Frontiers in RNA biology: advances from a small fly *Drosophila melanogaster*

James B. Brown (Lawrence Berkeley National Laboratory and University of California Berkeley, USA) **and Susan E. Celniker** (Lawrence Berkeley National Laboratory, USA)

In this article, we discuss emerging frontiers in RNA biology from a historical perspective. The field is currently undergoing yet another transformative expansion. RNA-seq has revealed that splicing, and, more generally, RNA processing is far more complex than expected, and the mechanisms of regulation are correspondingly sophisticated. Our understanding of the molecular machines involved in RNA metabolism is incomplete and derives from small sample sizes. Even if we manage to complete a catalogue of molecular species, RNA isoforms and the ribonucleoprotein complexes that drive their genesis, the horizons of molecular dynamics and cell-type-specific processing mechanisms await. This is an exciting time to enter into the study of RNA biology; analytical tools, wet and dry, are advancing rapidly, and each new measurement modality brings into view another new function or activity of versatile RNA. Since the dawn of sequence-based RNA biology, we have come a long way.

With the *Drosophila* genome sequence in hand in 2000, we commenced genome-wide structural and functional annotation. We used traditional methods of sequencing expressed gene tags (ESTs) and cDNAs[1] to identify portions of the genome that produce RNAs and embryonic spatial gene expression[2] to address the biological function of these RNAs. Although significant progress was made in improving the computed gene annotations, which miss 5′ and 3′ untranslated regions (UTRs) and struggle with joining protein-coding exons, we could only afford to sequence one transcript per gene – usually the most abundant among those of moderate size.

Fruits of the sequencing revolution

Next-generation sequencing revolutionized RNA biology, significantly reducing costs[3] (Figure 1) and allowing us to characterize the transcriptome in spectacular depth, revealing enormous diversity in alternative splicing, promoter and polyadenylation site selection. The modENCODE consortium produced the most comprehensive transcriptional map for a metazoan, surveying the transcriptome of whole animals including a 30 point developmental atlas, 29 dissected larval and adult tissues, 25 distinct cell lines and 21 environmental perturbations to discover condition-specific genes and transcripts (reviewed by Brown and Celniker[4]). Genes with enormously complex splicing patterns were discovered, probably producing thousands of transcript isoforms each via combinatorial usage of

