REVIEW

Advances in analyzing RNA diversity in eukaryotic transcriptomes: peering through the Omics lens [version 1; referees: 3 approved]

Sushant Bangru¹, Auinash Kalsotra ¹-³

¹Department of Biochemistry, University of Illinois at Urbana-Champaign, Illinois, USA
²Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Illinois, USA
³College of Medicine, University of Illinois at Urbana-Champaign, Illinois, USA

Abstract

Alternative splicing, polyadenylation, and chemical modifications of RNA generate astonishing complexity within eukaryotic transcriptomes. The last decade has brought numerous advances in sequencing technologies that allow biologists to investigate these phenomena with greater depth and accuracy while reducing time and cost. A commensurate development in biochemical techniques for the enrichment and analysis of different RNA variants has accompanied the advancement of global sequencing analysis platforms. Here, we present a detailed overview of the latest biochemical methods, along with bioinformatics pipelines that have aided in identifying different RNA variants. We also highlight the ongoing developments and challenges associated with RNA variant detection and quantification, including sample heterogeneity and isolation, as well as ‘Omics’ big data handling.
How did we get here?
The RNA world hypothesis, a widely prevailing idea among molecular biologists, describes self-replicating RNA as the precursor to all modern life forms. Besides recognizing its primordial origin, we have come to realize that RNA is a highly complex and diverse macromolecule, which plays central roles in protein synthesis by serving as the encoder (mRNA), the decoder (tRNA), and the catalyst (rRNA). But our understanding of RNA diversity, form, and function continues to evolve as we keep discovering new and exciting classes of RNAs that carry out unexpected functions. Recent transcriptome studies, which capture nearly every transcript in a cell, have revealed an overwhelming number of RNA variants resulting from multiple transcription initiation sites, alternative pre-mRNA splicing and polyadenylation, post-transcriptional editing, and direct chemical modifications of RNA (Figure 1). These studies have demonstrated that the variants play key roles in generating mRNA diversity by altering their coding sequences, half-lives, and translation efficiencies, allowing complex organisms to control temporal and tissue-specific transcriptome patterns.

The systematic study of RNA variants began in the 1990s when cloning of expressed sequence tags revealed the diversity of mRNA populations expressed in different cells and tissues of metazoans\(^1\). The advent of microarrays in the early 2000s expedited the discovery of new RNA variants\(^2,3\), but the technology had several limitations, including (i) high background levels due to dependence on hybridization, (ii) reliance on existing sequence knowledge, and (iii) a limited range of detection. In the last decade, the field has taken a significant leap forward with the development of next-generation sequencing. The timeline of major milestones that facilitated the discovery of novel RNA variants is provided in Figure 2.

First-generation sequencing required separate reactions for individual nucleotides and was limited by a gel electrophoresis step, which was further complicated by the need to use radioactive reagents\(^4,5\). Next-generation sequencing, however, uses advanced optics and flow cells to identify nucleotide sequences rapidly, making the analysis and sequence determination quicker and more accurate\(^6–8\). A number of next-generation sequencing platforms have been developed, and the most popular ones are Illumina Solexa\(^9\) (relies on fluorescence imaging post-nucleotide addition every cycle, approximately 100 nucleotide reads, 0.1% error rate), Roche 454 sequencing\(^10\) (long reads and short homopolymeric stretches sequenced in single-cycle, approximately 500 nucleotide reads, 1% error rate).

**Figure 1.** Alternative splicing, circular RNA formation, and mRNA modifications are the major mechanisms generating much of the mRNA diversity. Alternative splicing results in the selective inclusion or skipping of particular exons as well as alternate 5’ and 3’ untranslated region (UTR) selection within a processed mRNA transcript. In some cases, alternative splicing also results in intron retention events. Circular RNA formation occurs because of back splicing of the 3’ end of a downstream exon to the 5’ end of an upstream exon. Lastly, mRNAs can be modified through editing mechanisms converting A and C nucleotides to I and U nucleotides, respectively, as well as through the addition of functional groups on either the base or the sugar moiety of a nucleotide.
error rate), Ion Torrent\textsuperscript{11} (based on detection of H\textsuperscript{+} ion release due to polymerization of dNTP probes, approximately 200 nucleotide reads, 1% error rate), and SOLiD sequencing\textsuperscript{12} (relies on sequencing by ligation over synthesis, 35 nucleotide reads, less than 0.1% error rate). Of the above methods, the first three are classified as sequencing by synthesis methods whereas SOLiD is under the sequencing-by-ligation category. The field of sequencing by ligation has made rapid strides with the development of complete genomics technology that uses combinatorial synthesis methods, involving the use of multiple probe-anchor combinations to determine sequence\textsuperscript{13}. Additionally, Pacific Biosciences has recently introduced single-molecule real-time sequencing\textsuperscript{14} of full-length transcripts, and the average read length is 2.5 kb (non-stop sequencing and real-time imaging, more than 1% error rates).

