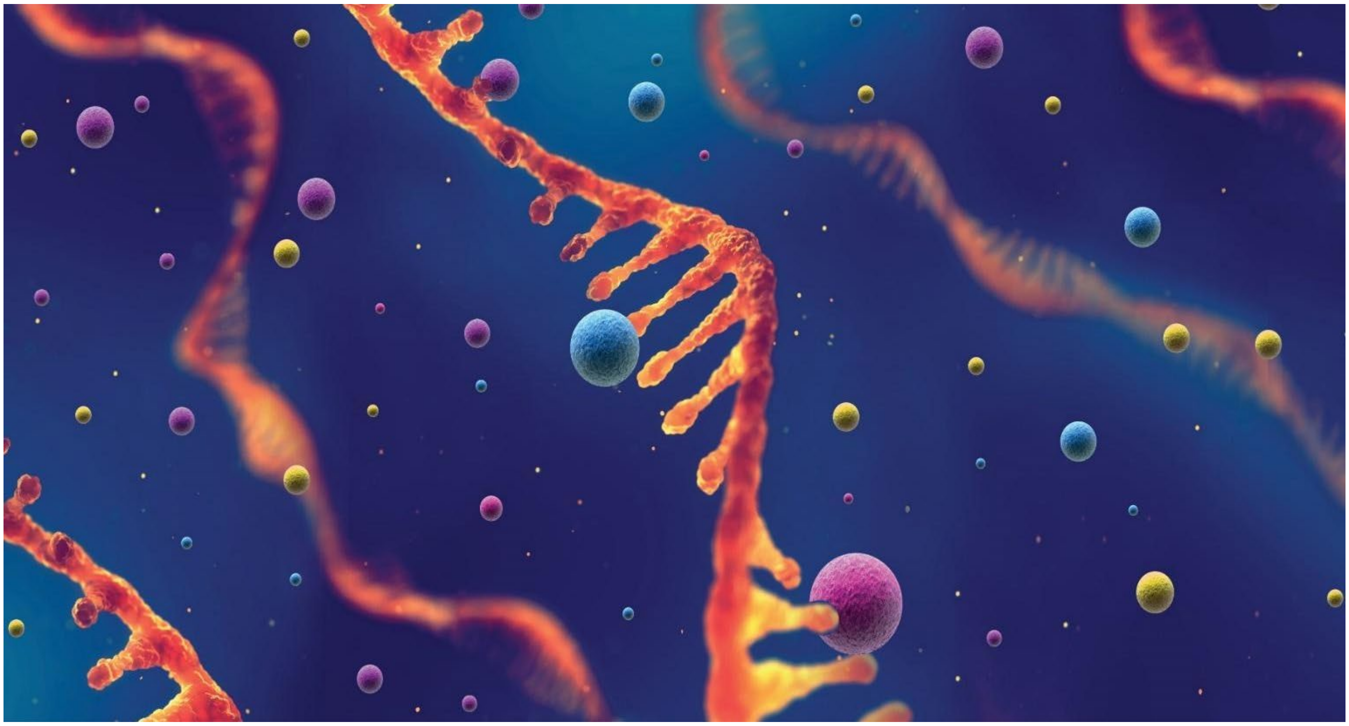

# Capturing RNA Sequence and Transcript Diversity, From Technology Innovation to Clinical Application

## May 24 – May 26, 2022

This report was developed by Avanti Corporation in collaboration with NIEHS and NHGRI meeting organizers.

National Institutes of Health • U.S. Department of Human Health and Services

## Table of Contents

## Workshop Overview

The National Institute of Environmental Health Sciences (NIEHS) and the National Human Genome Research Institute (NHGRI) hosted a virtual workshop titled ‘Capturing RNA Sequence and Transcript Diversity, From Technology Innovation to Clinical Application’ from May 24-26, 2022. The goal of the workshop was to determine the current capabilities, needs, and prospects for comprehensive characterization and understanding of the true diversity of all RNAs and their modifications at a chemical and structural level in relation to normal and disease states.

The RNA workshop had a number of critical objectives: 1) to explore the need for “RNomics”; 2) to identify technologies needed to comprehensively characterize RNA; 3) to determine what infrastructure, bioinformatics, and other resources are needed to support technology development, utilization, and validation; 4) to consider the steps needed to facilitate rapid adoption of these technologies by the wider scientific community; and 5) to identify how to best incorporate public outreach and workforce development in this area.

The three-day workshop was divided into the following six scientific sessions:

- Setting the Stage
- Impact on RNA structures and Biological Roles
- RNomics Applications from Research to Clinic
- Technologies for Direct Sequencing
- Infrastructure, Bioinformatics, and Critical Resources: What is Needed?
- Facilitating Technology Dissemination and Adoption

At the beginning of each session, a speaker provided a broad overview of the session topic. The keynote (Session 1) and plenary presentations (Sessions 2-6) were each followed by a moderated discussion. In Sessions 2-6, participants in concurrent breakout rooms discussed key questions related to the session theme. These breakouts provided an opportunity for smaller group discussions on more focused subject areas, guided by moderators and key participants.

This report provides an overview of the meeting presentations and key take-away messages. Please note that this report uses technical terminology. A glossary of relevant terms, all session recordings, and the workshop agenda can be found on the meeting web site. An executive summary is provided as part of the September 19, 2022, National Advisory Council for Human Genome Research Open Session Agenda.

## Opening Remarks

NHGRI Director Eric Green, M.D., Ph.D., welcomed all participants and reiterated the strategic vision of the NHGRI to create and maintain a robust foundation for genomics. He described some of NHGRI’s RNA-specific interests which include direct RNA sequencing, modified RNA bases, transcript isoforms, synthetic RNA with modifications, and multi-omics approaches to disease and risk. He referenced the NHGRI Technology Development Program that supports technology transfer, novel approaches to achieve orders-of-magnitude improvements, and refinement of current methods. The goal of the NHGRI Technology Development Program is to increase efficiency while decreasing cost and maintaining data quality through more robust, low-cost sequence determination and genotyping, along with epigenetic, functional, and synthetic genomic studies. Dr. Green also remarked on the need to further build NHGRI assets in RNA genomic analysis and for robust national and international partnerships to advance research in RNA biology.

