Y.-L.; Huang, R. H. *J. Mol. Biol.* **2006**, *358* (2), 571−579.

(26) Sterckx, Y.; De Gieter, S.; Zorzini, V.; Hadži, S.; Haesaerts, S.; Loris, R.; Garcia-Pino, A. *Protein Expression Purif.* **2015**, *108*, 30−40.

(27) Eighth International Conference on Methods in Protein Sequence Analysis. *J. Protein Chem.* **1990**, *9* (3), 247−368. .

(28) Samuelson, J. C.; Morgan, R. D.; Benner, J. S.; Claus, T. E.; Packard, S. L.; Xu, S. *Nucleic Acids Res.* **2006**, *34* (3), 796−805.

(29) Solivio, B.; Yu, N.; Addepalli, B.; Limbach, P. A. *Anal. Chim. Acta* **2018**, *1036*, 73−79.

(30) Krylov, S. N.; Dovichi, N. J. *Anal. Chem.* **2000**, *72* (12), 111−128.

(31) Azarani, A.; Hecker, K. H. *Nucleic Acids Res.* **2001**, *29* (2), No. e7.

(32) Mullard, A. *Nat. Rev. Drug Discovery* **2018**, *17*, 460.

(33) Rojo, M. A.; Arias, F. J.; Iglesias, R.; Ferreras, J. M.; Muñoz, R.; Escarmís, C.; Soriano, F.; López-Fando, J.; Méndez, E.; Girbés, T. *Planta* **1994**, *194* (3), 328−338.

(34) Addepalli, B.; Venus, S.; Thakur, P.; Limbach, P. A. *Anal. Bioanal. Chem.* **2017**, *409* (24), 5645−5654.

(35) Uchida, T.; Arima, T.; Egami, F. *J. Biochem.* **1970**, *67* (1), 91−102.

(36) Suzuki, A.; Yao, M.; Tanaka, I.; Numata, T.; Kikukawa, S.; Yamasaki, N.; Kimura, M. *Biochem. Biophys. Res. Commun.* **2000**, *275* (2), 572−576.

(37) Addepalli, B.; Lesner, N. P.; Limbach, P. A. *RNA* **2015**, *21*, 1746−1756.

(38) Taoka, M.; Nobe, Y.; Yamaki, Y.; Sato, K.; Ishikawa, H.; Izumikawa, K.; Yamauchi, Y.; Hirota, K.; Nakayama, H.; Takahashi, N.; Isobe, T. *Nucleic Acids Res.* **2018**, *46* (18), 9289−9298.

DOI: 10.1021/acs.analchem.9b01664
Anal. Chem. 2019, 91, 8500−8506