A call for direct sequencing of full-length RNAs to identify all modifications

For most organisms, DNA sequences are available, but the complete RNA sequences are not. Here, we call for technologies to sequence full-length RNAs with all their modifications.

Juan D. Alfonzo, Jessica A. Brown, Peter H. Byers, Vivian G. Cheung, Richard J. Maraia and Robert L. Ross

RNA determines cell identity and mediates responses to cellular needs. Such diverse cellular functions arise from the vast chemical composition of RNA comprising four canonical ribonucleotides (A, C, G and U) and more than 140 modified ribonucleotides (Fig. 1). Many years of RNA research laid the foundation for the development of RNA therapeutics as diverse as antisense oligonucleotide therapy for spinal muscular atrophy, and mRNA vaccines. These remarkable accomplishments were enabled by modified ribonucleotides, yet the 'true' sequence of RNA, i.e., the 'RNome', remains unknown. This key knowledge gap in understanding the building blocks of RNA must be filled. Here, we call for the development of high-throughput methods to sequence RNA directly on a transcriptome-wide scale and the necessary informatics to identify all RNA variants at the single-molecule level.

RNA is not an exact copy of DNA: processing steps such as splicing, editing, and base and sugar modification distinguish RNA sequences from their DNA templates. These modifications, including 140 known modified ribonucleotides and counting, influence RNA structure and function by affecting how the RNA interacts with other nucleic acids and regulatory proteins. Yet, how all the modified nucleotides are distributed in RNA transcripts remains unknown. This information gap stems from the lack of methods to sequence full-length RNAs directly. The technology that we call RNA sequencing is misleading; instead, a more accurate term would be complementary DNA (cDNA) sequencing, because RNAs are converted back to DNA by reverse transcription and then sequenced. In the RNA-to-DNA conversion, important