NIEHS Director Rick Woychik, Ph.D., welcomed all participants and set the stage for the workshop. He reminded attendees of how much simpler the central dogma proposed in 1950 was compared to the complexity of RNAs known today. Literature shows an increasing role for RNA in cells. Of particular interest to NIEHS, research demonstrates that some environmental exposures are impacting human health and phenotypes through the expression and post-transcriptional modification of different forms of RNA. Dr. Woychik stated that one of the big challenges and focuses of the workshop is to better define the types of RNA modifications and their resulting biological significance. He also reiterated the needs and challenges in the field of RNomics.

The Director of the Division of Genome Sciences at NHGRI, Carolyn Hutter, Ph.D., framed the goal of the workshop as a focus on capabilities, needs, and prospects for a broader and more comprehensive understanding of RNA. She defined the objectives of the workshop and the opportunities in the field RNA biology. Dr. Hutter requested that workshop participants freely share pertinent knowledge, discuss challenges, needs, and opportunities, address the current state-of-the-science, and identify potential areas of significance, innovation, and feasibility.

## Scientific Session 1: Setting the Stage

**Moderator:** Vivian Cheung, M.D.

Anna Marie Pyle, Ph.D., set the stage for the workshop with a presentation on the current state of the field of RNA science. She stressed the importance of studying the diversity of RNA sequences and modifications to advance the understanding of the role of RNA in health and disease. She provided an overview of the biological and technical challenges in the field, new strategies being developed, important unmet needs, and potential future directions.

Dr. Pyle began her talk by identifying biological challenges in the field. She highlighted that most RNAs are long, fragile, and of low abundance. Sequence data alone captures only a portion of the functionally relevant RNA information. This complexity is due to extensive alternative splicing of RNA transcripts into isoforms and the complicated secondary and tertiary structures of RNA. Pyle stated that there is only limited knowledge of the chemical reactions that modify the ribonucleobases and the sugar-phosphate backbone of RNA; furthermore, not enough is known about RNA-protein interactions. Additionally, RNA is highly modified by chemical reactions and thus its sequence differs across different cells even within an organism and during the life cycle of the RNA where the RNA is regulated by sensitive decay and surveillance pathways.

Studies of RNA are made difficult by both the complexity of RNA and by major technological limitations to current approaches. Currently, a majority of RNA sequencing is performed with cDNA-based short reads that are mapped to a DNA reference genome, thus we still cannot obtain the exact RNA sequence. As a result, we do not yet know how the RNA sequence and transcript abundance change in specific cellular responses and across cell types. Single-cell methods to study RNA biology presently do not yield

adequate information given the low abundance of many RNA transcripts and the fact that most single-cell methods only sequence the very 3’ end of transcripts. Identifying RNA modifications transcriptome-wide and with known stoichiometry is also not currently possible. The enzymes (e.g., RNA ligases, reverse transcriptases, RNA modification enzymes, etc.) used in RNA sequencing are inefficient. Enzymological research is needed to develop better molecular tools using basic mechanistic and evolutionary biology approaches. There is also a need for a platform to broadly communicate enzyme capabilities and efficiencies in different RNA analysis applications. Finally, data interpretation and management, machine-learning, and other computational methods to analyze RNA sequences and functions are still in their infancy.

There are a variety of new approaches in RNA sequencing, and these approaches should continue to be developed in parallel. Currently, direct RNA sequencing methods such as nanopore technologies have a high error rate, require high read depth, and are often not at single nucleotide resolution, but these methods are improving rapidly. Long reads of RNAs can be made with low error rates from cDNA synthesis with ultraprocessive reverse transcriptases, but this approach also has complementary shortcomings. Continued innovation in existing and novel single-cell RNA-seq[1] approaches and analyses will allow these techniques to become increasingly powerful and will be important for identifying rare cell populations and lineage trajectories, as well as addressing salient biological questions such as cellular responses to toxins. Other areas needing further development include methods to capture RNA modifications in cDNAs, methods to generate long cDNAs with mutational profiles (“fingerprints”) as well as training sets to understand their signatures, and software programs that handle RNA folding and accurately predict RNA structure and modifications.

The enormous potential of clinical applications of RNA research is only beginning to be realized. RNA-seq is expensive and there is little financial incentive for vendors to improve tool efficiency or quality. Thus, the innovation is only happening in a few well-funded environments. There are also too few effective incentives for partnerships between industry, government, and academia.

## Keynote Discussion Highlights

- Emerging strategies for direct and indirect long-read RNA sequencing need further development.
- Standards are needed for many aspects of RNA modification identification. There are technical developments needed in terms of hardware, software, and enzymes. A training set of RNAs with known modifications would be a useful community effort that would allow further development of many aspects of RNA biology research.
- Due to the rapid progression of technology, there is too little incentive to further develop better enzymes used to generate RNA sequencing libraries (e.g., reverse transcriptases, ligases, etc.). Facilitating enzyme tool development, characterization, and testing as a community effort could be of great assistance in developing necessary capabilities.
- New technologies are needed to introduce RNA modifications *in vitro*.

[1] For the purposes of this report, RNA-seq is defined as the conversion of RNA to DNA (cDNA) through reverse transcription before subsequent cDNA sequencing, and thus does not include direct RNA sequencing.

- A tissue and cell line bank developed to enable testing of new modification and sequencing methods would be ideal.
- Functionally relevant RNA information was discussed in the meeting and includes the m7G cap, poly-A site and tail length, each modification (in sequence context), exons, RNA structure, and RNA-binding protein (RBP) binding sites on a single-molecule scale.

## Scientific Session 2: Impact of Nucleotide Sequence on RNA Structure and Biological Roles

**Moderator:** Traci Hall, Ph.D.

Yunsun Nam, Ph.D., described how an RNA's sequence affects its structure and biological roles and highlighted how sequence and modification information is needed to determine RNA structure and function. The talk focused on the aspects of RNA diversity that impact function, starting with how base modifications - and backbone structure influence how an RNA is folded. These aspects affect nucleic acid and protein interactions and localization. To date, approximately 150 chemical modifications of RNA bases have been identified. These modifications impact how the RNA sequence affects structure. Further, quantitative biophysical and biochemical approaches have shown that protein-RNA interactions contribute to RNA structure, an important factor for deducing biological roles. Dr. Nam gave two examples that highlighted the impact of modifications on RNA: 1) m6A can destabilize RNA transcripts, and 2) m1A and m1G can induce RNA duplex unwinding. Modifications can also affect other physical parameters such as RNA folding. A better understanding of the regulatory mechanisms of RNA may reveal how stress and environmental toxins cause harmful effects.

