Myriad RNAs and RNA-Binding Proteins Control Cell Functions, Explain Diseases, and Guide New Therapies

BYUNG RAN SO AND GIDEON DREYFUSS

Department of Biochemistry and Biophysics, Howard Hughes Medical Institute, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

Correspondence: gdreyfuss@hhmi.upenn.edu

This summary of the 84th Cold Spring Harbor Laboratory (CSHL) Symposium on Quantitative Biology: RNA Control and Regulation, held in May 2019, highlights key emerging themes in this field, which now impacts nearly every aspect of biology and medicine. Recent discoveries accelerated by technological developments reveal enormous diversity of RNAs and RNA-binding proteins (RBPs) with ever-increasing roles in eukaryotes. Atomic structures and live-cell imaging of transcription, RNA splicing, 3′-end processing, modifications, and degradation machineries provide mechanistic insights, explaining hundreds of diseases caused by their perturbations. This great progress uncovered numerous targets for therapies, some of which have already been successfully exploited, and many opportunities for pharmacological intervention and RNA-guided genome engineering. Myriad unexplained RNAs and RBPs leave the RNA field open for many more exciting discoveries.

Foundational concepts of molecular biology established from studies of prokaryotes—one gene, one mRNA, one protein; the regulation of mRNA synthesis almost entirely by transcription; and cotranscriptional mRNA translation—could be captured in a single electron micrograph and summed up in a simple formula: DNA → mRNA → protein. The now-historical proceedings of CSHL Symposia from the 1950s and 1960s chronicle these insights. However, they could not explain or predict how mRNAs are made and function in eukaryotes, where primary protein-coding gene transcripts (pre-mRNAs, historically called hnRNAs) are not translation-ready mRNAs due to open reading frame (ORF)-disrupting introns, and the translation machinery (mRNA-decoding ribosomes and tRNAs) is in the cytoplasm, segregated from transcription sites on chromatin in the nucleus. Seminal discoveries in the 1970s and 1980s on transcription, splicing of protein-coding pre-mRNAs (to excise introns and join exons), self-splicing RNA catalysis, noncoding (nc) small nuclear RNPs (snRNPs), RNA-binding proteins (RBPs), 3′-end cleavage and polyadenylation, and mRNA translation regulation, transport, and surveillance uncovered an astonishing complexity of post-transcriptional gene regulation in eukaryotes.

The 84th CSHL Symposium showcased remarkable progress in understanding these processes, additional layers of mRNA regulation, numerous ncRNAs, and new RNA functions. Video recordings of 56 30-min oral presentations and discussions, chapters in this treatise, and hundreds of poster presentations provide a complete account of the five-day symposium. A traditional all-inclusive comprehensive summary would be superfluous. Instead, here we highlight some key themes, outstanding questions, and prospects. Presenters’ names are noted in parentheses in relevant contexts as starting points for deeper exploration of specific topics. The discussion does not necessarily reflect the presenters’ opinions or the main points they made, nor does their mention imply more significant contributions to the topic compared to other researchers in that area.

PRE-mRNA PROCESSING

The main pre-mRNA processing reactions are splicing and 3′-end processing by cleavage and polyadenylation (CPA). Decades of intensive research identified the RNA motifs (sequences and structural elements) and their cognate factors, the inventory of components that mediate these reactions, the chemistry of RNA intermediates, and the stepwise assembly of catalytic complexes. Splicing of each intron occurs in a large complex (spliceosome) that assembles through an elaborate series of binding and rearrangements of five snRNPs (snRNA–protein complexes, generally U1, U2, U4, U5, and U6; JA Steitz) and more than 50 proteins. The key signals, 5′ and 3′ splice sites (ss), are recognized by U1 base pairing and U2 associated proteins, respectively, which commence the assembly process. Various enzymes, particularly RNA helices, use ATP hydrolysis to effect conformational and compositional transitions necessary for splicing self-catalysis. CPA is specified by polyadenylation signals (PASs), consisting of a hexanucleotide sequence (AAUAAA and variants thereof) flanked by upstream and downstream motifs that recruit three main subcomplexes, comprising 20 proteins. The 3′-end cleavage and poly(A) addition are carried out by protein enzymes.