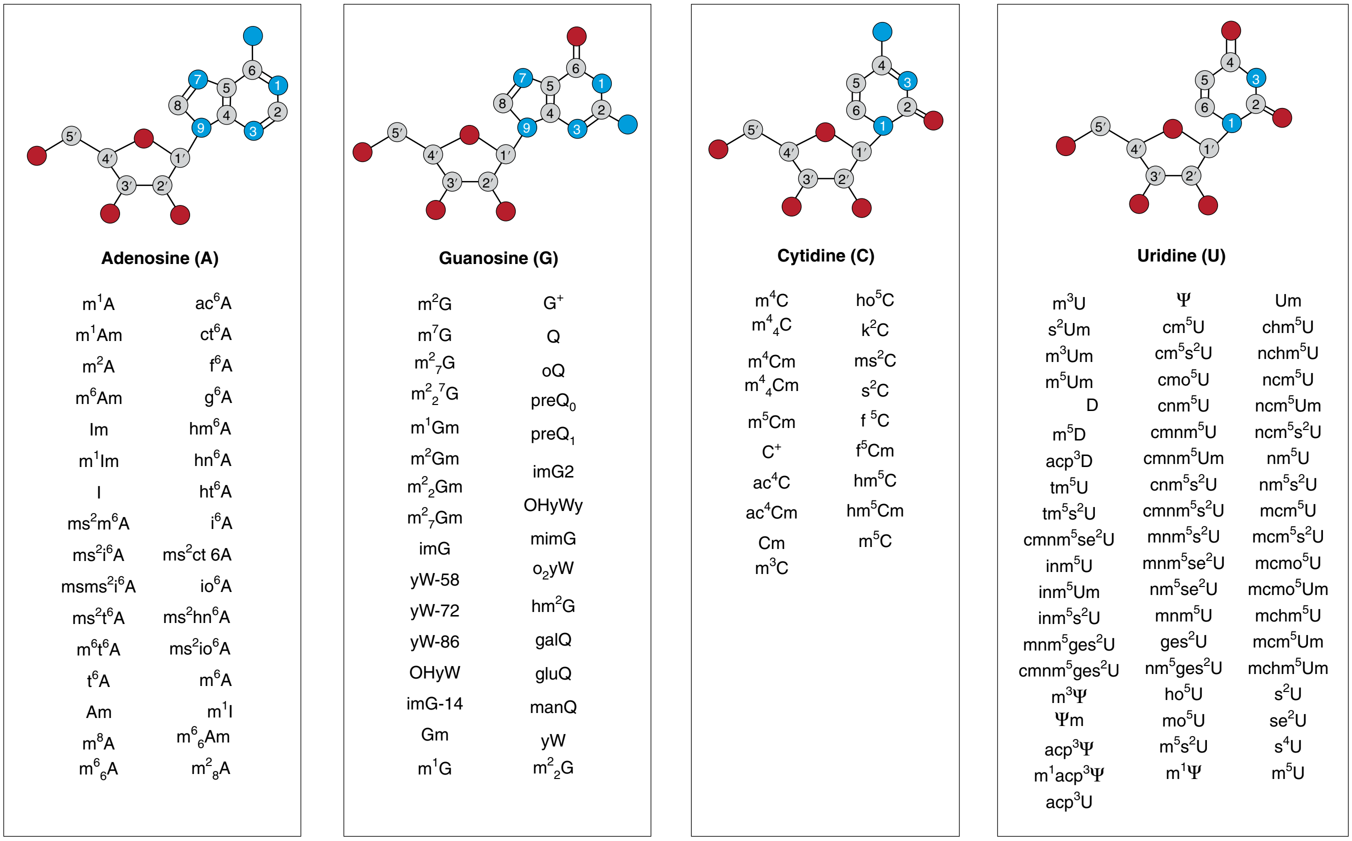

Fig. 1 | Chemical modifications of RNA. Of the more than 140 different modifications that occur in all types of RNAs, approximately ten can be mapped to specific sequence contexts through various methods discussed in this Comment. Methods are needed that can detect and quantify all the modifications to obtain complete RNA sequences. Modification nomenclature is as described in Modomics, http://genesilico.pl/modomics/.

information on nucleotide modifications is lost. No existing technology can determine the identity and position of all modifications simultaneously in full-length RNAs at both the single-molecule and transcriptome-wide scales. The reliance on cDNA sequences has led to a failure to obtain and understand key regulatory codes in the RNome of human cells and other organisms, including many infectious viruses with RNA genomes.

A collective effort is needed to develop the technologies necessary to sequence RNAs in ways that preserve and read the modifications. The resulting technologies should enable gathering of the necessary information on RNA sequence diversity among and within cells. Direct RNA sequencing would reveal the dynamics of modifications and provide a wealth of information, including modifications and splice sites that are pathogenic, and their consequences at the cellular and organismal levels.

Why do we need 'true' RNA sequences?

Base and sugar modifications, as well as splicing, affect RNA chemical properties, topology and function. Knowledge of the types and locations of the modifications and splice sites, their extent and their interrelationships is necessary for a basic understanding of how nucleic acids regulate cellular and organismal function and how dysregulation leads to diseases.

Defects in RNA modifications (which are distinct from splice-site alterations) account for more than 100 human diseases, including childhood-onset multiorgan failures, cancers and neurologic disorders. These conditions are now referred to as 'RNA modopathies'[1,2]. This number is likely to represent only a small percentage of the actual number of existing RNA modopathies.

To date, research has focused almost exclusively on DNA sequence analysis (usually whole exome) to identify the genetic causes of undiagnosed diseases. In this setting, causative variants have been identified in approximately 25–30% of individuals. Whole-genome sequences and long-fragment sequences can identify additional variants. Despite intense efforts, the genetic and molecular basis for many diseases remains unknown. Studies of gene expression have filled some gaps by improving the classification of diseases, such as different forms of lymphoma[3] and breast cancer[4]. Some of those studies have even guided treatments[5]. This current view of RNA, although incomplete, has led to the understanding that dysregulation of RNA processing underlies many conditions, such as amyotrophic lateral sclerosis, cancer and metabolic syndromes[6,7]. However, to pinpoint the RNA-processing steps that are dysregulated in diseases, we must know the different isoforms of each RNA within a transcriptome and the counts of individual transcripts. Direct RNA sequencing would enhance understanding of individual variations in gene expression, facilitate the determination of precise steps in RNA processing and reveal which processes are pathogenic when dysregulated.