### Plenary Discussion Highlights

- The synthesis of the specific RNA modifications and the generation of long RNA transcripts that include modified bases and sugars are major challenges.
- Specific chemical probes are needed to identify direct contacts between RNA bases, as well as between RNA bases and amino acids in bound proteins. As structural studies are low throughput, additional biochemical tools, e.g., crosslinking mass spectrometry, and high-throughput protocols for structural biology are needed.
- Improvements in RNA-binding protein binding motif quantification as well as the influence of RNA structure on RBP occupancy and gene expression are needed.
- There was strong support for developing structural prediction capabilities for RNA based on the approaches taken with successful protein efforts, though new methods and approaches will likely also be needed to address unique aspects of RNA structure.

### Breakout Session 2A: Association with RNA Binding Proteins (RBPs)

**Moderator:** Samie Jaffrey, M.D., Ph.D.

**Primary Participants:** Kathy Liu, Ph.D. and Chris Burge, Ph.D.

There was a discussion on RNA-protein interactions. Current capabilities can readily identify RBPs and some of the RNA sequences recognized by RBPs. However, many aspects of these interactions remain unknown, including the stoichiometry in regard to what fractions of a specific RNA are bound by RBPs, the duration of interactions between RBPs and specific RNAs, and the biological impact of the RNA-protein interactions. To address these questions, new high-throughput methods to study RBP-RNA interactions are needed. Understanding the biological relevance of RNA-protein interactions, the resident time of an RBP on an RNA, and how phase-separation in the nucleus alters and contributes to RNA binding properties will be critical. Challenges in studying RNA-protein interactions include the generally fast speed of enzymatic reactions that necessitates yet-to-be developed approaches to assign biologic function.

**Breakout Discussion Highlights:**

- A broader approach that takes a more holistic view of entire RNA transcripts rather than mapping individual binding sites is needed. Methods to determine all the proteins that are bound to an individual transcript at a specific moment of time, rather than an average of measurements, is also needed.
- Key questions include: How does the field develop capabilities to move from qualitative to quantitative global analysis of RNA-binding interactions for all RNAs? The stoichiometry of the binding events cannot be measured currently using Cross-Linking and Immunoprecipitation (CLIP)-seq methods. There is a need for new types of RBP/RNA mapping methods to quantify stoichiometry.
- Additional work is needed to characterize subcellular interactomes, and to understand interactions change in phase-separated compartments or during the RNA life cycle. Subcellular methods for detecting RNA-RBP interactions would be quite informative.
- Decay during the life cycle of RNA is a major component of regulation. Most current studies have looked at stable RNA-RBP binding while ignoring more difficult-to-detect, transient stages due to a lack of suitable exploration methods. Methods for mapping transient interactions will address an area of RNA biology where there is currently little insight.

## Breakout Session 2B: Implications of Heterogeneity and its Impact on Stress

**Moderator:** Lydia Contreras, Ph.D.

**Primary Participants:** Tom Begley, Ph.D. and Yu-Ying He, Ph.D.

Oxidative stress elicits a dynamic response that includes system-dependent processes such as RNA turnover, clearance, and metabolism. Determining the specific RNA modifications involved in the oxidative stress response resulting from environmental exposures and the role for those modifications should be a research focus. This breakout session addressed two fundamental questions: 1) What are the impacts of environmental exposures, oxidative stress by reactive oxygen species (ROS), and alkylation exposure on RNA structure and therefore disease? 2) How does stress impact different RNA species in terms of signaling, RNA-protein interactions, and trafficking? Variation in the half-life of RNA species, response to oxidative stressors, and the impact of RNA-protein interaction were also discussed.

**Breakout Discussion Highlights:**

- Different types of chemical modifications of RNA are induced enzymatically in response to ROS and oxidative stress, but only a limited number of them are understood. Specific enzymes that

deposit these stress-induced modifications on the RNA must be identified to evaluate the distinct impacts of oxidative stress.
- Significant effort is needed to develop an understanding of the impacts of the interactions between environmental exposure and RNA modifications on RNA structure. Exposure-induced modifications may be dose- and time-dependent which could lead to alterations of multiple RNA modifications on specific RNA species. There is a need to understand this stress-response mechanism and downstream effects of the regulation on RNA structure.
- Technological advances that facilitate the characterization and mapping of RNA modifications in biological systems, including tools to study specific RNA modifications in a variety of different cell types as well as methods to connect these changes to specific diseases that result from environmental exposures, are needed.
- The capacity to map temporal and spatial dynamics of RNA modifications inside a cell and tissue in relation to functional responses to stress or physiological conditions is needed. Advances will require technological innovation, development, and standardization.
- Sensitive methods to determine specific locations and types of ROS-induced modifications in RNA are needed. The effect of stress on different RNA species will depend on their 3D structures.

## Breakout Session 2C: Influences of Modifications on RNA Structure and Dynamics

**Moderator:** Blanton Tolbert, Ph.D.

**Primary Participants:** Dave Mathews, M.D., Ph.D. and Xiao Wang, Ph.D.

This breakout session addressed experimental and computational methods that consider base and sugar modifications in the prediction of secondary structure of RNA. In addition to the utilization of predictive algorithms, the prioritization of specific common modifications for standards development was also highlighted. The limitations of high-throughput SHAPE-type chemistry, which focuses on the 3D arrangement of atoms and chemical bonds in a molecule, and secondary structure prediction were discussed. Structural analyses are complicated by the dynamic nature of modifications including the different types of chemical reactions that can modify an RNA in different cellular conditions. A key challenge is developing approaches that will allow for identification of multiple different modifications on a single transcript. Better methods and enzymes that can identify site-specific modifications will be essential. There is also a need to understand how RNA modifications affect RBP occupancy. Transcriptome wide visualizations of RNA structural dynamics will help enhance understanding.