Perhaps the most significant breakthrough in the field was the development of massively parallel cDNA sequencing, commonly referred to as RNA-seq\textsuperscript{15}. It involves sequencing of cDNA fragments generated from total RNA after oligo-dT or random priming. For most transcriptome studies, mRNA is enriched from the total RNA pool by using either a poly(A) pull-down or selective depletion of rRNA. Following sequencing, the reads are aligned to a pre-annotated reference genome or assembled de novo to provide both the qualitative and the quantitative information of entire transcriptomes. This technique now can be used to identify new transcripts without having a priori knowledge of all gene models\textsuperscript{16}. Recent mapping tools can also identify splice junctions by using reads spanning across two exons, bolstering identification of alternative splice isoforms. Initially, RNA sequencing (RNA-seq) studies examined only the steady-state levels of RNA populations; however, the relative contribution of variable transcription (synthesis) and RNA degradation (stability) in those studies was unclear. New techniques such as Net-Seq\textsuperscript{17,18} and BRU-seq\textsuperscript{19} address this problem. Net-Seq employs DNA-RNA-RNA polymerase complex stability to isolate 3’ ends of nascent transcripts. Bru-Seq uses pulse-chase labeling of newly transcribed RNA with bromouridine to determine rates of RNA synthesis and stability.

The past five years have witnessed a surge of new techniques employing RNA-seq to identify many new RNA variants, and isolation/enrichment methods are based on each variant type. In parallel, significant advances in computing power have expanded our ability to annotate and quantify these variants with high confidence from a single cell to complex tissues, along with the advent of consortium projects that have given rise to databases for their comparative and meta-analyses\textsuperscript{20–22}. The increasing sensitivity of detection, wherein extremely minute quantities (picograms) of RNAs can be amplified and sequenced, has also spurred development of single-cell RNA-seq (scRNA-seq)\textsuperscript{23} for identifying heterogeneity within cell populations and uncovering transcriptome differences that may be hidden in bulk RNA-seq\textsuperscript{24}. Some of the commonly used sequencing technologies and bioinformatics packages for the detection of different RNA variants are provided in Table 1.

In this review, we discuss different types of mRNA variants commonly found in metazoans, consider the latest advances in techniques to identify them, and provide a brief overview of the recent developments in bioinformatics pipelines for their quantitative analyses.

### Decoding the splice variants

Pre-mRNA splicing is an essential step for the expression of most genes in metazoans. Splice sites within a pre-mRNA facilitate the splicing and ligation of introns and exons, respectively, producing a mature mRNA transcript. Alternative splicing involves differential usage of splice sites to yield multiple mRNAs from a single
The development of strand-specific and paired-end sequencing each splice junction provided a direct measure of splicing efficiency, cataloging the number of mapped reads against individual exons plus found to have alternatively spliced isoforms. Transforming our view of the extent and complexity of eukaryotic gene expression, efficiency, stability, and localization add the addition of all potential splice junctions to the reference genome. Clever computational approaches, however, overcame these limitations through the alignment of the alternative exon-exon junctions. Clever computational pipeline to accurately detect, visualize, and quantify complex local splice variants that are tissue-specific or change dynamically across different experimental conditions.

In addition to the discovery of variably spliced regions, RNA-seq has been an incredible asset for the investigation of splicing intermediates and novel splice products. For instance, Vaquero-Garcia et al., have recently developed a new computational pipeline to accurately detect, visualize, and quantify complex local splice variants that are tissue-specific or change dynamically across different experimental conditions.

For instance, Vaquero-Garcia et al., have recently developed a new computational pipeline to accurately detect, visualize, and quantify complex local splice variants that are tissue-specific or change dynamically across different experimental conditions.

The initial RNA-seq studies for detecting splicing patterns suffered from short read lengths, limiting the reliable genomic alignment of the alternative exon-exon junctions. Clever computational approaches, however, overcame these limitations through the addition of all potential splice junctions to the reference genome and accounting for them within the alignment mix. The results transformed our view of the extent and complexity of eukaryotic transcriptomes, as more than 95% of human multi-exon genes were found to have alternatively spliced isoforms. Importantly, calculating the number of mapped reads against individual exons plus each splice junction provided a direct measure of splicing efficiency that allowed precise quantification of isofrom ratios.