The past year has brought acute awareness of diseases caused by viruses with RNA genomes. As we have learned, RNA viruses replicate and mutate quickly. To help identify these viruses in early phases of disease outbreaks and to eradicate them, obtaining their 'true' sequences would be valuable; the indirect cDNA sequences currently available present incomplete information. Direct RNA sequencing would help molecular epidemiologists determine mutation types, propagation patterns and the phylogeny of viruses, to understand how new viruses arise and to design effective eradication programs[8,9].

RNA sequences that include modifications would also enhance drug development. For RNA-based therapeutics, such as antisense oligonucleotides and mRNA vaccines, RNA sequences could be used to improve the design of probes that inactivate target sequences or mimic sequences for vaccine development. Designing RNA therapeutics on the basis of cDNA sequences provides only rough guidance; therefore, greater precision is required.

Although RNA is single-stranded, its three-dimensional structure is often complex and poorly characterized. Among RNAs, the structures of tRNAs are the best characterized. Those characterization studies have shown that modified nucleotides in the anticodon loops affect structure and function[10]. Beyond tRNAs, modified nucleotides in mRNAs are just beginning to be identified, and modifications are increasingly being reported. Long noncoding RNAs are also modified, and these modifications have been linked to physiological function[11]. Currently, most structural analyses of RNA are performed with polymers of the four canonical nucleotides (A, C, G and U) without any modifications. Because relatively little is known about how the different modified nucleotides pair with one another or the four canonical ribonucleotides, the prediction of RNA structure remains rudimentary. Technologies such as cryo-electron microscopy offer unprecedented opportunities to examine the structures of RNA–protein complexes, so that the 'true' RNA sequences better reflect in vivo interactions. As imaging technologies improve, the shapes of the RNA molecules instead of their cDNA surrogates will be able to be visualized.

In summary, RNA sequences with their modifications constitute the 'true' information content of RNA. The RNome is needed to usher in an era of molecular and clinical studies based on a solid foundation of sequences and structures.

Establishing the 'true' sequence of full-length RNA

Methods to map modifications can be classified into two groups: (1) indirect, which require manipulation of the RNA before sequence determination and (2) direct, in which the nucleotides, including their modifications, are identified by inspection of the RNA.

Indirect methods usually rely on sequencing by synthesis, wherein RNA is converted to cDNA via reverse transcriptase. This approach has limited ability to identify modified nucleotides. One exception is inosine, which is read in this context as guanosine and can be faithfully identified as an A-to-G mismatch by comparison with the corresponding DNA sequence. In some instances, because the reverse transcriptase cannot bypass the modified nucleotides, the reactions prematurely terminate at those positions, and modifications at the more 5′-located positions are not identified.

Some indirect sequencing approaches identify base and ribose modifications; however, each type of modification is studied with unique methods[12,13]. Consequently, the modifications are identified separately rather than together in the same experiment, and sequence context is not captured. Some methods involve manipulation of bulk RNA samples with chemicals or enzymes that specifically target a particular modification and can differentially reveal the presence of the modification during the subsequent sequencing step. For example, bisulfite treatment of C5-methylated nucleic acids (m^5C in RNA and 5mC in DNA, per current notation convention) leads to the efficient deamination of the unmodified cytosines to form uridine, whereas m^5C remains intact. After sequencing of bisulfite-treated nucleic acids, any C that remains unchanged is assumed to be m^5C in the original sequence. Recent improvements in this technique include formamide during bisulfite treatment, which has greatly increased the accuracy of base-calling[14].