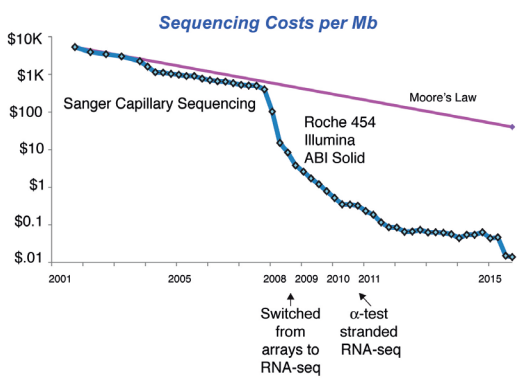


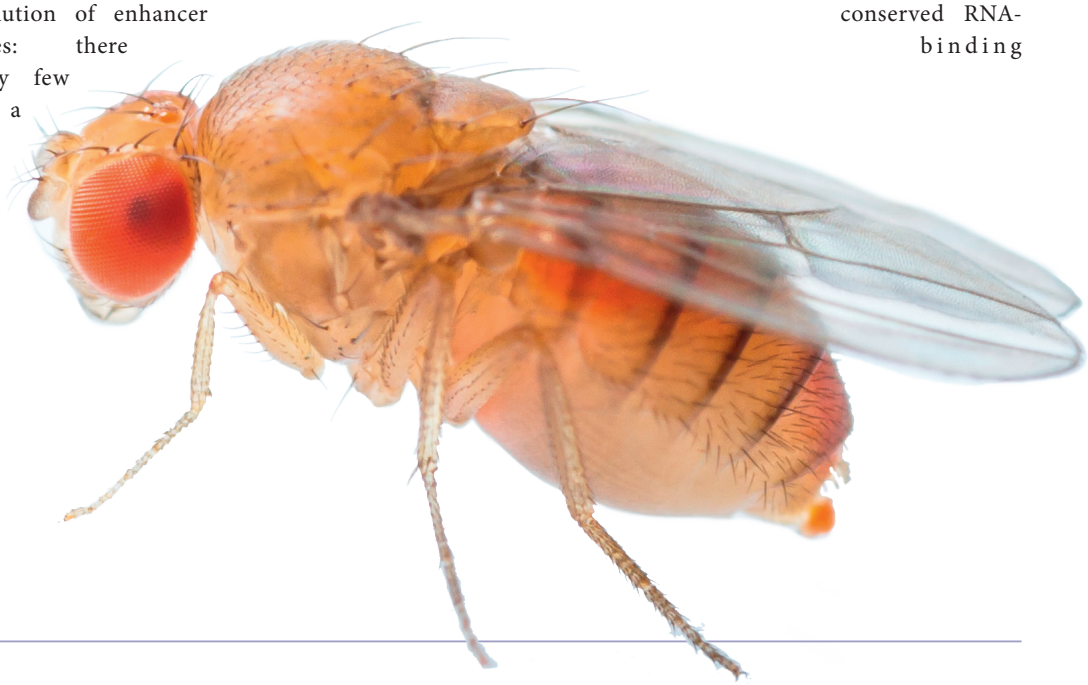
Figure 1. A paradigm shift in sequencing has reduced costs faster than Moore's Law. Our ability to interrogate RNAs radically improved with next-generation sequencing strategies. We used the Roche 454 system to sequence RLM-RACE products to characterize promoters, the Illumina system to sequence CAGE, RNA-seq and PAS-seq and the ABI SOLiD to produce the first total stranded embryonic RNAs (reviewed by Brown and Celniker[4]).

alternative promoters, splice and polyadenylation sites (reviewed by Brown and Celniker[4]). Many new examples of RNA-editing[5], zero-nucleotide recursive splicing[6] and thousands of circular RNAs[7] were also uncovered during the project. Associated primarily with neural tissues, we found ubiquitous alternative polyadenylation site selection – most neural genes encode multiple, sometimes extremely long 3′ UTRs in both the developing and adult CNS[8]. These and related studies in other organisms (reviewed by Mayr[9]) have revealed that the vast majority of genes in many metazoan species admit multiple isoforms – challenging the notion of 'alternative' transcripts, a moniker that specifically implies a dominant isoform. We now view this terminology as a relic of cDNA sequencing; it poorly reflects the reality of transcriptional complexity in metazoans, particularly in neural tissue.

Origin of complex splicing

The discovery of the reality of combinatorially complex RNA processing raises many questions: what is the molecular basis of ultra-complex splicing in neural tissue – i.e. is Down's syndrome cell adhesion molecule 1 (*DSCAM1*)[10] a general model or a specific case? Most alternative splicing is not generated by mutually exclusive splicing cassettes, as in *DSCAM1*, indicating that multiple mechanisms are most likely. We found some time ago, but have yet to publish, that many genes with ultra-complex splicing patterns in fruitflies are similarly complex in humans, but, notably, not the genes with insect-specific (so far as we know) splicing mechanisms (such as *DSCAM1*). The evolution of splicing complexity may be as complex and multifactorial as the evolution of enhancer sequences: there are very few (perhaps a handful) of genes with conserved intron–exon structure between nematodes, fruitflies and humans[11]. Similarly, at least one recursively spliced gene in *Drosophila* (out of 115) is also recursively spliced in humans (out of four) – splicing regulation appears to be conserved even when the positioning of introns in the parent protein-coding sequences has been effectively randomized during evolution[6]. It seems likely that future studies examining the properties of genes with complex RNA processing patterns in multiple species may reveal new mechanisms of regulation and targeting for regulatory proteins.