The development of strand-specific and paired-end sequencing led to improved mapping, the discovery of pervasive anti-sense transcription, and many novel splice variants in vertebrate species. Emerging technologies are geared toward sequencing the full-length transcripts (for example, Iso-seq) such that variably spliced regions from different areas of a single gene can be accurately connected to provide information about the entire mRNA isoforms (that is, from the 5’ end to the poly[A] tail). There has also been a strong push to improve the existing algorithms to capture the full complexity of splicing patterns within existing RNA-seq data. For instance, Vaquero-Garcia et al., have recently developed a new computational pipeline to accurately detect, visualize, and quantify complex local splice variants that are tissue-specific or change dynamically across different experimental conditions.

In addition to the discovery of variably spliced regions, RNA-seq has been an incredible asset for the investigation of splicing intermediates and novel splice products. For instance, it led to the identification of recursive splicing where long introns require specific exon contexts and are spliced out in phases rather than the usual two-step process. Recursive splicing uses cryptic sites within the intron, known as recursive splice (RS) site/ratchet points, which are surprisingly more conserved than the regular 5’ and 3’ splice sites. However, the identification of these RS sites required extensive deep sequencing with over a billion paired-end reads, without poly(A) selection in an attempt to capture pre-mRNA and nascent RNA transcripts. Likewise, RNA-seq has renewed interest in studying circular RNAs (circRNAs). The discovery of circRNAs, though known for over two decades, remained mostly serendipitous. Typical RNA-seq experiments fail to conserve RNA circularity because fragmentation and the absence of a...

<table>
<thead>
<tr>
<th>Variant analysis classification</th>
<th>Biochemical methods (methods for feature enrichment)</th>
<th>Bioinformatics tools</th>
<th>Pre-processing tools</th>
<th>Analysis tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative splice variants</td>
<td>a. RNAseq post Poly(A) enrichment for mRNA pulldown b. Isoform sequencing (Iso-Seq) and SLR-RNA-seq (synthetic long-read) for full-length transcript sequencing c. RNAseq enrichment for circular RNA</td>
<td>- Quality check: fastQC and HTSeq - Processing: Fastx, Trimmomatic, and cutadapt - Mapping: STAR, OLeGo, and TopHat</td>
<td>rMATS, MISO, IRCall, CIRCexplorer, MAJIQ, and DEXSeq</td>
<td></td>
</tr>
<tr>
<td>5’ and 3’ end variants</td>
<td>a. 3’ end enrichment libraries: 3P-Seq, PAPERCLIP, and 3’-Seq b. poly(A) tail length: TAIL-seq, PAL-seq, and LM-PAT c. 5’ end enrichment libraries: CAGE and deepCAGE</td>
<td>- Quality check: fastQC and HTSeq - Processing: Fastx, Trimmomatic, and cutadapt - Mapping: STAR, OLeGo, and TopHat</td>
<td>MISO and DaPars</td>
<td></td>
</tr>
<tr>
<td>RNA modification variants</td>
<td>a. antibody-based IP: m6A, m1A, m5C, and hm5C and b. Selective RNA chemistry: m5C and Ψ</td>
<td>- Processing: flexbar and Fastx - Mapping: STAR, OLeGo, and TopHat</td>
<td>MACS, HOMER, CIMS, and CITS</td>
<td></td>
</tr>
</tbody>
</table>
poly(A) tail or free 5′ or 3′ end cause de-enrichment of circRNAs during most library preparations. However, the lack of exposed 5′ or 3′ ends also stabilizes the circRNAs, making them resistant to exonuclease-mediated degradation. The recent development of biochemical methods to enrich for circRNAs (for example, RNaseR treatment to digest single-stranded RNA) along with new computational strategies (CIRCexplorerR) to mine the existing Encyclopedia of DNA Elements (ENCODE) Ribozero RNA-seq data has aided their de novo detection in humans, mice, and Caenorhabditis elegans. The results have revealed that most circRNAs are non-coding splice variants containing multiple exons located in the middle of genes that are often expressed in a tissue- and developmental stage-specific manner and that these variants may act as sponges to antagonize microRNA activity.