Progress in understanding these pre-mRNA processing reactions—and other aspects of RNA biology—has been
greatly facilitated by technological developments. Extensive biochemical and genetic studies, including reconstitutions with RNA substrates in vitro, purifications from cells, and the use of mutations to arrest reaction intermediates, exemplify discovery schemes used in the field. In addition, crystal structures of many individual constituents and subunits have been determined by X-ray crystallography. Yet, with few exceptions (e.g., ribosomes), the inability to obtain high-quality diffracting crystals of entire complexes left essential information obscure. Single-particle cryo-electron microscopy (cryo-EM), which does not require crystalization, produced in rapid succession a profusion of atomic or near-atomic resolution three-dimensional views of molecular machines that transcribe, process, degrade, and regulate RNAs. This includes spliceosomes at various assembly stages (K Nagai, presented by M Wilkinson), CPA complexes (LA Passmore), exosomes (E Conti), and miRNA and other complexes (JA Steitz, BL Bass).

**COTRANSCRIPTIONAL RNA PROCESSING**

Despite their intracacies, spliceosomes and CPA complexes are generic machines that are highly conserved in eukaryotes. They attach cotranscriptionally, aided by interactions with the carboxy-terminal domain (CTD) of RNA polymerase II (Pol II) (XD Fu, KM Neugebauer, A Aguilera). This enhances processing efficiency and shortens the lag time (compared to uncoupled in vitro), confers 5′ to 3′ polarity with respect to signal usage (first come, first served), and introduces kinetic effects through the interplay between Pol II speed and signal strength (A Kornblihtt). This coupling has many important consequences. For example, by default, cotranscriptional processing makes CPA most likely at the TSS-proximal PAS. In thousands of genes, cryptic PASs in the first or second introns (which are typically the longest in humans) can elicit transcription-terminating premature CPA (PCPA). This is generally suppressed by U1 snRNP binding upstream nearby (e.g., at the first 5′ss [which also initiates splicing]). This U1 activity (telescripting) is necessary for full-length transcription and used for regulating PAS usage throughout pre-mRNAs or other nascent transcripts that can be modulated by U1 snRNP availability (G Dreyfuss). A lower ratio of U1 snRNP binding sites to PASs in upstream antisense transcripts in many protein-coding genes (which are typically bidirectional) makes them susceptible to PCPA, which enhances sense transcription directionality (PA Sharp).

Without complicating circumstances such as long introns, generic machineries would be sufficient to execute constitutive splicing and CPA. Splicing has some benefits—for example, for eliminating nonsense mutations (LE Maquat)—although it is debatable if the benefits outweigh the costs (intronless genes are perfectly functional). Still, constitutive processing limits the number of mRNAs and proteins a gene can theoretically make to one, regardless of how many introns and PASs it has. However, this is not the case in complex organisms. Their existence depends on making multiple mRNAs and protein isoforms from the same gene. This mRNA diversity is generated primarily by alternative splicing and alternative CPA of pre-mRNAs. It is a compelling rationale for introns and splicing because it obviates the need to increase the number of genes. For example, humans have approximately 20,000 genes, but produce many-fold greater number of protein isoforms, which is necessary for making many different cell types.

**RNA-BINDING PROTEINS (RBPs)**

Alternative pre-mRNA processing is controlled by a very large assortment of RBPs. There are around 1000 genes in humans. RBPs are part of and play a role in all steps in the life of mRNAs and in every aspect of RNA biology; all RNAs in cells exist as complexes with RBPs (RNPs). Numerous RBPs associate with pre-mRNAs cotranscriptionally (historically called hnRNP proteins), which helps prevent long pre-mRNAs from misfolding and sculpt them for processing. RBPs regulate alternative RNA processing by binding to splicing enhancers and suppressor motifs and thereby promote or hinder the assembly of spliceosomes and CPA factors (COPAFs) at nearby splice sites and PASs. Changes in RBP and snRNP repertoires and activities regulate cell- and development-stage-specific alternative processing and cells’ responses to environmental changes (K Lynch).