Alternatively, the known propensity of reverse transcriptase to make mistakes at modified nucleotides has been exploited to map certain modifications. Here, the mutational 'signature' in the ratios of canonical-nucleotide misincorporation changes according to the modification type[15]. This signature has been used in the mapping of pseudouridine, 1-methyladenosine (m^1A) and 5-methylcytosine (m^5C) in bulk RNA (as recently reviewed in ref. [16]). Calling of modified nucleotides via reverse transcriptase nucleotide misincorporation, when performed in tandem with enzymatic dealkylation, has also been demonstrated to be useful in mapping several methylation sites[17]. However, the above techniques are not currently applicable to most modifications, and they rely heavily on a priori knowledge of a given modification. In addition, because they apply to populations rather than individual molecules, the extent of transcript diversity is unknown. Finally, this approach makes de novo mapping of previously known and unknown modifications at new positions in RNA challenging at best.

Currently, two direct methods are used to identify the positions and types of modification in RNA: mass spectrometry[18] and nanopore technology[19,20]. These methods can identify known modifications and even previously unknown modifications. Liquid chromatography and tandem mass spectrometry (LC–MS/MS) is considered the 'gold standard' for modification analysis, because it directly measures RNA. Modification mapping of RNA by LC–MS/MS can handle only small RNA oligonucleotides (~17-nucleotide maximum length), and the approach involves digestion of the RNA sample with nucleotide-specific nucleases. This procedure is followed by ionization and fragmentation, which yield the mass of the RNA strand as well as the sequence context, with modified nucleotides mapped to their respective positions within the strand. Technological advances in mass spectrometers and chromatographic systems have improved the sensitivity of this analytic technique, thus resulting in the acquisition of more precise data while consuming less sample. However, with the existing instruments and software, weakly abundant transcripts within a population cannot be identified, and longer RNA fragments cannot be sequenced; thus, technologic developments are much needed. The generation of manageable fragments for mass spectrometry sequencing is limited by the current availability of known nucleotide-specific nucleases. The lack of robust and specific nucleases currently presents a larger challenge to this method than the actual accuracy and sensitivity of the detection platform. Additionally, informatics will be required to process the resulting LC–MS/MS data.

Nanopore sequencing is a promising technology that analyzes nucleic acids as single molecules (approximately 20 kb)[19,20]. The technology involves the detection of ion currents generated, in the context of a polynucleotide chain, as single nucleotides pass through small pores. For example, α-hemolysin and other proprietary pore-forming proteins have been used[21,22]. The samples are processed at approximately 400 nucleotides per second, and the identities of the nucleotides are determined on the basis of unique current signals from different nucleotides. Nanopore sequencing is a computationally intensive enterprise; base-calling is achieved by neural networks, but direct base-calling can generally yield errors as high as 10–15%. A recent report[23] has shown that, similarly to indirect sequencing, a combination of direct base-calling and signatures from 'base-calling errors' may improve accuracy. This combination method has been used to map canonical nucleotides, as well as several methylated derivatives and pseudouridines[20,23–30]. Nonetheless, the technology reads signals from several nucleotides at a time and thus has not achieved single-nucleotide resolution. In addition, the extent of false positives when this technology is used for transcriptome-wide studies is unknown. Much progress has been made in using nanopores for DNA sequencing[31–33] and identifying several DNA modifications[34–39]. However, the diversity of modifications found in RNA poses a greater challenge, because this technology relies on base-calling algorithms to assign modifications. These algorithms are currently limited by the general lack of RNA-modification standards needed to train the algorithms. Most modifications cannot be synthesized chemically or enzymatically to provide such training standards. The generation of standards for all modifications is a major challenge in training nanopore or any other technology heavily relying on computational methods to map modifications in RNA accurately. The implementation of new synthesis protocols to generate modification standards is currently very limited but offers a favorable avenue for exploring new chemistries. Ultimately, the limitations of nanopores for mapping RNA modifications cannot be fully ascertained until standards are established for all modifications. Notwithstanding some limitations, this technology has potential because of its ability to sequence single long transcripts in mixed populations. Support for the development of RNA-based sequencing and the detection of nucleotide modifications should extend the promise of these techniques.

