Messenger RNA modifications: Form, distribution, and function

Wendy V. Gilbert,1 Tristan A. Bell,1,2† Cassandra Schaeining1,3†

RNA contains more than 100 distinct modifications that promote the functions of stable noncoding RNAs in translation and splicing. Recent technical advances have revealed widespread and sparse modification of messenger RNAs with N\(^6\)-methyladenosine (m\(^6\)A), 5-methylcytosine (m\(^5\)C), and pseudouridine (Ψ). Here we discuss the rapidly evolving understanding of the location, regulation, and function of these dynamic mRNA marks, collectively termed the epitranscriptome. We highlight differences among modifications and between species that could instruct ongoing efforts to understand how specific mRNA target sites are selected and how their modification is regulated. Diverse molecular consequences of individual m\(^6\)A modifications are beginning to be revealed, but the effects of m\(^5\)C and Ψ remain largely unknown. Future work linking molecular effects to organismal phenotypes will broaden our understanding of mRNA modifications as cell and developmental regulators.

The first modified RNA nucleoside was identified almost 60 years ago by analyzing salt-soluble RNA from yeast (1). Since then, more than 100 chemically distinct modified nucleotides have been characterized, most of which were identified in tRNAs and other abundant noncoding RNAs from diverse organisms (2). N\(^6\)-methyladenosine (m\(^6\)A) was the first internal mRNA modification discovered (3, 4), and at ~1 to 3 m\(^6\)A residues per message, it is abundant enough to be readily detected by bulk mRNA analysis. Next-generation sequencing approaches have allowed mapping of the locations of m\(^6\)A and less-abundant modified nucleosides. Here we summarize these methods and the mRNA modification landscape they reveal. We then discuss the enzymes responsible for installing m\(^6\)A, m\(^5\)C, and pseudouridine (Ψ) at specific sites and emphasize questions regarding the basis for target specificity and regulation. Finally, we describe the remaining challenges in determining the functions of mRNA modifications, which have the potential to regulate genes with widespread consequences for development and disease.

Transcriptome-wide mapping of m\(^6\)A, m\(^5\)C, and Ψ

All high-throughput methods for locating modified nucleosides rely on one of two approaches: Either antibodies are used to isolate modified RNA fragments for sequencing or some modification-selective RNA chemistry is exploited (Fig. 1, A to C). The first genome-wide RNA modification maps were generated using antibodies against m\(^6\)A to identify thousands of ~100 nucleotide RNA fragments containing the modification in mammalian cells (5, 6). This approach has been adapted to give single-nucleotide–resolution m\(^6\)A maps by cross-linking the antibody-RNA complexes and determining the sites of cross-link–induced ethylcarbodiimide p-tosylate (CMCT), a bulky covalent adduct that creates a block to reverse transcriptase (RT) (10). The use of a “click” chemistry–compatible CMC coupled to biotin allows pre-enrichment of Ψ-containing RNA fragments (11). Different computational strategies have been used to identify Ψ sites from analysis of CMC-dependent RT stops in yeast and human cells (11–14), which may explain some discrepancies between Ψ annotations.

The potential effect of an mRNA modification depends on both the molecular consequences and the percentage of transcripts that are modified. For example, a modification that leads to accelerated mRNA decay is unlikely to have much biological effect if only 1% of transcripts are modified, whereas a modification that causes an alternative protein variant to be produced could be functionally important, even at very low levels. A limitation of current m\(^6\)A and Ψ profiling methods is the lack of quantitative information about the extent of modification. Changes in the relative enrichment of a particular sequence in m\(^6\)A pull-downs from different growth states have been used to infer regulation of modification, but the absolute fraction of mRNA that is modified cannot be determined from these data. Similarly, differences in the abundance of CMC-dependent reads can indicate relative changes in pseudouridylation when comparing the same mRNA site in different conditions but cannot be compared between different sites due to sequence-dependent capture biases in library preparation. A method has been developed to enable absolute quantitation of modified nucleosides at specific mRNA sites (15), but this technique does not scale to allow parallel measurements at many sites. High-throughput methods to quantitatively profile m\(^6\)A and Ψ would considerably advance the field.