Many of the most abundant prototypical RBPs, the hnRNP proteins, and numerous RBPs with the same domain structure and features that go by other nomenclature are removed with introns, as are snRNPs. However, many others remain associated with mRNAs after splicing, during export to the cytoplasm, where they have additional roles in translation, transport to specific locations, and mRNA stability. They are generally dislodged during the first round of translation and shuttle back to the nucleus (except those on 3′ UTRs). Additional RBPs are added during splicing, thereby marking spliced junctions (exon junction complex), providing positional information that helps detect nonsense codons (LE Maquat). Thus, RBPs link the steps from transcription on chromatin sites to translation on ribosomes and the cell extremities. In another technological feat, single-molecule imaging vividly traces the journey of actin mRNA guided by the RBP ZBP1 to fibroblasts’ leading edges, where they translate (RH Singer). The mRNP complexes that guide other long-distance mRNA transport and localized translation, which are crucial in neurons and early development (RB Darrell), are also assembled during processing in the nucleus. Additional RBPs associate with mRNAs in the cytoplasm, such as the poly(A) binding protein, which has roles in translation and mRNA turnover.

**GRANULES AND CONDENSATES: THE POWER OF RNP SELF-ASSEMBLY**

Hallmark features of RBPs explain their wide range of functions and their roles organizing granular micro-compartments that function in RNA metabolism. Early se-
quencing described RBPs modular structures, multi-
RNA-binding domains (typically two or more RBD/RRMs, KH domains, etc.), and “auxiliary domains” noted for their simple amino acid composition (now referred to as low complexity [LCDs]). Auxiliary domains harbor transport signals and LCD repeats (such as RGGs, pro-
lome-rich regions) that can enhance RNA binding and in-
teractions with other RBPs. Similar LCDs are found in hundreds of RBPs and transcription factors (many RBPs have dual functions in transcription). LCDs are relatively disordered and high IDR index and have a high propensity to engage in homotypic, albeit weak, interactions, which makes them form gels at high concentration (S McKnight). RNA lowers the RBP concentration of gelling (also known as phase separation), which high concentra-
tions of RNAs contribute to by RNA–RNA interactions (R Parker, A Gladfelter). Almost any RNA over a few hundred nucleotides that has sufficient single-stranded re-
gions can do this, because RBPs have a certain generic “nonspecific” RNA binding (counter to the misconception that specificity is binary). RNA binding makes LCDs multivalent, which potentiates weak interactions (e.g., NEAT1 RNA [T Hirose]). The result is a massive coalescence of RNAs with many RBPs bound to them. They have a granular appearance that has fascinated mi-
icroscopists for some time and now have biophysical un-
derpinning, and their significance in many molecular processes is becoming better understood. The forces of self-assembly can give rise to large, readily visible gran-
ules in both the nucleus (e.g., Cajal bodies, gems, speck-
les/interchromatin granules) and cytoplasm (e.g., stress granules). Granules that serve as storage depots of maternal RNP components play important roles in early de-
velopment (R Lehmann, G Seydoux). The tendency to describe granules as membraneless organelles is some-
what questionable as they can be highly dynamic, and there is a lack of evidence that specific functions can only occur inside of them (as opposed to the idea that they represent a high local concentration of components that can function inside or outside of them).

The same RNP-driven self-assembly principles operate on smaller scales to concentrate transcription factors at super-
enhancers (PA Sharp, II Cisse, RA Young), chromatin silencer HP1α to H3K9 methylated chromatin (GJ Narlikar), and mRNAs encoding secretory proteins to the cytoplasmic face of the endoplasmic reticulum (C Mayr).

NONCODING RNAs

Classes of RNAs that do not encode proteins nor are
directly involved in translation and instead have defined functions in other processes have been well established. For example, snRNAs (discussed above), snoRNAs (guide RNA modifications/rRNA processing), miRNAs (CS Pikaard), tRNAs (O Rando), transposable elements (M-E Torres-Padilla), and piRNAs (transposable element sup-
pression; J Brennecke, L Joshua-Tor) illustrate the range of roles RNAs have in cell regulation. miRNAs have roles in regulating mRNA translation and stability (D Bartell, JA Steitz, RA Martienssen). A more nebulous grouping, called long noncoding RNAs (lncRNAs), is based on a somewhat arbitrary size cutoff (>250 nt) and absence of conventional translation ORF. By this definition, there are tens of thousands of lncRNAs. A small fraction of those have spe-
cific functions. For example, Xist is essential for X-chromo-
some silencing (J Lee, N Brockdorff, HY Chang, A Akhtar). Some other lncRNAs have clear and important biological effects (C Dean, J Mendell, J Wysocka)—such as MALAT1, which is highly expressed in and contributes to cancers—and are valuable prognostic markers (DL Spect-
tor). However, presently most lncRNAs have no specific functions ascribed to them. The expectation of specific func-
tions, based on proximity to genes of interest, significant expression level correlation with certain conditions, and other considerations, motivates extensive research on indi-
vidual lncRNAs. On the other hand, the lack of significant sequence conservation among lncRNAs in the same genomic locations in closely related mammals and their very short half-lives (I Ulitsky, NJ Proudfoot) suggest that they may not each have specific or unique functions.

