Trends in Cell & Molecular Biology
Vol. 11, 2016

Mini-Review

# The life of mRNA and its genome-wide analysis

**Geetha Durairaj[1,#], Sukesh Bhaumik[1] and David A. Lightfoot[2,*]**

[1]Department of Biochemistry and Molecular Biology, Southern Illinois University School of Medicine, Carbondale, IL, 62901; [2]Genomics Core Facility, Department of Plant Soil and Agricultural Systems, and Center for Excellence in Soybean Research, Teaching and Outreach, Southern Illinois University, Carbondale, IL 62901-4415, USA.

## ABSTRACT

Expression of genes transcribed by RNA Polymerase II is vital in maintaining various cellular processes including differentiation and development. Gene expression involves the synthesis, processing and export of mRNA from nucleus to cytoplasm for translation to proteins. Misregulation of any of these steps or altered gene expression is associated with cellular malfunctions. Therefore, a large number of studies have been focused on understanding these steps towards regulation of gene expression. Here, we discuss the steps associated with the life of mRNA and the application of high-throughput methods in genome-wide analysis of mRNA.

**KEYWORDS:** mRNA, gene expression, RNA polymerase II, export, translation

## INTRODUCTION

Life of a protein starts with the synthesis of messenger RNA (mRNA) in the nucleus by RNA Polymerase II [1]. Subsequently, mRNA is processed, and then gets exported from the nucleus to cytoplasm *via* nuclear pores in a transcription-dependent manner. In the cytoplasm, mRNA is translated to proteins, and finally, gets degraded. Therefore, gene expression includes a number of steps that starts inside the nucleus and ends in the cytoplasm. Here, we discuss the life of mRNA right from its synthesis until its degradation, and the high-throughput methods in mRNA analysis.

### The life of mRNA

The journey of an mRNA starts with transcription. When the nascent mRNA is about 20-24 nucleotides long, it gets capped at its 5'-end to be associated with the cap-binding complex. As mRNA synthesis progresses, various RNA-binding proteins (RBPs) that aid in 3'-end processing, splicing, nuclear exosome-mediated surveillance and export get loaded onto mRNA [2, 3]. While many of these factors are co-transcriptionally associated with the transcribing mRNA, some factors are recruited post-transcriptionally as well. These are the evolutionarily conserved factors, which dictate the pattern of expression of mRNA. These factors involved in the synthesis of functional mRNA are physically and functionally coupled, and ensure that functional mRNAs exit the nucleus. Together with mRNA, these proteins constitute the messenger ribonucleoparticle (mRNP).

The voyage of mRNA inside the nucleus is the preparatory phase in the making of functional mRNAs. While 5' cap and 3'-end processing machineries target almost all transcripts that are synthesized, splicing is limited to intron-containing mRNAs. Splicing is mediated by spliceosome [4]. Once introns are removed, a complex known as EJC (Exon Junction Complex) gets loaded onto

*Corresponding author: ga4082@siu.edu

[#]Present address: Department of Biological Chemistry, University of California, Irvine, CA 92697, USA.

specific positions on mRNA just upstream to the exon junctions. This complex contains many factors that are involved in mRNA export and nuclear surveillance mechanisms [5]. Evidence exists to prove that not a single step in nuclear mRNA journey is distinct [6, 7]. Hence, it is impossible to look at any of the above-mentioned steps in isolation. Once mRNPs are ready to be exported out of the nucleus, a type of quality control check is carried out on their sequences with the aid of nuclear surveillance complex, known as exosome [8]. It degrades the aberrant messengers and only properly processed transcripts exit the nucleus with the aid of export receptors and chaperone proteins.

In the cytoplasm, most of the mRNAs enter the translationally active pool that encodes proteins. The nuclear cap structure that is bound to newly exported mRNA interacts with translation initiation factor, elF4G [9, 10]. This interaction aids in the recruitment of small ribosomal subunit, which starts scanning for the presence of start codon. Once a start codon is identified, the large ribosomal subunit is recruited, ultimately leading to the formation of an active polysome complex. mRNAs are threaded through the space between the ribosome complex to undergo a pioneering round of translation [11, 12]. This process removes any associated hnRNPs (heterogeneous nuclear ribonucleoproteins) that might hinder the translational activity. At this stage, nuclear 5' cap and 3' Poly (A) binding nuclear protein (PABNP1 in yeast) are replaced by elF4E (cytoplasmic cap binding protein) and Poly (A) binding cytoplasmic protein (PABPC), respectively. All these events result in efficient translation of the message [13, 14].