In contrast to m\(^6\)A and Ψ, the level of m\(^5\)C at specific sites in RNA or DNA can be quantitatively determined by bisulfite sequencing. For transcriptome-wide analysis of m\(^5\)C, isolated RNA is treated with bisulfite to convert unmethylated cytosines to uracils before cDNA synthesis. The extent of modification at each cytosine residue is then determined by observing the rate of nonconversion—the fraction of reads that do not show the expected C to T sequence change—assuming complete conversion of unmethylated C nucleosides. The first map of m\(^5\)C claimed more than 8000 candidate sites in mRNAs from human (HeLa) cells (16); however, these sites could include other cytosine modifications known or suspected to interfere with bisulfite conversion (17) and may include false positives from stochastic nonconversion events. More targeted high-throughput methods exploit the catalytic mechanism of m\(^5\)C methyltransferases to trap covalent intermediates formed between these enzymes and their target sites, which are then identified by RNA sequencing after immunoprecipitation of the methyltransferase (18). Targeted bisulfite sequencing at candidate m\(^5\)C sites allows verification and quantitation of modification and could be scaled to monitor hundreds

---

1Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. 2Graduate Program in Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. 3Graduate Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. *Corresponding author. Email: wgilbert@mit.edu †These authors contributed equally to this work.

"...mRNA modifications have the potential to affect most posttranscriptional steps in gene expression."

mutations within enriched RNA fragments (7, 8). Antibody-based modification profiling has the advantage of concentrating sequencing efforts on sites of interest.

There is some evidence for artifactual enrichment of mRNA fragments lacking m\(^6\)A. In budding yeast, where the only known m\(^6\)A-generating enzyme is not essential for cell viability, it was possible to rigorously determine the background association of unmethylated mRNA fragments with anti-m\(^6\)A antibodies. Notably, almost half of the putative m\(^6\)A peaks were found to be methyltransferase independent (9), suggesting a potentially widespread problem with the current m\(^6\)A maps in many systems. Higher-confidence m\(^6\)A sites can be identified by performing methylome mapping after genetic manipulation of methyltransferases and/or demethylases or by overlaying m\(^6\)A peak profiles with maps of methyltransferase interaction sites obtained by cross-linking and immunoprecipitation (CLIP) approaches.

Pseudouridine mapping relies on chemical strategies to selectively derivatize Ψ nucleosides with N-cyclohexyl-N'-beta-(4-methylmorpholinum)
of sites using microfluidics-based multiplex polymerase chain reaction and deep sequencing (20).

Although noncoding RNAs are extensively modified in all organisms, the modification landscape of prokaryotic mRNAs has barely begun to be explored. m6A has been reported in mRNA from *Escherichia coli* and *Pseudomonas aeruginosa* (20), and m5C has been mapped in *Sulfolobus solfataricus* mRNAs (21). Because there are substantial differences between prokaryotic and eukaryotic mRNA metabolism, the functional consequences of mRNA modifications are likely to differ as well. It will be interesting to see how these ancient and conserved RNA modifications are harnessed for posttranscriptional gene regulation in organisms with widely differing lifestyles.

Reproducibility and orthogonal validation give confidence in mark annotations. The m6A methylomes of different human cell lines are highly overlapping (5, 22), genetic manipulation of methyltransferases and demethylases produces the expected changes in m6A signals (9, 22–24), and thousands of m6A target RNAs have been shown to cross-link to the modifying enzymes in vivo (24, 25). In contrast, for m5C and Ψ, there are still few transcriptome-wide profiles and almost no independent analysis of similar cell types or growth states by different groups. Because functionally important sites are likely to be identified in multiple studies, a critical next phase will be the determination of m5C and Ψ sites whose detection is robust to technical variation.

mRNA-modifying enzymes and target specificity

In mammalian mRNA, m6A is primarily produced by the methyltransferases METTL14 and METTL3, orthologs of yeast IME4. METTL3 and METTL14 associate with the regulatory subunit WTAP (Wilms’ tumor 1–associating protein) to form a 200-kDa methyltransferase complex (24, 25). Knockout mouse embryonic stem cell lines lacking METTL3 or METTL14 display up to 99% reductions in bulk mRNA methylation and reduced m6A signal at thousands of sites (22, 26). Most mammalian m6A sites are found within the consensus sequence Rm6ACH (R = G or A, H = A, C, or U), which is consistent with the enriched binding motifs observed in CLIP studies of METTL14, METTL3, and WTAP (GGAC, GACU, and GACU, respectively) (24).