During the period of the modENCODE consortium, the RNA-binding protein (RBP) embryonic lethal abnormal vision (ELAV) was identified as an effector of 3′ UTR extensions in the nervous system[12], providing a hint of the molecular basis for complex polyadenylation site selection in these tissues. Transcriptome-wide association profiles for 20 sequence-specific RNA-binding factors were mapped using both MS/MS (RIP-MS/MS, proteomics) and sequencing (RIP-seq)[13]. We found that the RBP Mushasi (MSI) is even more strongly associated with 3′ UTR extensions than ELAV, supporting multiple inputs to complex RNA processing. In addition, genes with ultra-complex splicing patterns are highly enriched for interaction with several heterogeneous nuclear ribonucleoproteins (hnRNPs) compared with expressed genes in general – consistent with the hypothesis that the complex splicing observed in neural tissue is regulated, rather than a product of dysregulation. RNA interference (RNAi) and RNA-seq were used to knock down 56 RNA-binding proteins, many known splicing factors and some unknown proteins with conserved RNA-binding



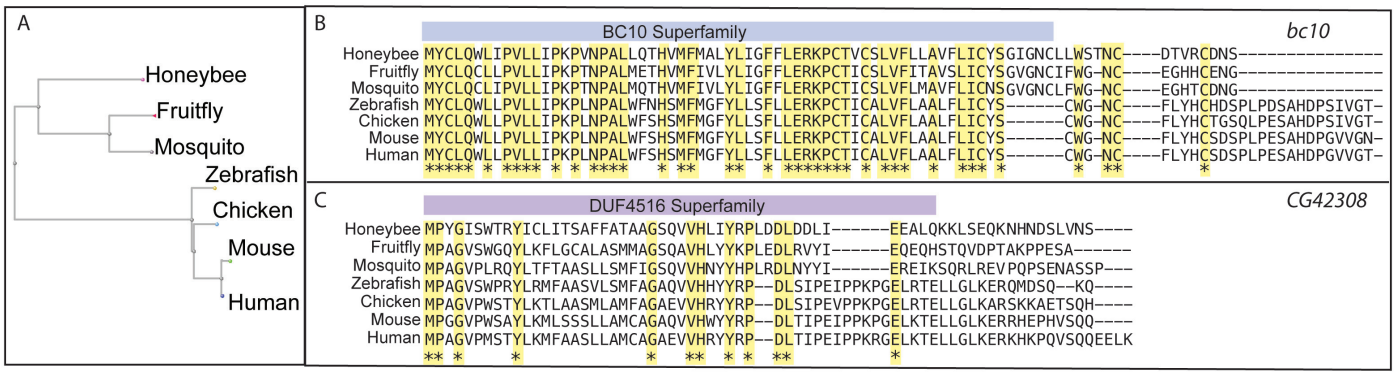


Figure 2. Ultra-conserved smORFs in the human genome with orthologues in *Drosophila*. (**a**) Phylogenetic tree constructed using the BC10/BLCAP locus. (**b**) Multiple sequence alignment of the BC10 peptide in a broad selection of bilaterian species. (**c**) Multiple sequence alignment for an unstudied ultra-conserved smORF gene. Both genes are also conserved in arthropods.

domains[14]. To our considerable surprise, nearly all of the proteins that were knocked down altered promoter selection. This could relate to feedback through multiple layers of gene regulation, i.e. indirect effects, or could indicate a role for RBPs in promoter selection, perhaps via an iterative process where subsequent rounds of transcription stabilize the choice (or choices) of active promoter at any given time, as is the case with the transcriptional reinforcement of H3K36me3 (trimethylated Lys[36] on histone H3) chromatin marks during splice site selection (reviewed by Braunschweig et al.[15] and Naftelberg et al.[16]). Pulse–chase assays (e.g. BruChase-seq[17]) and imaging techniques enable the study of the directionality of these processes[18,19], and this is likely to remain an active area of work for years to come.