PERVASIVE TRANSCRIPTION, RNA PROCESSING, AND MODIFICATIONS

In our view, the enormous complexity of RNAs that mamalian cells make is puzzling. Deep RNA-sequenc-
ing (RNA-seq) shows a much greater diversity of RNAs generated at every step than had been anticipated or can be explained in the foundational framework of molecular biology. Many more will undoubtedly be discovered from single-cell RNA-seq. Pol II transcription is highly pervasive, initiating bidirectional transcription from nu-
merous promoters throughout the genome, unless access to them is blocked. Many alternative TSSs are used in the same genes. Transcription can continue for enormous dis-
tances, requiring barriers, checkpoints, and other impedi-
ments (as mentioned above and SM Gasser, S Grewal) to slow it down and RNA cleavage to effect termination (K Adelman, JE Wilusz, NJ Proudfoot). There are multiple alternative PASs and other 3’-end processing signals, result-
ning in numerous RNAs from the same gene and a variety of 3’ ends and tails. Annotation-independent splice junction calling identified many more splicing events than had been anticipated based on what is represented in steady state mRNAs, now in the hundreds of thousands. There is splicing in lincRNAs, recursive splicing in long introns, resplicing in mRNAs, and back splicing, which makes circular RNAs (circRNAs) (L Chen). Enzymatic editing (B Bass) and modifications (C He, S Jaffrey) alter the chemistry of nucleotides. Many of these regulate some aspect of RNA function, but many others may just be bystanders. One can speculate why such complexity arises and what purpose it might have (as G Dreyfuss had at this symposium’s conclusion), but it is exciting that so much remains to be deciphered.

MANY DISEASES AND NEW THERAPIES

The large number of components and processes that make and regulate RNAs increase opportunities for some-
thing to go wrong. Mutations in RNA processing signals, mRNA untranslated regions, miRNAs, snRNA, RBPs, and RNA processing factors cause minimally hundreds of human diseases. For example, a synonymous mutation in a splicing regulating sequence in \textit{SMN} causes frequent exon 7 skipping, resulting in a SMN deficiency that causes spinal muscular atrophy (SMA). Mutations in splicing factors, such as U2 associated SF3B1, cause hematologic malignancies and other cancers (O Abdel-Wahab, R Bradley), and expanded G-rich repeats in a C9orf72 intron, which sequesters hnRNPF thereby decreasing its availability and impairing splicing, are linked to amyotrophic lateral sclerosis (ALS) (J Manley). Mutations in hnRNPA1 and similar RBPs underlie many neurodegenerative diseases.

Such advances in understanding pathogenic mechanisms have helped develop diagnostic tools and have revealed many targets and approaches for repairing, replacing, or removing them as potential therapies. Remarkable successes have already been achieved—for example, using antisense oligonucleotides to correct SMN splicing (to enhance exon 7 inclusion) for SMA (A Krainer). Improvements in RNA-guided CRISPR–Cas technologies, originally discovered from studies on adaptive immunity in bacteria, enable genome and transcriptome engineering with numerous research and clinical applications (J Doudna and F Zhang). Having atomic 3D maps of RNP complexes reveal numerous potential targets for small molecules and the ability to monitor RNA structure changes (K Weeks) holds promise for pharmacological intervention.

\textbf{CONCLUSION}

Great advances in the RNA field now impact nearly every aspect of biology and medicine. Technological developments, including in RNA-seq and data analysis, mass spectrometry, structure determination by cryo-EM, high-resolution imaging in cells, and genome editing, have greatly accelerated the pace of discovery. The wealth of new information has provided insights, inspired useful applications, and uncovered puzzling complexities for future research.