A call for action

The number of sequenced genomes and transcriptomes has rapidly expanded in recent years, and massive databases rich in information are freely available for analysis. After the DNA sequence and transcriptome were characterized, the resultant information had originally been expected to be sufficient to connect genotype to phenotype. We now realize that this is far from the case.

RNAs are highly processed co-transcriptionally and post-transcriptionally. Therefore, the RNA modifications must be known to understand how they contribute to function. We argue that RNA modifications may be a form of missed variants, and thus incorporating them in genetic analyses would enhance the mechanistic understanding of diseases. The success of mRNA vaccines opens the door to using more RNA therapeutics to treat infectious diseases and replacing missing or mutant transcripts in genetic disorders[40]. To realize the full potential of RNA in disease prevention and treatment, technologies are needed to sequence RNA directly at the single-molecule level.

Like technologies for DNA sequencing, those for direct RNA sequencing must be high throughput. Although each cell contains only two copies of DNA, it contains many copies of RNA. Compounding the issue, the modifications are likely to be highly dynamic. Thus, methods are needed to sequence RNA easily and cost effectively.

As mentioned above, DNA-sequencing methods can already identify modifications. Determining whether those approaches can be readily applied to RNA would be prudent. Nanopore sequencing is a promising path forward, although this technology will require the development of standards for each modification in a different sequence context to improve the accuracy of base-calling algorithms so that real-time transcriptome-wide experiments become feasible.

Here, we call for an investment of funds and infrastructure to develop technologies to sequence full-length RNA and the informatics to detect and identify all modifications. Innovations to develop standards for each modification, instruments that sequence RNA directly, and computational methods that support those instruments and analyze the results are all needed. Admittedly, the resources, technology and informatics necessary to sequence RNA directly will be on the scale of the Human Genome Project. By building on the Human Genome Project's success, complete RNA sequences should advance understanding of gene regulation and lead to new frontiers in health and medicine.

Juan D. Alfonzo[1], Jessica A. Brown[2], Peter H. Byers[3,4], Vivian G. Cheung[5]✉, Richard J. Maraia[6] and Robert L. Ross[7]

[1]Department of Microbiology; Center for RNA Biology and Ohio State Biochemistry Program, Ohio State University, Columbus, OH, USA. [2]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA. [3]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA. [4]Department of Medicine (Medical Genetics), University of Washington, Seattle, WA, USA. [5]Department of Pediatrics, Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA. [6]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA. [7]Department of Cancer and Cell Biology, Metabolomics Mass Spectrometry Laboratory, University of Cincinnati, Cincinnati, OH, USA.