All mRNAs have a limited life span which is largely dictated by the efficiency of mRNA degradation machineries. In eukaryotes, mRNA degradation mainly occurs through two pathways, each of them requiring a gradual shortening of the poly (A) tail [15]. Following the tail reduction, cap is removed by de-capping enzymes and mRNA is degraded by the action of exonucleases or exosomes [1, 15]. The degradation of mRNA occurs in certain cytoplasmic foci, known as P-bodies (processing bodies) [16]. P-bodies are enriched with numerous proteins involved in mRNA degradation [16-19]. While the major pathways of mRNA degradation require exonucleases, endonucleolytic degradation pathways also exist in cytoplasm. Endonucleolytic degradation occurs by sequence-specific mRNA cleavage *via* Dicer-associated miRNAs (microRNAs) and siRNAs (small interfering RNAs) [20, 21]. The sequence-specific interactions of mRNAs with miRNAs also target mRNAs to P-bodies. In addition to degrading translationally active mRNAs, the degradation machinery is also crucial in destroying mRNAs that lack a proper translational stop signal (Nonstop decay) or that possess premature stop signal (nonsense-mediated decay) [15].

While above-mentioned decay mechanisms target the mRNAs in the translationally active pool, not all mRNAs enter into such pools upon export. Rather, they are held in translationally quiescent state inside P-bodies, thereby adding up to the complexity in the study of these bodies [18, 19, 22]. Silenced mRNAs are generally observed in metazoan embryos where the cell stores mRNAs that are required for development. Such silent storage of mRNAs is mediated by shortening of their Poly (A) tails which is mediated by CPEB, a protein that binds to cytoplasmic polyadenylation element (CPE) in 3'UTR [22]. CPEB interacts with some proteins (4E-binding proteins) that compete with elF4G for binding to elF4E, thereby masking translation initiation. There are many proteins like CPEB that mask translation initiation. This effect can be reversed under conditions that favor the translation of these mRNAs [1]. Together these steps define the lifetime of mRNA right from its birth in nucleus to its degradation in cytoplasm.

**Analysis of mRNA**

Isolation of mRNAs in every step of its life presents a significant challenge. Although it is easy to use high-density sucrose gradient sedimentation of cell fractions followed by isolation of mRNAs associated with each fraction, the involvement of the same RBPs in various steps in the life of mRNA limits this approach. A combination of genomic tools with biochemical approaches has evolved into a field, known as Ribonomics [23, 24]. Ribonomics employs the use of modern technologies to isolate and study the fraction of mRNAs associated with RBPs.

To understand the abundance of mRNA from its initiation to degradation, it is important to measure mRNA during initiation, hnRNPs' accumulation, transport, polysome formation, and early and late phases of degradation. Identification of mRNAs with RBPs was limited to *in vitro* methods, and was challenging until the advent of immunoaffinity method developed in Jack Keene's lab [23-25]. This method relies on the biochemical purification of mRNP complex containing the target RBP along with the associated mRNAs followed by microarray profiling of mRNAs. This method is similar to chromatin immunoprecipitation (ChIP)-based microarray analysis or ChIP on Chip, but with mRNA. Tagged or endogenous mRNA-binding proteins are isolated using specific antibodies, and associated mRNAs are characterized by *en masse* assay. This method has been named as RIP-Chip (RNA-binding protein immunoprecipitation-microarray profiling) [25-27]. Another recently developed method for immunopurifying RBP-associated mRNAs involves UV crosslinking of RNA-protein complex followed by affinity purifications [28, 29]. This method is highly stringent in purifying specific RNA-protein complex of interest, and is known as CLIP (UV-crosslinking-immunoprecipitation). CLIP has been widely used in the characterization of specific mRNAs in complex with RBPs of interest. Further, this technique can be modified to mediate the isolation of mRNAs associated with specific steps of their life. However, many RBPs are common in different steps in the life of mRNA in the nucleus. For instance, cap-binding complex associates with nascent mRNA during early phase of elongation and remains bound to it until it is replaced by cytoplasmic cap. This is just one example of various RBPs including the conserved hnRNPs that bind to mRNA during different steps of its journey. This makes the process of isolation of mRNAs associated with specific steps during its expression pretty challenging.