**Breakout Discussion Highlights:**

- Many modifications of RNA occur in clusters, and it is important to understand how the different combinations of modifications affect RNA structures.
- RNA secondary and tertiary structure prediction is an important opportunity. Few RNA structure/prediction methods take chemical modifications into consideration. There is a strong need to predict the effect of base and sugar modifications on RNA structure and their impact on translation to proteins.

- Selective 2'-hydroxyl acylation analyzed by primer extension and sequencing (SHAPE-seq) provides a snapshot about RNA structure but fails to capture dynamic information. Incorrect prediction of secondary structure suggests additional methods and refinements are needed.
- There are several existing methods for quantitative sequencing of specific types of RNA modifications. However, inefficient methods for RNA ligation and barcoding are major factors limiting innovation and development, single-cell approaches, and study of structural dynamics.
- Software and laboratory tools that address hundreds of modifications at high throughput remain to be developed.
- Development of methods for comprehensive analysis of RNA-protein interactions is needed to better understand these dynamic biological complexes. It will be essential to solve the even more complex problem of determining base modifications in relation to the complexes at the individual RNA and sequence level. Effort in this area needs to be coordinated.

## Scientific Session 3: RNomics Applications from Research to Clinic

**Moderator:** Wendy Gilbert, Ph.D.

Jeannie Lee, M.D., Ph.D., discussed clinical and therapeutic applications of RNomics. Her presentation covered two main approaches – RNA as the drug itself and RNA as the target of the drug. Although most of the genome (98%) is actively transcribed, current drugs only target about 700 proteins. Many diseases, including Huntington's Disease and Amyotrophic Lateral Sclerosis, are linked to toxic RNAs, indicating that current drug development can be vastly expanded to include RNA targets. She shared her experience working with Merck on a proof-of-concept project to identify small molecule targets for a lncRNA required to maintain X-inactivation. A small molecule was identified that bound to the ncRNA, Xist, and blocked its interaction with PRC2 and SPEN. Results from this project suggest that specific RNAs can be systematically targeted by small molecules using an unbiased screening approach even without advanced knowledge of RNA's 3D structure. She also discussed future directions for the RNA therapeutics field.

### Plenary Discussion Highlights

- There is a need to create and curate small molecule (SM) libraries targeting RNA structures with multiple SMs for each target RNA with varying binding efficiencies. It will be necessary to probe more thoroughly into the different possible RNA conformations (i.e., secondary or tertiary structures) to identify potential targets for SMs. It is also critical to characterize RNA-protein complexes to identify additional druggable targets.
- It will be important to explore RNA-modifying enzymes as potential drug targets.
- High-resolution RNA structures will still be needed to identify SMs for therapy; differences in the sequence and structure of RNAs are often subtle. Structural information will be helpful in drug development efforts, particularly for identifying small molecules with increased efficacy and less toxicity.
- There was a general discussion about the need to understand endogenous modifications in a range of tissues to discern disease-related modifications in human cells and those in the genomes of RNA viruses.

- Participants expressed the need for reference RNA databases utilizing tissues from a range of donors that will need to be flexibly developed to account for development and innovation of technologies for direct RNA sequencing and modification detection of low abundance RNAs.

## Breakout Session 3A: RNA Therapeutics

**Moderator:** John Cooke, M.D., Ph.D.

**Primary Participants:** Li Li, Ph.D. and Samie Jaffrey, M.D., Ph.D.

Two primary topics were covered in this breakout session: 1) the therapeutic implications of RNA biology research and 2) the impact of RNA modifications on development of RNA therapeutics. There was consensus on the significant need to translate fundamental RNA biology knowledge into treatment and clinical applications while communicating the importance of this research to the public, especially given the recent success in RNA vaccines. There is also a need to distinguish between RNAs themselves as therapeutics and RNAs as targets of therapeutics. Among the current challenges are the high manufacturing and transportation costs of RNA therapies, especially for developing countries and rare diseases.

**Breakout Discussion Highlights:**

- There is a need to establish therapeutic development pipelines and diverse and integrated teams that encompass the range of expertise needed for all steps of the process including initial discovery, synthesis of high-grade oligonucleotides, animal studies, and good laboratory practice manufacturing for pre-clinical and clinical studies. This may include both academic labs and industry partners.
- Experts in delivery technologies, for example lipid nanoparticles, are critical to this effort, particularly in delivering RNAs to specific tissues. This may also involve modifications to RNAs, and innovative delivery systems including microinjection technologies.
- RNA therapy development has the advantage of being an “informational drug” in that once a delivery platform is developed, the RNA sequence can be changed for different targets.
- Standardized cell lines for testing immunogenicity of RNA therapies, for example human dendritic cell lines, could be an important resource in this effort.

## Breakout Session 3B: RNA Diagnostics

**Moderator:** Matt Disney, Ph.D.

**Primary Participants:** Jason Watts, M.D., Ph.D., Blerta Xhemalçe, Ph.D., and Tao Pan, Ph.D.

This breakout session addressed two main topics: 1) the emerging uses for RNA sequences in diagnostics and 2) ways to enhance current RNA sequencing technologies for application to diagnostics. Small RNAs, specifically tRNAs, have four properties that can be used in diagnostics: abundance, modifications, charging, and fragmentation. Current methods for tRNA sequencing and detection of modification are already applicable to multiple biological samples (e.g., saliva and nasal swabs) and could be applied to disease prognostics and guiding therapeutic approaches, an example of this is use in colon cancer staging. Likewise, RNA modifying proteins are emerging as disease targets, and improved methods for detecting disease-related modifications can aid in biomarker development for both disease diagnostics and monitoring treatment. There are already diagnostic tests based on mRNA sequence. Including analysis of mRNA modifications will add a new layer of complexity to enhance biomarker development. New drugs are currently being developed for alternate splice forms of RNA, and there is a need to detect the range of alternative splice forms in different tissues. Cell-free RNAs are also being explored for cancer diagnosis.