Voyage of discovery

In addition to discoveries about RNA processing, the application of RNA-seq to tissues, and single cells, has led to the discovery of new transcripts and genes. Thousands of new antisense transcripts and hundreds of putative long non-coding RNAs (lncRNAs) were identified during our analysis of the modENCODE data. We also found a few hundred genes that encode new conserved small open reading frames (smORFs). Many smORF genes are highly conserved from fruitflies to humans (Figure 2). Indeed, the threshold to designate a new gene non-coding compared with coding was necessarily arbitrary, and the generation of ribosome profiling data by other laboratories in the community has since raised questions about how we draw these lines between mRNAs and lncRNAs[20]. Is sequence conservation and/or the presence of known conserved domains sufficient to make such determinations? We think not. A re-analysis of community ribosome profiling data identified smORFs in roughly 5% of our modENCODE lncRNAs that we now believe to be translated (J.B. Brown and S.E. Celniker , unpublished work), but which have evaded detection by MS/MS. For RNA biology to progress, we need to develop a realistic model of how ribosomes selectively translate some RNAs and not others – and this means developing far more comprehensive maps of spatially localized RNA–protein interactions than yet exist. Beyond RNA biology, it probably also means community investment in direct detection and quantification of peptides. In conjunction with cytoplasmic RNA binding and interaction assays, there is tremendous potential to begin to unravel new layers of ribosome-proximal translational regulation. We found that most RNA-binding factors (of those we have studied) associate with mRNA and protein products from the same gene, suggestive of ribosome-proximal regulatory interactions across many classes of RNA-binding factors[13]. We posit that there are fundamental mechanisms of feedback for biological systems that have yet to be discovered. Perhaps the genetic code is not "…Nature's last elegant solution" after all (quote from S. Brenner cited by Kornberg[21]).

The role RNAs play in metazoan biology is obviously a function of when and where they are expressed in the animal. We've worked to determine spatial gene expression profiles in the Drosophila embryo for a number of years[2,22,23], and this work has led to some insights to the control of gene expression during development, particularly in collaboration with Mark Biggin[24,25]. Our comprehensive map of transcription factor (TF) spatial expression was the first in an animal[22] and from it, we have begun to explore the role of TFs in organogenesis,

in collaboration with Erwin Frise[22,26]. We've also dissected *cis*-regulatory modules that control gene expression[25,27], studied chromatin state and its role in the control of gene expression[28] and hormonal control of gene expression in 41 diverse cell types[29]. The picture that emerges from this work is one of exceptional complexity: each tissue, organ system and likely cell type in an animal admits a unique profile of regulatory proteins and RNAs – and concomitant diversity in promoter activation and RNA processing. We will know the identity of each transcript encoded in a metazoan genome the day we can assay each cell throughout the life cycle of an animal (single cell RNA-seq). Until then, we will necessarily make due with a partial picture, an incomplete parts-list.

Our view of the hierarchy of gene regulation, beginning with initial chromatin remodelling, reinforced through protein-mediated RNA–chromatin interactions, and culminating in translational and post-translational control of protein and RNA gene products continues to evolve. Another frontier is the role of RNA structure in modulating both protein-mediated and direct molecular interactions. Riboswitches[30], upstream ORFs[31] have been known for decades, and it is now clear that chemical modifications to RNAs also alter structure, and therefore RNA–protein interactions[32]. New computational tools are needed that leverage recent advances in machine learning to make substantive progress in the identification of functional and conserved RNA domains, particularly where the conservation of primary sequence is poor or non-existent. It is increasingly apparent that RNA structure broadly influences gene expression[33], or at least is broadly modulated during gene regulation. Perhaps in another 15 years, we will see the emergence of the first quantitative mechanistic models of gene regulation, from initial chromatin remodelling, reinforced through protein-mediated RNA–chromatin interactions, culminating in translational control and feedback.

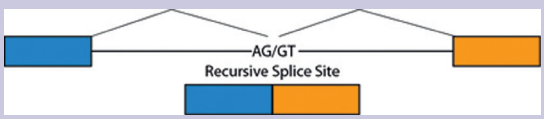
The studies described in this article and many others in which our group has been involved over the years were designed to generate hypotheses and to build the datasets and sufficiently comprehensive maps of gene expression and hierarchies of gene regulation to provide the insights biologists require to formulate innovative new models of gene regulation in the context of development. Fifteen years after the genome sequence, the pace of discovery in RNA biology has never been higher. ■

Glossary

EST or expressed sequence tag is a short sequence usually from the 5′ or 3′ end of a cDNA clone. They are used to identify and characterize gene transcripts.

Mutually exclusive splicing is a form of alternative splicing in which one of two or more exons is retained in the mature mRNAs after splicing, but not both.

Recursive splicing is a multi-step process of large intron removal that consists of juxtaposed 3′ and 5′ splice site sequences, with no exon between them.



RIP-MS/MS RBPs are immunoprecipitated (RIP) and the associated proteins are characterized by tandem mass spectroscopy (MS/MS). MS/MS is an analytical technique used to identify proteins in a complex mixture from the mass and charge of its peptide fragments. A sample is ionized, accelerated and analysed. Ions from the first spectra are then selectively fragmented and analysed by a second stage of mass spectrometry to generate the spectra for the ion fragments. The fragments are used to match to predicted peptides.