Cracking the terminal variants

Whereas the gene body variants in mRNAs promote proteome diversity, variability at the terminal ends is known to serve more of a regulatory role in protein synthesis. This realization has led to a strong surge in the development of new methodologies to identify all the 5′ and 3′ ends variants within eukaryotic transcriptomes. For instance, recent methods have revealed that alternative transcription start sites (TSSs) are prevalent in the human genome and that approximately 30 to 50% of human genes have more than one promoter. A substantial number of these alternative TSSs alter the 5′ UTRs, in turn modulating the translational output of mRNA. 5′ UTRs are known to comprise many regulatory features—5′ cap structure, internal ribosome entry sites, G-quadruplexes, and so on—that can drastically affect the ability of ribosomes to assemble and initiate translation. Apart from forcing the gain or loss of such regulatory features, alternative TSSs generate RNA variants with different upstream open reading frames (uORFs). Ribosome footprinting studies have uncovered a significant fraction of such short translated uORFs within 5′ UTRs that potently repress the translation of downstream ORFs.

Apart from providing proper signals for mRNA 3′ end formation and poly(A) site selection, the 3′ UTR is considered to be a regulatory hub for gene regulation. Proper 3′ end formation is essential for nuclear export and stability of mature mRNAs and their efficient translation. Analysis of existing RNA-seq data with new computational pipelines (MISO and DaPars) or sequencing of the libraries enriched for 3′ ends of mRNAs (3P-Seq, 3′-seq, 3′READS, and PAPERCLIP) have documented that over 50% of human mRNAs express variable 3′ UTRs—generated by alternative cleavage, splicing, and polyadenylation. Importantly, these 3′ UTR variants are differentially expressed during differentiation, development, and disease, and they often influence the accessibility of microRNAs and RNA-binding proteins (RBPs) to the 3′ UTR cis-regulatory elements affecting mRNA stability, localization, and translation efficiency.

Poly(A) tail length is another important 3′ UTR feature that impacts the post-transcriptional fate of mRNAs. Although some biochemical methods (LMPAT and ePAT) can accurately measure the poly(A) tail lengths of individual genes, precise measurements at a genome-wide scale are hampered by difficulties in sequencing homopolymeric sequences longer than 30 nucleotides. However, new advances in poly(A) tail length profiling methods (TAIL-seq and PAL-seq) use specialized 3′ adapters, RNase T1 treatment (digests the mRNA body but not the poly(A) tail), and the spike in controls in library preparation. These biochemical improvements in conjunction with refined machine-learning algorithms to analyze the intensity of raw fluorescence signals allow the precise estimation of poly(A) tail lengths. These new methodologies have started to provide the first high-resolution, global views of poly(A) tail length variants and their relevance to mRNA stability and translational efficiency in cells and tissues from a variety of species, including yeast, Drosophila, zebrafish, Xenopus, mouse, and human.

Marking the variants

Though originally discovered in highly abundant rRNAs, tRNAs, and snRNAs, newly tailored mRNA sequencing methods have revealed that base modifications and editing sites are also highly prevalent in mRNAs. The study of mRNA modifications (methylation, hydroxy-methylation, and pseudouridylation) plus editing (A–I or C–U) is now commonly referred to as “epitranscriptomics”. Much like the epigenome, the epitranscriptome expands the information content of nucleic acids, rendering them additional structural and functional flexibility. Recent advances have led to the profiling of millions of modification sites for each of the known mRNA modifications, including methylation, as well as base editing (such as A–I or C–U) and editing (such as A→I or C→U). Whereas most of these have been observed across all genes, some are limited to specific classes of transcripts.

Remarkably, the genome-wide mapping of modifications and editing sites has not only expanded the landscape of this new class of mRNA variants within eukaryotic transcriptomes but also unraveled their dynamics in abundance and site selection in response to intra- and extra-cellular stimuli. The impact of these post-transcriptional marks is now recognized in almost every step of gene regulation, including mRNA processing, stability, and translation. An interesting example is the recent work on ADAR (adenosine deaminase acting on RNA) enzymes, showing their role in protecting double-stranded RNA (dsRNA) against recognition by dsRNA response pathways. Importantly, the writers, erasers, and reader proteins for some of these modifications have been identified, and the phenotypic and molecular consequences of their gain or loss of function have been demonstrated.