**Breakout Discussion Highlights:**

- New technology needs include improved reverse transcriptase enzymes to generate accurate sequences from longer RNAs along with accompanying chemical and enzymatic approaches that convert modifications to mutations during transcription. In the short term, these methods could enhance current long and short-read sequencing approaches.
- A centralized resource of high-quality, searchable RNA sequences and modifications in typical and diseased tissues is needed. This could involve integrating existing RNA data sets (e.g., GTEx, ENCODE, HCA, and HuBMAP) into a single resource. Such a resource could provide baseline RNA data in typical tissues that could be used for comparisons with RNA profiles in disease studies.
- A library of RNA features in surrogate tissues (e.g., lymphocytes or other WBCs) that could be a useful source of surrogates for neuronal cells or other target tissue types.
- There is need to develop lower cost, high-throughput methods for application of RNA biomarkers.
- Mass spectrometry approaches for RNA sequencing and modification detection at single nucleotide resolution are needed to complement direct RNA sequencing approaches.

## Breakout Session 3C: Viral RNome

**Moderator:** Marcos Morgan, Ph.D.

**Primary Participants:** Stacy Horner, Ph.D. and Dirk Dittmer, Ph.D.

This breakout session focused on the need, scope, and feasibility of a viral RNome project. Studying the viral RNome has distinct advantages including the public health benefit of developing antiviral drugs, gained understanding of the basic biology of viral RNA modifications, and knowledge of how this may affect virus evasion of the innate immune system. In addition, viral genomes are generally small making this a more feasible area of research. Modifications (e.g., N6- and 2’-O-methylation of adenosine) of viral RNA affect viral replication and may help the virus to evade the innate immune system, but currently this knowledge is incomplete as it is limited by the inability to detect all the modifications in viral RNA. An unbiased catalog of viral RNA modifications and the ability to quantify these modifications will be essential for a broader understanding of modifications in the RNA field. This effort would entail combining mass spectrometry (MS) approaches, to compile a complete catalog of modifications, with direct RNA sequencing approaches to determine the sequence specific locations of the modifications. New tools, including new chemical methods for probing RNA 3D structure and enhanced imaging methods, to characterize new and emerging RNA viruses with respect to sequence, structure, and modifications are essential for this proposed pilot project.

**Breakout Discussion Highlights:**

- A viral RNome project would include comprehensive mapping of RNA structures within the virion and infected cell as well as a resource cataloging viral transcriptomes. It is also critical to understand the modifications of the RNA throughout the life cycle of viruses.
- A centralized approach could be created to analyze viral RNAs from multiple labs. A possible workflow could entail combining the MS identification of RNA modifications, usage and synthesis of training standards, and modification-sensitive direct RNA nanopore sequencing of viral transcriptomes.
- It will be necessary to develop sequencing methods that are not dependent on poly-A tails.
- A comprehensive program for studying the RNome of many viruses would be difficult, and it may be wise to concentrate on viruses that have the greatest public health significance.

## Scientific Session 4: Technologies for Direct Sequencing

**Moderator:** Chuan He, Ph.D.

Meni Wanunu, Ph.D., spoke of current technologies for direct RNA sequencing including nanopores and mass spectrometry (MS) in relation to current prospects and future needs. The most prevalent method uses cDNA-based sequencing..This is an indirect method and most modification data is lost through the conversion of information-rich RNA into the four canonical DNA bases. cDNA sequencing is predominantly performed on short read platforms. Newer approaches utilize single-molecule methods to obtain long reads with nanopore, zero-mode waveguides, and other emerging technologies that are able to better sequence isoforms. Comparisons of abundant short read cDNA information with rarer long read information have enabled identification of many gene isoforms, but much work is still necessary to make this a routine laboratory procedure suitable for clinical implementation.

Direct sequencing of long RNA molecules with simultaneous modification identification is both a need and an opportunity. Development of low-cost and high-throughput methods to sequence full-length RNA directly will be required to advance the field. Methods for routine identification of both common and uncommon RNA modifications in sequence context are needed. The emerging field also requires the development of gold standard RNAs with known modifications in different sequence contexts spanning the diversity of possible RNA sequence variation.

Technologies and approaches to address secondary structure and RBPs are another area of opportunity in RNA research. The success of protein folding prediction (e.g., alpha fold) suggest that similar efforts for RNA with bound RBPs may be fruitful. The interplay between advances in mass spectrometry and RNA analysis approaches will also be important for advancing the field, especially as these techniques inform and guide interpretation of RNA modifications. Overall, the needs Dr. Wanunu outlined can be addressed by emerging technologies but will require focused efforts and collaboration from both academia and industry.

### Plenary Discussion Highlights

- There is strong need for technologies to measure the more common RNA modifications, especially as they change with time on individual RNA molecules, to better understand their co-maturation, function, and the mechanisms involved.

- The field needs sequencing technologies that allow for capture of the transcription start site, poly(A) sites, poly(A) length, the many RNA modifications, and transcript isoforms. RNA structures also need to be elucidated, and the sequences bound by RBP should be determined. Sensitive methods are needed given that many transcripts are expressed in low abundance.
- RNA standards with specific known modifications are essential to enable large scale efforts to develop and mature detection technologies and enable subsequent application to important questions. The size and surrounding sequence context of the standards should be appropriate for the technology (e.g., for nanopore detection technology, varying the 3-4 bases on either side of the modification in all sequence combinations if possible).
- There are many existing and developing technologies that each provide a useful perspective. In the short and long term, cross-validation studies of different technologies and data types are needed for the diversity of common RNA modifications and isoforms. The community would benefit from cross-comparison of existing methods, development of new ones, and establishment of best-practices guidelines for the diversity of opportunities to address RNA biology and biomedicine.

## Breakout Session 4A: Technologies on the Horizon

**Moderator:** Jens Gundlach, Ph.D.

**Primary Participants:** Patrick Limbach, Ph.D. and Ben Garcia, Ph.D.

The discussion was primed to address the key technologies that are at the early innovation and development stages and other technologies that hold promise for direct RNA analysis and sequencing. Both mass spectrometry (MS) and direct RNA sequencing were viewed as having significant current capabilities with considerable near-and long-term room for innovation and development. Discussion focused on a great need for multiple RNA analysis technologies since there are often performance tradeoffs. Both fragmentation and non-fragmentation, molecular analysis methods that strike a balance between quantification capabilities and sensitivity requirements are needed.