Despite the strong consensus, only a small fraction of RACH sites are detectably methylated in vivo, arguing that the sequence motif is not sufficient to determine the distribution of m6A. Furthermore, m6A sites are strongly biased toward the 3′ ends of transcripts in organisms as diverse as yeast and humans (5, 6, 9). This distribution suggests a functional coupling between modification and other RNA processing steps. There is also substantial enrichment of m6A sites in the vicinity of stop codons in a variety of mammalian cell types (5, 6, 22), which hints at a role for the ribosome. A simple model can predict which RGAC sites will be methylated in yeast using...
just three features: the sequence flanking the site, the proximity of the site to the 3′ end, and the predicted secondary structure at the site (unstructured favored). Expanding the consensus sequence to ANRACNNNU yielded the greatest predictive power; nevertheless, only ~10% of sites that matched this consensus were observed to be methylated (9). It remains unclear how much of the observed site specificity of m6A modifications is due to the intrinsic substrate preferences of these enzymes, and it would be useful to systematically and quantitatively test modification of diverse methyltransferase substrates.

Additional methyltransferase enzymes may also modify mRNAs with m6A. METTL4 is closely related to METTL14 and METTL3 and is predicted to have catalytic activity (27). Although knockdown of METTL4 had no detectable effect on bulk m6A levels in mRNA from HeLa cells (24), the sensitivity of this assay is limited and would not detect changes in modification of a small subpopulation of mRNA substrates. Human cells express at least one active m6A methyltransferase in addition to METTL14 and METTL3, as suggested by a study of an m6A site on the human U6 small nuclear RNA (snRNA). The sequence context of this site does not match the RACH consensus, and methylation of this site cannot be outcompeted by an mRNA target of METTL3-METTL14 (28).

Pseudouridine sites in mRNA are much less abundant than m6A, yet the enzymology is more complex (Table 1). There is evidence for mRNA pseudouridylation by 8 out of 10 yeast Ψ synthases, whose canonical targets are mostly sites in tRNA but also include snRNA and ribosomal RNA sites (12–14). A plurality of identified mRNA substrates in yeast have been genetically assigned to two tRNA-modifying enzymes: Pus1 (12), whose human ortholog also modifies mRNAs (II), and Pus7 (I3). The basis for Pus1’s target site specificity has long puzzled those studying tRNA modification. Pus7 modifies mRNA sites found in the specific sequence context UGGAR, in agreement with its known tRNA target site preferences. However, the structural context of this motif is critical for pseudouridylation of noncoding RNAs (29), which may explain why only a tiny fraction of RNAs with UGUAR motifs are modified. Pseudouridylation by all Pus enzymes likely involves recognition of RNA secondary structures, the details of which remain to be discovered. Transcriptome-wide RNA structure probing techniques should facilitate identification of the structural determinants for mRNA pseudouridylation, similar to work on m6A (30). mRNA pseudouridylation has been determined in only a few eukaryotic cell types thus far. Because the Ψ synthases that modify mRNAs in yeast and human cells have homologs in all domains of life, mRNA pseudouridylation is likely to be widely distributed.

Comparatively little is known about the enzymology of mC deposition in mRNA. Although mC is common in noncoding RNAs from all domains of life, there are marked species differences in the reported prevalence of mC in mRNA (~8000 mC sites in human mRNAs compared with a single mRNA mC site in budding yeast (16, 21). Saccharomyces cerevisiae expresses three known mC methyltransferases: Ncl1, which modifies multiple tRNAs at several positions, and Nop2 and Rem1, which target the ribosomal RNA. In human cells, the methyltransferases Dnmt2 and Nsun2 have been shown to modify certain mRNAs (16, I8), but their verified targets do not include most reported mC sites, suggesting that additional enzymes may be active toward mRNA substrates. Humans have seven known proteins in the Nsun methyltransferase family (Nop2 and Nsun2 to Nsun7), of which five have been shown to have catalytic activity, and all seven have the predicted active-site cysteine residues. No sequence motifs are common among reported mC sites in human mRNAs, which is expected if multiple enzymes are responsible. In contrast to humans, all known mC sites in S. solfataricus exactly match the sequence of a ribosomal mC site, suggesting a common enzyme (21).

**Table 1. mRNA-modifying enzymes and modification readers.** m, mammalian; y, yeast.

<table>
<thead>
<tr>
<th>Factors</th>
<th>m6A</th>
<th>Ψ</th>
<th>m5C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Writers</strong></td>
<td>METTL14 (m) (22, 26), METTL3 (m) (22, 26), METTL4? (m) (27)</td>
<td>Pus1 (m, y) (II–I3), Pus2 (y) (I2, I3), Pus3 (y) (I2, I3), Pus4 (y) (II–I3), Pus6 (y) (I2, I3), Pus7 (m, y) (II–I3), Pus9 (y) (I2, I3), Cbf5 (y) (I2, I3)</td>
<td>Dnmt2 (m) (I6, I8), Nsun2 (m) (I6, I8)</td>
</tr>
<tr>
<td><strong>Erasers</strong></td>
<td>FTO (m) (23), ALKBH5 (m) (31)</td>
<td>None identified</td>
<td>None identified</td>
</tr>
<tr>
<td><strong>Readers</strong></td>
<td>YTH family (m) (32, 36, 40), YTHDF2 (m) (5, 32)</td>
<td>None identified</td>
<td>None identified</td>
</tr>
</tbody>
</table>