Today, the transcriptome-wide mapping of modified nucleosides relies on either antibody-based immunoprecipitation (m6A, m1A, m5C, and hm5C) or a modification-selective RNA chemistry approach (m5C or Ψ) to enrich for the target modification during library preparation. However, further optimization of these methods is needed to simultaneously obtain single-nucleotide resolution and stoichiometry of RNA modifications, as fractional modification represents another mechanism to generate functional diversity within individual mRNA pools. Current epitranscriptome profiling techniques accurately identify the modification sites within RNA transcripts but fail to provide information about the modified fraction for each site. Given that a modification may alter the mRNA secondary structure or binding site for RBPs, quantitative fractional information about each modified site in the transcriptome is highly desired.
Translating the variants

Early dogma predicted that the amount of mRNA for a given gene directly correlated with the amount of protein made. However, we now know that through sequence variations discussed above and along with other RNA control elements, the translation of specific transcripts can be fine-tuned for the purpose of responding to cellular needs in a variety of contexts41. Therefore, estimating average gene-level translation by ribosome profiling—a technique that uses short, approximately 28 nucleotide ribosome footprints on mRNA as a means of quantifying protein output—may not provide an accurate measure of the translational potential for individual RNA variants or isoforms. This is especially true for splice variants, which can exhibit variable coding capacities because of (i) sequence differences, (ii) altered mRNA half-lives, or (iii) differential nucleo-cytoplasmic export. Indeed, several recent efforts to measure isoform-specific translation have revealed discrete effects of 5′ end42, 3′ end43, and coding sequence55,58 diversity.

Using translation-associated transcript leader sequencing (TATL-seq), Arribere and Gilbert successfully mapped transcript leader boundaries across the yeast genome and discovered hundreds of new alternative transcript leader variants, the majority having differential translation efficiencies46. Complementing this work, Spies et al. acquired decay and translation rates for proximal and distal 3′ UTR variants using poly(A)-primed sequencing (2P-seq) but found only a modest regulatory influence for alternative 3′ UTR sequences68. By comparing isoform ratios in cytoplasmic and polyribosomal extracts from human cells (Frac-seq), Sterne-Weiler et al. observed that not all splice isoforms are similarly loaded into polyribosomes and that the presence of microRNA binding sites or premature termination codons can greatly affect the loading of individual isoforms into polyribosomes69. Similarly, a major impact of mRNA isoforms on protein output was observed by Floor and Doudna, who recently adapted the classic approach of polysome profiling to measure isoform-specific translation (TrIP-seq)47.

Where are we headed?

Emerging technologies have given us a powerful toolbox to detect and quantify RNA variants from simple unicellular organisms to complex vertebrates. These new techniques allow for previously unimaginined access to compare transcriptome patterns across species, tissues, developmental stages, and disease states. We now can isolate and sequence picogram amounts of RNA from fresh and formalin-fixed paraffin-embedded tissue specimens and even identify variants within single cells that are isolated from a wide range of biological samples. Moreover, the ENCODE project has generated unprecedented information about the many different RNA variants present in model organisms and human cells. Systematic discovery of cis-acting RNA elements, trans-acting RBPs that bind to these elements to control the production of variants, and their expression and subcellular localization is paving the way to uncovering the underlying code for generating much of the RNA diversity.47,55,58 If sequencing costs continue to decline as anticipated, clinical measurements of normal and aberrant RNA variants in body fluids or tissue biopsies in the not-so-distant future may become part of personalized medicine, making for more accurate diagnoses and offering better-informed treatment plans. However, as with most opportunities, certain challenges must be overcome to realize their full potential. What are the common challenges facing global profiling of RNA variants?

The authors declare that they have no competing interests.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We thank Sayeepriyadarshini Anakk and the members of the AK lab for constructive criticism and suggestions.
References


Open Peer Review

Current Referee Status: ✔ ✔ ✔

Editorial Note on the Review Process
F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty and are edited as a service to readers. In order to make these reviews as comprehensive and accessible as possible, the referees provide input before publication and only the final, revised version is published. The referees who approved the final version are listed with their names and affiliations but without their reports on earlier versions (any comments will already have been addressed in the published version).

The referees who approved this article are:

Version 1

1 Gordon Carmichael, Department of Genetics and Developmental Biology, University of Connecticut Health Center, Farmington, CT, USA
   Competing Interests: No competing interests were disclosed.

2 Emanuele Buratti, Molecular Pathology, International Center for Genetic Engineering and Biotechnology, University of Trieste, Trieste, Italy
   Competing Interests: No competing interests were disclosed.

3 Graziano Pesole, Dipartimento Di Bioscienze, Biotecnologie e Biofarmaceutica, Università degli Studi di Bari, Bari, Italy
   Competing Interests: No competing interests were disclosed.