**Breakout Discussion Highlights:**

- Comprehensive analysis by MS needs size-specific RNA digestion methods that could be based on new or improved nucleases or the development of physical methods like acid hydrolysis.
- Current MS methods use positive ion technology that is adapted for negative ion RNA applications. This is an opportunity for the innovation and development of new, high-throughput, multiplexed negative-ion MS technology from the ground up.
- Nearly every aspect of direct RNA analysis needs major improvements in cost, throughput, and accuracy of base modification determinations and calls to reach comprehensive capabilities. Methods to examine larger RNA transcripts that reflect the true complex nature of the transcriptome are particularly needed as promising new MS and sequencing approaches are on the horizon.
- Structural analysis of protein-RNA complexes and cross-linking MS studies would benefit from development of new photoactivatable MS cleavable cross-linkers for quantitative analysis.

- A consortium approach would give experts the opportunity to work collaboratively, enabling wet and dry lab advances by focusing on technology innovation and development to enable RNA-based advances in biology and medicine.

## Breakout Session 4B: Leveraging Existing Technologies – Opportunities for Further Innovation

**Moderator:** Stirling Churchman, Ph.D.

**Primary Participants:** Chris Mason, Ph.D. and Qi Chen, Ph.D.

The session addressed technologies that are being applied to direct RNA analysis and areas in need of innovations. Existing technologies are leading to significant progress in the understanding of RNA metabolism and processing, but much work is needed to decrease analyte requirements while increasing throughput and capabilities.

**Breakout Discussion Highlights:**

- Better enzymes for RNA manipulation and analysis are broadly needed for uses that include cDNA reverse transcription and ligation of single-stranded RNA.
- Currently, there is limited capacity to synthesize most of the ribonucleosides and ribonucleotides with modifications in the bases and sugar. Focus on providing gold standards and training sets for the common modifications that can be synthesized would bring direct RNA sequencing capabilities closer to those available for DNA today.
- There is a great need to map modifications to individual RNA bases using nanopore, MS, and other de novo sequencing and analysis technologies.
- Nanopore based RNA sequencing has a lot of promise, but well-defined base modification standards that span the sequence detection space are critically needed to train sequence base-callers.
- Technologies that are being applied today to understanding RNA metabolism such as SLAM-seq, Time lapse-seq or 4SU labeling for RNA metabolism studies would benefit from further development.
- There is a strong need to simultaneously identify the multiplicity of RNA modifications in single molecules for a stoichiometric view of isoforms.

## Breakout Session 4C: Unmet Challenges and Needs

**Moderator:** Vivian Cheung, M.D.

**Primary Participants:** Tatjana Trcek, Ph.D. and Hagen Tilgner, Ph.D.

This session addressed current roadblocks to progress in the RNA field and the critical questions that need new analysis strategies and tools. Discussion topics included technology development, pilot projects to understand the function of specific RNAs in the context of modifications and splicing isoforms, the need for nomenclature for all the modified bases and sugars, and the need for searchable databases. Ways to push the field forward were discussed, including targeted analysis of the transcripts of a limited set of specific genes as well as a single viral RNA genome to provide comprehensive information on a few RNA transcripts and/or a small RNA genome.

**Breakout Discussion Highlights:**

- Sequencing and other technologies are needed that capture functionally relevant RNA information at the tissue, single-cell, and sub-cellular level in humans, RNA viruses, model organisms, and cell lines to further understand changes over time and function in normal and disease states.
- Existing databases address some aspects of functionally relevant RNA information, but concerted efforts are needed to integrate existing and expand RNA knowledge utilizing standardized nomenclature. Future efforts should move towards comprehensive, integrated, and useful knowledgebases.
- Today's sequence file formats (e.g., BAM) will need significant updating to accommodate advances in RNA biology and knowledge. The full annotations of functionally relevant RNA information likely will require entirely new nomenclature and new file formats to accommodate all the RNA modifications.

## Scientific Session 5: Infrastructure, Bioinformatics, and Critical Resources: What is Needed?

**Moderator:** Phil Bevilacqua, Ph.D.

Christopher Burge, Ph.D., presented an overview of the current state of the field for bioinformatic analysis of complex RNA data. He pointed out that all sequencing technologies have biases originating from library preparation or the sequencing itself, which can differ between samples and batches. In addition, Dr. Burge emphasized the importance of recognizing the origins of these biases, developing robust statistical procedures to help minimize their impact, and using insights about biases to inform future technologies. It is also important to account for cell type composition in tissues and mixtures of cells when analyzing mRNA features and recognizing different isoforms. Dr. Burge called for more consistent nomenclature for exons, isoforms, and modifications, and stressed the crucial nature of their revision, pointing to the importance of updating gene annotations with new knowledge.

### Plenary Discussion Highlights

- It is necessary to be aware of sequencing biases influenced by GC content, transcript lengths, transcript abundance, and isoforms.
- A difficult task of coordination will be required to devise and utilize standards for RNA molecule characterization across databases and analysis platforms.
- While standard databases do a good job of annotating coding regions, annotation quality of the 5' and 3' ends of RNA transcripts, alternative splicing, and other RNA features varies widely. More comprehensive efforts are needed to fill in the gaps.
- It is important to view the biological importance of transcripts while taking stochastic mis-splicing errors and splicing intermediates into account.
- There is a need to facilitate bridges between research communities including those focusing on single-cell analysis of RNA splicing, 3' UTR polyadenylation, and transcription/translation machinery.

## Breakout Session 5A: Infrastructure and Bioinformatics

**Moderator:** Angela Brooks, Ph.D.

**Primary Participants:** Manolis Maragkakis, Ph.D. and Kin Fai Au, Ph.D.

This breakout session addressed the tools and datasets necessary for analysis of direct RNA sequencing data and ways artificial intelligence (AI) and machine learning (ML) can be utilized and further developed. It was pointed out that the Long-Read RNA-seq Genome Annotation Assessment Project (LRGASP) is evaluating methods for identification and quantification of RNA transcripts. Preliminary findings indicate a technology-dependence bias on long-read sequences.