During initiation of mRNA synthesis, ser-5-phosphorylated (Ser-5-P) form of RNA polymerase II recruits the capping enzymes to the 5'-end of mRNA [30]. This occurs immediately after the birth of the nascent transcript. Once mRNA is capped, various factors that aid in post-transcriptional processing and export get loaded onto the active gene. For instance, an important factor that gets co-transcriptionally loaded onto the active gene is TREX (Transcription/Export) complex that couples transcription elongation to export [6, 7]. TREX is a multi-protein complex that is involved in early steps of mRNA export from nucleus to cytoplasm. Apart from TREX, numerous 3'-end processing and splicing factors also get associated with the active gene co-transcriptionally. Due to tight functional and physical couplings of numerous RBPs, it is a big challenge to demarcate and isolate mRNA during every single step of transcription and associated processing events.

hnRNPs are functional proteins that prepare mRNA for export. Accumulation of hnRNPs marks that the mRNA is ready to be exported out of nucleus. In yeast, four main hnRNP proteins have been well-studied (Hrb1, Hrp1, Npl3 and Nab2) while ~20 hnRNP proteins have been documented in metazoans. Different hnRNPs bind to different transcripts and hence it is difficult to isolate mRNAs from a particular hnRNP complex. hnRNPs sediment at 30S on a velocity gradient, [31] and hence associated mRNAs with hnRNP complexes can be isolated from the sedimented fraction. mRNAs associated with nuclear pore complex proteins (Nups) during its export process can be specifically purified using RIP or CLIP techniques by generating antibodies against specific components of nuclear pore complex or by epitope tagging strategy.

In the cytoplasmic phase, mRNAs associated with polysomes can be isolated using the sucrose density gradient. In this method, cell extracts are fractionated through a sucrose gradient and absorbance of various fractions are monitored at 254 nm [32]. Fraction that corresponds to polysomes can then be isolated based on its high absorbance and mRNAs can be obtained. Alternatively, specific subunits of the ribosome can be immunopurified and the associated mRNAs can then be isolated [33]. Biggest challenge in specific mRNA isolation in cytoplasm lies in the separation of early and late degradation stages of mRNAs. All mRNAs from different stages of translation are de-capped and de-adenylated before undergoing degradation and aggregate in P-bodies with their associated degradation enzymes [18]. However, isolation of

mRNAs that are targeted for degradation in P-bodies needs deeper strategies as biochemical purification and sedimentation of P-bodies is itself a challenge for biochemists.

**Genome-wide mRNA analysis**

Genome-wide mRNA analysis is based on hybridization and sequencing methods to quantitate the mRNA levels globally [34-36]. Hybridization-based mRNA analysis on a genome is carried out using DNA microarray. This methodology is based on the principle of hybridization between the probe and target molecule (i.e., cDNA) from the sample. Probes are sequences that correspond to each gene in the organism and hence represent the genome of the organism to be tested. Depending upon the probes, there are two types of DNA microarrays: cDNA microarray and oligonucleotide microarrays [34, 35]. cDNA microarrays contain fragments of cDNA that are around 600-2400 nucleotides in length. cDNAs are constructed by amplifying the genes from cDNA libraries using PCR (polymerase chain reaction) and then spotted on to the glass slides using robotic techniques. The cDNA microarrays have high detection sensitivity due to longer fragments. However, these arrays suffer from cross-hybridization problems. Besides using cDNA probes, short oligonucleotide sequences representing a single gene can be spotted on a slide. They may be longer (60-mer probes in case of Agilent arrays) or shorter (20-mer probes from Affymetrix with each gene containing a dozen 20-mers scanning different locations in gene). These arrays are known as high density oligonucleotide or primer arrays. Unlike one or several copies of cDNA in cDNA microarrays, the oligonucleotide microarrays contain two times a set of probe pairs for each gene. While one set has a perfect match of oligonucleotides to the gene, other set has a mismatched nucleotide at the middle of the sequence [36]. Expression is then determined by calculating the difference between the hybridization of a sample to the matched probe minus the mismatched one. Oligonucleotide arrays usually hybridize to one labeled sample per array with appropriate internal controls. Various oligonucleotide microarray platforms are commercially available from vendors like Agilent, Affymetrix and Illumina. The differences between the platforms from various vendors lie in the size of the probes and the procedures for hybridization of the sample. Due to high information content of these probes, they are used in hybridization-based detection of mutation analysis as well as gene expression studies.

cDNA and oligonucleotide arrays measure expression of mRNA differently. Choice of the platforms solely depends on the features spotted on the arrays and also economical aspects involved in the experiments. A comprehensive comparison of various platforms was carried out by many groups and the outcomes show a difference in concordance of results across platforms [37, 38]. Hence, good internal controls and data normalization strategies are necessary in carrying out these experiments. While microarrays are the preferable choice of global gene expression analysis of well-annotated genomes, these methods are biased to already existing knowledge about the genome, and hence cannot be applied to less characterized genomes. This can be overcome by sequencing-based next generation technologies as discussed below.