**Regulation of mRNA modifications**

m6A is a truly dynamic mRNA modification that can be enzymatically removed by demethylase “erasing” enzymes. Fat mass and obesity-associated protein, FTO, was the first mammalian RNA demethylase implicated in the dynamics of m6A modifications (23). ALKBH5 is a second, conserved eraser of m6A marks that is highly expressed in the testes and is required for spermatogenesis and fertility in mice (37). The sequence or structural preferences of the demethylases may account for some or all of the observed specificity in the distribution of m6A among individual RACH sites.

Alternatively, the m6A “writing” and “erasing” enzymes may have only minimal sequence requirements (for RACH motifs), allowing the accessibility of potential sites and/or the availability of competing RNA-binding proteins to shape the landscape of m6A modification. An example of this mode of regulating m6A comes from studies of the heat shock response in mammalian cells (MEF and HeLa) (32). It is plausible that m6A modification is somehow coupled to cotranscriptional RNA processing events, given the pronounced 3′ bias of modified sites. Moreover, shifts in the relative abundance of m6A in 5′ compared with 3′UTRs (3′ untranslated regions) (e.g., upon heat shock (32, 33)) could arise from a global change in recruitment of methyltransferases or demethylases to promoters.

Changes in enzyme or substrate localization may also play an important role in regulating mRNA modifications. In yeast, heat shock induces relaxation of the Ψ synthase Pus7 from the nucleus to the cytoplasm. This shift correlates with a 10-fold increase in the number of Pus7-dependent pseudouridylated mRNA target sites (I3), which could reflect an increased window of opportunity for Pus7 to interact with potential mRNA substrates. Similarly, conditions that decrease the rate of mRNA export, either globally or for specific messages, could increase mRNA modification by nuclear enzymes. In S. cerevisiae, the Ime4-Mum2-Slb1 m6A methyltransferase complex localizes to nucleoli at the stage of meiosis in which m6A mRNA levels peak, and mutations in SLZ1 that prevent nucleolar accumulation of Ime4 also reduce m6A levels by about a factor of 3 (9). Another apparent localization-dependent regulatory mechanism involves heat shock-induced nuclear accumulation of the mammalian m6A “reader” protein, YTHDF2, which competes with the FTO demethylase for binding to m6A sites (32). The observation that heat shock causes a substantial shift in the distribution of m6A from 5′ to 3′ sites implies that 5′ sites are more efficiently targeted by demethylases during normal growth and may be more inherently dynamic.

Pseudouridylation is thought to be irreversible, so its observed dynamics are likely mediated by the production or degradation of pseudouridylated messages. It is also possible that the
enzymes are nuclear during normal growth. This m6A shows highly dynamic changes during mitotic potential. Consistent with this possibility, mRNA modifications may have similar regulation as reported in other contexts. mRNA turnover rapidly, reversibly and irreversible does not work in the nucleus and that mRNAs can be turned over rapidly, reversible and irreversible mRNA modifications may have similar regulatory potential. Consistent with this possibility, m5A shows highly dynamic changes during mitosis in S. cerevisiae, despite the apparent lack of an m5A demethylase (9).

It is notable that most known mRNA-modifying enzymes are nuclear during normal growth. This localization allows m5A, and likely other modifications, to influence mRNA biogenesis from the earliest stages and may enable coupling between transcriptional control circuits and mRNA modification state.

**Functional consequences**

At the molecular level, mRNA modifications have the potential to affect posttranscriptional steps in gene expression. Some broad regulatory themes have emerged from mechanistic studies of mRNA regulation by m5A, which has been linked to control of mRNA stability (26, 36, 37), splicing (38, 39), and translational efficiency (32, 33, 40) and to pri-miRNA (miRNA) processing (41) (Fig. 2). m5A, and likely other modifications, can mediate diverse effects on mRNA metabolism by affecting interactions with RNA binding proteins. YTHDF2—the first modification-specific RNA binding protein, or “reader,” identified (5)—increases turnover of m5A-modified mRNA by promoting colocalization with decay factors (36). This destabilizing function of m5A plays an important biological role in stem cell differentiation by regulating key pluripotency factors (26, 37). Individual YTH proteins interact with distinct subsets of m5A sites and produce different effects on gene expression when perturbed in different cellular contexts (32, 36, 40), suggesting combinatorial control by uncharacterized co-regulators.