✉*e-mail: vgcheung@med.umich.edu*

Published online: 15 July 2021
https://doi.org/10.1038/s41588-021-00903-1

References

1. Jonkhout, N. et al. *RNA* **23**, 1754–1769 (2017).
2. Suzuki, T. *FASEB J* https://doi.org/10.1096/fasebj.2020.34.s1.00132 (2020).
3. Alizadeh, A. A. et al. *Nature* **403**, 503–511 (2000).
4. van 't Veer, L. J. et al. *Nature* **415**, 530–536 (2002).
5. Rosenwald, A. et al. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
6. Nussbacher, J. K., Tabet, R., Yeo, G. W. & Lagier-Tourenne, C. *Neuron* **102**, 294–320 (2019).
7. Goodall, G. J. & Wickramasinghe, V. O. *Nat. Rev. Cancer* **21**, 22–36 (2021).
8. Moya, A., Holmes, E. C. & González-Candelas, F. *Nat. Rev. Microbiol.* **2**, 279–288 (2004).
9. Connallon, T. & Clark, A. G. *Genetics* **190**, 1477–1489 (2012).
10. Colby, D. S., Schedl, P. & Guthrie, C. *Cell* **9**, 449–463 (1976).
11. Torsin, L. I. et al. *Int. J. Mol. Sci.* **22**, 581 (2021).
12. Schwartz, S. et al. *Cell* **159**, 148–162 (2014).
13. Meyer, K. D. et al. *Cell* **149**, 1635–1646 (2012).
14. Khoddami, V. et al. *Proc. Natl Acad. Sci. USA* **116**, 6784–6789 (2019).
15. Potapov, V. et al. *Nucleic Acids Res.* **46**, 5753–5763 (2018).
16. Anreiter, I., Mir, Q., Simpson, J. T., Janga, S. C. & Soller, M. *Trends Biotechnol.* **39**, 72–89 (2021).
17. Zheng, G. et al. *Nat. Methods* **12**, 835–837 (2015).
18. Sutton, J. M., Guimaraes, G. J., Annavarapu, V., van Dongen, W. D. & Bartlett, M. G. *J. Am. Soc. Mass Spectrom.* **31**, 1775–1782 (2020).
19. Cozzuto, L. et al. *Front. Genet.* **11**, 211 (2020).
20. Garalde, D. R. et al. *Nat. Methods* **15**, 201–206 (2018).
21. Song, L. et al. *Science* **274**, 1859–1866 (1996).
22. Derrington, I. M. et al. *Proc. Natl Acad. Sci. USA* **107**, 16060–16065 (2010).
23. Liu, H. et al. *Nat. Commun.* **10**, 4079 (2019).
24. Drexler, H. L., Choquet, K. & Churchman, L. S. *Mol. Cell* **77**, 985–998.e8 (2020).
25. Parker, M. T. et al. *eLife* **9**, e49658 (2020).
26. Jenjaroenpun, P. et al. *Nucleic Acids Res.* **49**, e7 (2021).
27. Aw, J. G. A. et al. *Nat. Biotechnol.* **39**, 336–346 (2020).
28. Smith, M. A. et al. *Genome Res.* **30**, 1345–1353 (2020).
29. Motorin, Y. & Marchand, V. *Genes (Basel)* **12**, 278 (2021).
30. Begik, O. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00915-6 (2021).
31. Clarke, J. et al. *Nat. Nanotechnol.* **4**, 265–270 (2009).
32. Miga, K. H. et al. *Nature* **585**, 79–84 (2020).
33. Jain, M. et al. *Nat. Biotechnol.* **36**, 338–345 (2018).
34. Wallace, E. V. B. et al. *Chem. Commun. (Camb.)* **46**, 8195–8197 (2010).
35. Gigante, S. et al. *Nucleic Acids Res.* **47**, e46 (2019).
36. Simpson, J. T. et al. *Nat. Methods* **14**, 407–410 (2017).
37. McIntyre, A. B. R. et al. *Nat. Commun.* **10**, 579 (2019).
38. Ni, P. et al. *Bioinformatics* **35**, 4586–4595 (2019).
39. Rand, A. C. et al. *Nat. Methods* **14**, 411–413 (2017).
40. Servick, K. *Science* **370**, 1388–1389 (2020).

Acknowledgements

We thank the following colleagues with whom we shared this Comment and who agreed to add their names in support of it: M. Akeson, M. G. Bartlett, S. J. Baserga, B. L. Bass, T. J. Begley, S. L. Berger, A. J. Berman, P. J. Blackshear, J. M. Bujnicki, N. K. Conrad, T. K. Dayie, P. C. Dedon, W. V. Gilbert, M. Gorospe, M. Hafner, A. E. Hargrove, A. Hasse, C. He, M. Helm, J. E. Jackman, K. Koutmou, A. Lal, A. T. Lamm, B. Lee, J. T. Lee, N. Lee, S. A. Leidel, E. Y. Levanon, P. A. Limbach, J. Lis, A. Makarov, R. M. Mitton-Fry, C. W. Mueller, A. Nath, K. M. Neugebauer, T. Pan, T. Pederson, A. M. Phillippy, K. J. Riley, J. J. C. Rosenthal, S. Schwartz, J. A. Steitz, M. F. Summers, T. Suzuki, J. Sztuba-Solinska, B. S. Tolbert, T. Trcek, N. G. Walter, J. A. Watts and J. Zhang.

Author contributions

All authors contributed equally to the writing of this Comment.

Competing interests

The authors declare no competing interests.