Sequencing-based genome-wide analysis is a technique for studying the expression analysis of known as well as unknown genomes. This methodology also quantifies the expression of genes that are under-represented in hybridization-based arrays due to low mRNA abundance. A tool for high-throughput sequencing-based gene expression analysis is SAGE (Serial Analysis of Gene Expression) which is widely used to get a snapshot of the mRNAs that are expressed under a given set of conditions. In this method, cDNAs are collected and digested into smaller fragments using specific restriction enzymes. The fragments from digested cDNAs of different genes are then ligated to form concatamerized tags which can be cloned into plasmid vector to generate SAGE library. Sequencing and counting the tags reveal which genes are expressed and how often. A variation of this method which is being widely used now is Super-SAGE. Super-SAGE is similar to SAGE except for the length of the sequence tag that is produced. Super-SAGE synthesizes 26-bp sequence tags in contrast to 20-bp tags in SAGE [39].

While SAGE and Super-SAGE have contributed much to the field of genomics, the advent of additional high-throughput sequencing technologies have resulted in dramatic reduction in the cost of sequencing. Three platforms are widely used in today's next generation transcriptome sequencing: Roche-454 pyrosequencing, Illumina analyzer and SOLiD sequencing [40, 41]. The Roche-454 pyrosequencing technique is based on pyrosequencing strategy, where incorporation of single nucleotide releases a pyrophosphate, which in turn produces light by downstream reactions. In Illumina analyzer-based method, the fluorescently labeled nucleotides with their 3'-OH group blocked are added simultaneously and the incorporation is recorded by an image detector. In SOLiD sequencing-based method, four different fluorescently labeled di-base probes compete for ligation to the primer sequence complementary to the adaptor sequence on DNA. Ligation of specific probe to the primer sequence results in chemical cleavage of the octamer to release the label which is recorded, and the steps are repeated for desirable lengths.

High-throughput genome methods result in a massive amount of data. They have to be analyzed in an efficient way to interpret the results of these techniques. Data analysis is done using various software packages that accompany the array chips from commercial vendors. Analysis of microarray data involves pre-processing, inferential statistics and explorative statistics. Background signal subtraction (global or local) and various statistical tests like t-test, cluster analysis and ANOVA (analysis of variance) are employed in analyzing the data [42]. However, before analysis of the data, it has to be normalized due to differences in the experimental approaches like unequal quantities of starting RNA, efficiency of the labeled-dyes and the systematic bias in the expression levels. Normalization includes a scaling factor and then calculates the normalized ratios between the sample and the reference sets. The difference in expression patterns of the same sample set across various platforms and experiments is mainly due to the difference in normalization. Hence, certain standards were set up to analyze and interpret the microarray data. MIAME (Minimum Information About a Microarray Experiment) protocol was developed to interpret and verify the data obtained from microarray analysis [43]. Standard microarray data model and format MAGE (MicroArray and Gene Expression) were also developed with contributions from many organizations like Affymetrix, Agilent and Iobion.

## CONCLUSION

Expression of RNA Polymerase II genes is correlated with mRNA levels. Altered levels of mRNA (and hence gene expression) are linked to a variety of cellular disorders and diseases (2). Thus, a better understanding of the control and contribution of every step in the life of mRNA are crucial, and such knowledge will help to develop therapies towards maintaining normal cellular functions. Towards this goal, we have discussed here the life of mRNA from its birth in the nucleus to its degradation in the cytoplasm, and various techniques for its genome-wide analysis.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## REFERENCES

1. Moore, M. J. 2005, Science, 309, 1514-1518.
2. Durairaj, G., Garg, P. and Bhaumik, S. R. 2009, RNA Biol., 6, 531-535.
3. Jimeno, S. and Aguilera, A. 2010, J. Biol., 9, 6.
4. Fleckner, J., Zhang, M., Valcarcel, J. and Green, M. R. 1997, Genes Dev., 11, 1864-1872.
5. Le, H. H., Izaurralde, E., Maquat, L. E. and Moore, M. J. 2000, EMBO J., 19, 6860-6869.
6. Bergkessel, M., Wilmes, G. M. and Guthrie, C. 2009, Cell, 136, 794-794.
7. Aguilera, A. 2005, Curr. Opin. Cell Biol., 17, 242-250.
8. Lebreton, A. and Séraphin, B. 2010, Biochimica et Biophysica Acta - Gene Regulatory Mechanisms, 1779, 558-565.
9. Lejeune, F., Ranganathan, A. C. and Maquat, L. E. 2004, Nat. Struct. Mol Biol., 11, 992-1000.
10. Prevot, D., Darlix, J. L. and Ohlmann, T. 2003, Biol. Cell, 95, 141-156.