**Breakout Discussion Highlights:**

- There is an urgent need for appropriate training sets, including larger data sets, for validation and benchmarking.
- There is an immediate need for more standardized formats for data storage and standardized pipelines or software to manage raw data.
- Participants called for journals to require data be deposited in either raw form or standardized fast5 formats.
- A critical problem involves the limitations in existing bioinformatics applications that hinder their use in direct RNA sequencing. The building of ML/AI-based tools to capture modifications may incentivize increased direct RNA sequencing and stimulate the development and use of new bioinformatics systems.

## Breakout Session 5B: Critical Computational Resources

**Moderator:** Tom Begley, Ph.D.

**Primary Participants:** Avi Ma’ayan, Ph.D. and Phil Bevilacqua, Ph.D.

This session covered cloud computing needs and the current state of RNA databases along with upstream and downstream data analysis needs. Overall discussion focused on data integration, data reproducibility, and training. Approaches to facilitate sharing and combining data from different cell types and formats, as either an average or single data sets, are needed to facilitate sharing across groups. A preference was expressed to minimize data processing and thereby the overall cloud computing costs.

**Breakout Discussion Highlights:**

- There is a need to outlined strategies to facilitate data processing and analysis including electronic (e.g., Jupyter) notebooks to streamline workflows and software platforms to facilitate analysis and reproducibility.
- Software code should be included with the data deposition along with better and accurate descriptions of biological context, experimental conditions, and analysis.
- The overall integration, harmonization, and analysis capabilities of current databases are inadequate. A centralized RNA knowledgebase that harmonizing information across multiple resources and accounts for the full diversity of RNA types could address this gap.

- Due to accessibility concerns for RNA-seq data from dbGAP, which lacks global analysis capabilities, efforts should focus more on generating a facile, sequence-level search engine that protects patient privacy.
- There is a need for additional training in data science so that researchers have analytic capabilities and understanding for the rapidly advancing RNA field.

### Breakout Session 5C: Critical Biological Resources

**Moderator:** Pete Dedon, M.D., Ph.D.

**Primary Participants:** Yinsheng Wang, Ph.D., Sara Rouhanifard, Ph.D., and Harald Schwalbe, Ph.D.

This session focused on identifying needed RNA modification resources. Highlighted key topics in the field included: 1) the status of formulation of synthetic RNA standards for known RNA modifications; 2) how the available standards are employed; 3) prioritization of RNA modifications for additional standards generation; and 4) who will generate the standards and how will this be accomplished in a timely fashion. It was recognized that current efforts for standards have been less robust in terms of synthetic controls for nanopore sequencing and there is a new sense of urgency focused on establishing new synthesis methods that incorporate surrounding sequence context. Emphasis was placed on 1) the need for reliable, appropriate, and verifiable standards for the more prevalent modifications, 2) attention to variations within and across tissue types, 3) standards developed from non-templated synthesis (e.g., solid-phase chemical and chemoenzymatic synthesis), and 4) commercialization of standards for wider use by the community. There was discussion about developing a toolbox of RNA-modifying enzymes that can be used to construct RNA modifications in their proper sequence context, as well as engineering enzymes for more customized and flexible synthesis of standards.

**Breakout Discussion Highlights:**

- More attention should be given to high occupancy modification sites and to regions that are frequently shared across cell types/lines.
- There is an opportunity to incorporate more information on RNA modifications into structure-function analyses in order to better understand RNA-protein and small molecule interactions resulting from different types of modifications.
- Nuclear Magnetic Resonance (NMR) studies could facilitate elucidation of biological function and RNA drug discovery by more accurately measuring sites of modification and allowing measurement of binding affinities. NMR combined with mass spectrometry would allow for more quantitative analyses.
- As a standardized resource, the use of appropriate cell lines for purification of large quantities of bulk RNA for subsequent analyses by multiple investigators with a variety of approaches would help advance the field.

## Scientific Session 6: Facilitating Technology Dissemination and Adoption

**Moderator:** Mark Adams, Ph.D.

Brenton Graveley, Ph.D., discussed facilitating technology dissemination and adoption. He divided his presentation into instrumentation, dissemination, standards, and training. Focused attention is needed to democratize the availability of instrumentation, disseminate technology, create data standards, and increase training in computational biology. A common goal of efforts in these areas should be to facilitate the rapid adoption of new methods and technologies.

## Plenary Discussion Highlights

- The need for funding to purchase appropriate equipment was emphasized. This funding might include supplements to grant awards to purchase needed equipment and grant programs for equipment purchase targeted to resource-challenged and undergraduate institutions.
- Technology dissemination and adoption can occur in many different forms, from incorporating a coordination component in large research consortia/centers to spearheading dissemination efforts, to running dedicated outreach sessions at regional and national meetings.
- Both data and experimental standards are important. ENCODE and other large consortia have established good practices and standards that serve as models to enhance experimental replicability in future efforts.
- Additional training support by various mechanisms is needed in such disciplines as genomic science and computational biology. Summer internships with industry for graduate students can be an effective training opportunity.

## Breakout Session 6A: Technology Dissemination and Adoption

**Moderator:** Kate Meyer, Ph.D.

**Primary Participants:** Mark Adams, Ph.D. and Mark Akeson, Ph.D.

This breakout session focused on identifying barriers to the dissemination and adoption of the latest genomics technologies. Two critical factors could significantly improve technology dissemination and adoption – 1) quality training videos on technology use and data analysis, and 2) development of widely-accepted data and experimental standards. The expansion of technology adoption, transitioning from "early adopters" to "early majority," is critical for maintenance of any technology.

**Breakout Discussion Highlights:**

- Barriers to adoption of the latest, state-of-the-art sequencing technologies such as nanopore sequencing include the biological material input amounts, throughput limitations, accuracy, costs, training, intellectual property restrictions, and publication difficulties. These barriers are compounded in resource-challenged institutions.
- The dissemination of new technologies could be a double-edged sword if those using the data/methods do not know how to use them properly and end up generating and publishing low-quality data. Methods/technologies in the wrong hands can be damaging to science. If graduate students, post docs, and junior investigators are not getting the proper training, they may not know how to correctly analyze a dataset. Widely accepted data and experimental standards would aid in mitigating this risk.
- Funded competitions and collaborative efforts could be effective tools for technology dissemination and adoption. These mechanisms would also provide training opportunities.