Dedicated reader proteins have not yet been identified for m5C or Ψ, but the ribosome has the potential to function as a universal reader of mRNA modifications. Incorporation of single m5A, m5C, or Ψ modifications in specific codon contexts reduced protein production in E. coli by 20 to 70% (42), a level of repression exceeding that of many conserved microRNA target sites (43). There is also the possibility of encoding truncated proteins through site-specific ribosome stalling at modified codons, depending on how the stall is resolved. Perhaps the most exciting potential for mRNA modifications to affect gene function is through regulated rewiring of the genetic code, and there are limited but intriguing observations supporting this possibility. Insertion of m5C led to 4% recoding of proline as leucine in E. coli (42), and Ψ-containing stop codons were efficiently mistranslated as specific amino acids in budding yeast (44). The effects of modified nucleosides on translation have been tested in very limited contexts, and the mechanisms responsible for noncanonical decoding events are unknown. If coding-sequence modifications lead to the production of alternative protein variants, even low-occupancy modifications could have substantial biological effects.

RNA modifications may broadly influence mRNA metabolism through their effects on RNA structures. m5A destabilizes RNA duplexes in vitro (45), and m5A sites in mRNA tend to be unstructured in vivo (30). Both m5C and Ψ affect tRNA folding and are likely to affect mRNA structures as well (46). By changing the accessibility of binding sites for regulatory
factors, the effect of RNA structure may be direct (e.g., inhibiting splicing or translation initiation by blocking access to functional sites) or indirect, as was recently shown for m^2A structural switches affecting splicing (39). Given the stabilizing effects of Ψ on RNA structure and the central role of RNA structure in prokaryotic gene regulation, it will be particularly interesting to see if eubacterial and archaean mRNAs are also stabilizing effects of m^2A. m^2A addition is a trans-acting factor that works similarly from anywhere within the 3′UTR. On the other hand, modification sites that control alternative splicing or protein recoding should be conserved at the nucleotide level. Thus, examining the patterns of conservation could illuminate the likely molecular function of a modification, as well as its biological importance.

With the use of CRISPR-based genome engineering, it is currently possible to directly test the influence of altering any modification site in many organisms. CRISPR multiplexing strategies could also potentially permit interrogation of many sites in parallel and hasten functional discoveries. As a complementary approach, it would be valuable to promote de novo modification at specific sites by engineering fusions between catalytic domains (e.g., methyltransferase, Ψ synthase) and a programmable RNA binding domain scaffold such as Pumilio.

The expanding epitranscriptome
Just how complex is the epitranscriptome? High-throughput sequencing methods were required to reveal the presence of sparse Ψ modifications in mRNA, suggesting that new approaches will discover still more previously unseen mRNA modifications. In fact, while this Review was in preparation, the first maps of 5′-hydroxymethylcytidine and N^4-methyladenosine sites were reported using new sequencing techniques to examine mRNA from flies, yeast, and mice (48–50). Given the enormous diversity of RNA modifications found in tRNA and the demonstrations that many tRNA-modifying enzymes also have mRNA substrates (11–14, 16, 18), the mRNA-modification landscape is likely to be very rich indeed.

The past 5 years have witnessed a revolution in our understanding of the extent and diversity of mRNA modifications. Many clever enrichment and chemical modification strategies coupled to high-throughput sequencing have produced maps of m^2A, m^{2-5}C, and Ψ across the transcriptomes of many organisms and demonstrated the dynamics of these modifications across different cellular growth states. Now that diverse molecular functions of mRNA modifications are beginning to emerge, the field is poised for breakthroughs in understanding the most important cellular and organismal functions of mRNA modification.

REFERENCES AND NOTES
38. X. Zhao et al., Cell Res. 24, 1403–1419 (2014).

ACKNOWLEDGMENTS
We thank members of the Gilbert lab for discussion. This work was supported by grants from the NIH (GM01316 and CA187236) and the American Cancer Society (RSG-13-396-01-RMC) to W.V.G., a NIH Pre-Doctoral Training Grant (T32GM07287) to T.A.B., and an NSF Graduate Research Fellowship to C.S.
Messenger RNA modifications: Form, distribution, and function

Wendy V. Gilbert, Tristan A. Bell and Cassandra Schaening

Science 352 (6292), 1408-1412.
DOI: 10.1126/science.aad8711