11. Maquat, L. E., Tarn, W. Y. and Isken, O. 2010, Cell, 142, 368-374.
12. Gao, Q., Das, B., Sherman, F. and Maquat, L. E. 2005, Proc. Natl. Acad. Sci. USA, 102, 4258-4263.
13. Sachs, A. B. and Varani, G. 2000, Nat. Struct. Biol., 7, 356-61.
14. Mangus, D. A., Evans, M. C. and Jacobson, A. 2003, Genome Biol., 4, 223.
15. Houseley, J. and Tollervey, D. 2009, Cell, 4, 763-776.
16. Fillman, C. and Andersen, L. J. 2005, Curr. Opin. Cell Biol., 17, 326-331.
17. Parker, R. and Sheth, U. 2007, Mol. Cell, 5, 635-646.
18. Eulalio, A., Ansmant, B, I. and Izaurralde, E. 2007, Nat. Rev. Mol. Cell Biol., 8, 9-22.
19. Wickens, M. and Goldstrohm, A. 2003, Science, 300, 753-755.
20. Liu, J., Valencia-Sanchez, M. A., Hannon, G. J. and Parker, R. 2005, Nat. Cell Biol., 7, 719-723.
21. Filipowicz, W., Bhattacharyya, S. U. and Sonenberg, N. 2008, Nature Reviews Genetics, 9, 102-114.
22. Richter, J. D. and Sonenberg, N. 2005, Nature, 7025, 477-480.
23. Tenenbaum, S. A., Lager, P. J., Carson, C. C. and Keene, J. D. 2002, Methods, 26, 191-198.
24. Tenenbaum, S. A., Carson, C. C., Lager, P. J. and Keene, J. D. 2000, Proc. Natl. Acad. Sci. USA, 97, 14085-14090.
25. Penalva, L. O., Tenenbaum, S. A. and Keene, J. D. 2004, Methods Mol. Biol., 257, 125-134.
26. Baroni, T. E., Chittur, S. V., George, A. D. and Tenenbaum, S. A. 2008, Methods Mol. Biol., 419, 93-108.
27. Keene, J. D., Komisarow, J. M. and Friedersdorf, M. B. 2006, Nat. Protoc., 1, 302-307.
28. Wang, Z., Tollervey, J., Briese, M., Turner, D. and Ule, J. 2009, Methods, 48, 287-293.
29. Ule, J., Jensen, K., Mele, A. and Darnell, R. B. 2005, Methods, 37, 376-386.
30. Ahn, S. H.., Kim, M. and Buratowski, S. 2004, Mol. Cell, 13, 67-76.
31. Reed, R. and Magni, K. 2001, Nat. Cell Biol., 9, 201-204.
32. Inada, T., Winstall, E., Tarun, S. Z. Jr., Yates, J. R. 3rd, Schieltz, D. and Sachs, A. B. 2002, RNA, 8, 948-958.
33. Halbeisen, R. E., Scherrer, T. and Gerber, A. P. 2009, Methods, 48, 306-310.
34. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. 1995, Science, 270, 467-470.
35. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. L. 1996, Nature Biotechnol., 14, 1675-1680.
36. Butte, A. 2002, Nature Reviews, 1, 951-961.
37. Tan, P. K., Downey, T. J., Spitznagel, E. L. Jr. and Xu, P. 2003, Nucleic Acids Res., 31, 5676-5684.
38. Woo, Y., Affourtit, J., Daigle, S., Viale, A., Johnson, K., Naggert, J. and Churchill, G. 2004, J. Biomol. Tech., 15, 276-84.
39. Matsumura, H., Krüger, D. H., Kahl, G. and Terauchi, R. 2008, Curr. Pharm. Biotechnol., 9(5), 368-74.
40. Mardis, E. R. 2008, Annual Review of Genomics and Human Genetics, 9, 387-402.
41. Nagarajan, N. and Pop, M. 2010, Methods Mol. Biol., 673, 1-17.
42. Quackenbush, J. 2002, Nature Genetics Supplement, 32, 496-502.
43. Brazma, A. 2001, Nat. Genet., 29, 365-371.