## Breakout Session 6B: Training

**Moderator:** Mike Summers, Ph.D.

**Primary Participants:** Brenton Graveley, Ph.D. and Jamie Cate, Ph.D.

This breakout focused on training needs and noted that lack of adequate training in several areas is a significant barrier to research progress in the genomic sciences. There is a need for improved training in computational biology and bioinformatics in the biomedical sciences, generally.

**Breakout Discussion Highlights:**

- The availability of a library of quality, easily accessible, low or no-cost online training videos would be a significant driver in technology adoption. Such a resource would require a home as well as a long-term commitment to funding and maintenance. Specific NIH support was proposed as one way to encourage the creation of training videos.
- Students should be taught how to handle large datasets. This should include hands-on training, since people are frequently more engaged in the analysis process when they are using their own data.
- Currently training awards are predominantly made to large, research-intensive institutions. Training efforts should be extended to include more diverse and smaller institutions as well as professional schools.
- Expanded curricula that includes more didactic training on computational biology/bioinformatics/data analysis should be developed and distributed. Curriculum development is frequently driven by pre-professional programs which can be slow in incorporating new ideas and developing new courses.
- Individual career awards and diversity supplements are considered important sources of NIH support to encourage workforce diversity. Outreach components should be required in large NIH-funded projects such as consortia and center projects. Institutions can also support efforts to build diversity.
- “Permanent scientist” award programs should be expanded to support academic scientists who are not in tenure-track positions but who represent a critical component of the research enterprise.

## Wrap-up Session

**Moderators:** Blanton Tolbert, Ph.D., Vivian Cheung, M.D., Pete Dedon, M.D., Ph.D., and Brenton Graveley, Ph.D. (Workshop Executive Committee Members)

The workshop concluded with a summary of the sessions presented by Blanton Tolbert, Ph.D. on behalf of the executive committee who moderated this session. Dr. Tolbert highlighted key areas covered and identified needs for the RNomics field.

The impact of nucleotide sequence and modifications on RNA structure, dynamics, and biological roles were discussed; topics included associations of RNA with RBPs, the vast diversity of RNA in terms of sequence, structure, and functions, and the impact of cellular and environmental stress on RNA processing. From the genomic medicine perspective, participants considered how to leverage lessons learned from the COVID-19 pandemic, think about new drug targets, and maximize the potential of RNA therapeutics. The group emphasized how complete RNA sequences and technologies for RNA sequencing will benefit disease diagnostics, development of biomarkers, and therapeutics. The participants also discussed how sequences of RNA viruses may be a valuable resource for understanding RNA and protein-RNA biology.

To address the challenges of RNA sequencing, participants highlighted existing technologies as well as those on the horizon. Conversation centered around the application of direct RNA sequencing, mass spectrometry, and cDNA sequencing to gain an understanding of the stoichiometry, identity, and position of modifications in the transcriptome in the context of splicing isoforms. There are also challenges in the infrastructure, bioinformatics, and computational spaces including significant training needs. Inadequate biological standards were highlighted frequently as a barrier to future RNomics work. The final scientific session covered topics including technology dissemination and adoption, diversity and inclusion, training the next generation of scientists, and communicating work in this field to the public so it can have broader impact.

To make progress in these areas, Dr. Tolbert noted that it will be essential to work on multiple fronts collaboratively and in parallel. Simultaneously, the field will need to generate fundamental understanding, develop innovative technologies, and translate knowledge into platforms for diagnostics and therapeutics. He also emphasized the need to identify, recruit, and support new talent, workforce diversity, and career options in RNomics. A fundamental goal is to understand the modifications on all transcript isoforms in sequence, structural, and RBP-binding context and make the data available in user-friendly formats. Resources, innovation, personnel, and leadership will be fundamental to the success of these efforts.

## Key Needs

1. Standards, including cell lines and those used for calibration.
2. Reagents including enzymes for writing/erasing site-specific modifications, new chemicals for synthesis of modified bases in synthetic RNAs, antibodies for detection of all known RNA modifications, and specific RNA protein bifunctional chemical cross-linkers.
3. Extraction methods and tissue sources for quantitative sequencing of intact RNA.
4. Quantitative and single-molecule direct RNA sequencing methods.
5. Methods that identify structures, the impact of modifications on structure, and the importance of structure to RNA biology.
6. Intracellular RNA visualization methods.
7. Data analysis algorithms and databases.
8. Collaborations, partnerships, and crosstalk between academia, industry, and government and across disciplinary bounds.
9. Broad interdisciplinary training, including computational training and retraining.
10. Mechanisms for community engagement and dissemination of knowledge.

## Wrap-up Discussion Highlights

- Any RNomics efforts should consider all RNAs in a cell.
- There is a strong need for innovation and development of new technologies and to decrease costs to make them broadly applicable to more research avenues and bring them into widespread use by the field and larger scientific community.
- A federated, collaborative model of organization may be an effective way to tackle some of these challenges. Support should be spread geographically and across a diversity of institutions and individuals to mirror and leverage relevant expertise and scientists. The framework should foster collaboration and communication, like NHGRI's Technology Development Coordinating Center (TDCC), and continue to bring in new investigators and ideas.
- There may be a need for resource centers around topics such as standards, mass spectrometry expertise, and directed protein evolution efforts for better enzymes.
- No single approach will solve these problems and efforts will need to be large in scope. This work will require collaborations between molecular biologists, structural biologists, computational biologists, fundamental researchers, clinicians, and technologists. There is a need for many areas of expertise in these efforts, including from outside of the RNA community.
- The National Institute of Standards and Technology (NIST) may play a potential role in RNA standards. Currently, NIST is generating a public set of diverse RNA sequencing data on the Genome in a Bottle lymphoblastoid cell lines and induced pluripotent cell lines available from Coriell which may form a basis for a community effort in RNA.
- Advancements in RNA research have enormous implications for human health. The ability to map modifications in sequence context for clinical applications would be transformative. There are opportunities for RNA as a therapeutic, drug delivery system, or diagnostic tool. Studying the off-target effects of RNA in the drug discovery pipeline will also be critical.