RIP-seq RNA immunoprecipitation (RIP) is a method to map *in vivo* RNA–protein interactions. The RNA-binding protein (RBP) of interest is immunoprecipitated (IP), using either an antibody or a tagged protein, together with its associated RNAs. Transcripts are detected by next-generation sequencing (seq).

RNA-binding proteins (RBPs) are a class or proteins that bind to double- or single-stranded RNAs that then form ribonucleoprotein complexes. Many RBPs have characteristic conserved RNA-recognition motifs (RRMs) that act as the binding site for interacting with RNAs. RBPs play important roles in a number of RNA processes including splicing, polyadenylation, mRNA stabilization, mRNA localization and translation.

H3K36me3 is a variant of the histone 3 (H3) nuclear protein, one of the core histones. The lysine (K) at position 36 contains three methyl groups. The methyl groups allow regulation through the binding of other chromatin proteins. H3K36 methylation is a feature of transcribed genes.

Pulse–chase assays are standard assays in biochemistry traditionally using radioactivity to monitor a cellular process over time by exposing cells to a labelled compound (pulse) and then to the same compound in an unlabelled form (chase).

Transcription factors (TFs) are proteins that contain one or more DNA-binding domains and recognize and bind to specific DNA sequences in the genome to control gene expression.

Ben Brown is a Staff Scientist in the Environmental, Genomics and Systems Biology Division at Lawrence Berkeley National Laboratory and head of the Department of Molecular Ecosystem Dynamics. His laboratory develops statistical machine learning tools to study biological responses to ecologically adverse conditions. His research programme aims to establish exposure biology as a foundational science, viewing toxicology as systems-level genetics, with the potential to play a major role in the elucidation of processes ranging from development to speciation. He is the analysis lead for the Consortium for Environmental Omics and Toxicology. email: jbbrown@lbl.gov

Susan Celniker is a Senior Staff Scientist, Deputy Division Director for Environmental, Genomics and Systems Biology Division at Lawrence Berkeley National Laboratory (LBNL) and an Adjunct Professor in Comparative Biochemistry at the University of California, Berkeley. As co-Director of the Berkeley Drosophila *Genome Project she led the LBNL efforts in collaboration with UC Berkeley and Celera Genomics to sequence the* Drosophila *genome. She also led the modENCODE* Drosophila *transcriptome consortium. Her research is focused on understanding the control of gene expression during development. Currently, she is investigating the role of the host/microbiome responses to environmental perturbagens. email: celniker@fruitfly.org*

References

1. Stapleton, M. et al. The *Drosophila* Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. Genome research **12**, 1294-1300 (2002).
2. Tomancak, P. et al. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. Genome biology **8**, R145, (2007).
3. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (2016).
4. Brown, J. B. & Celniker, S. E. Lessons from modENCODE. Annu Rev Genomics Hum Genet **16**, 31-53, (2015).
5. Graveley, B. R. et al. The developmental transcriptome of *Drosophila* melanogaster. Nature **471**, 473-479, (2011).
6. Duff, M. O. et al. Genome-wide identification of zero nucleotide recursive splicing in *Drosophila*. Nature **521**, 376-379, (2015).
7. Westholm, J. O. et al. Genome-wide analysis of *drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell reports **9**, 1966-1980, (2014).
8. Smibert, P. et al. Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. Cell reports **1**, 277-289, (2012).
9. Mayr, C. Evolution and Biological Roles of Alternative 3'UTRs. Trends in cell biology **26**, 227-237, (2016).
10. Schmucker, D. et al. *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell **101**, 671-684 (2000).
11. Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. Nature **512**, 445-448, (2014).
12. Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3' UTR extension in the *Drosophila* nervous system. Genes & development **26**, 2259-2264, (2012).
13. Stoiber, M. H. et al. Extensive cross-regulation of post-transcriptional regulatory networks in *Drosophila*. Genome Res **25**, 1692-1702, (2015).
14. Brooks, A. N. et al. Regulation of alternative splicing in *Drosophila* by 56 RNA binding proteins. Genome research **25**, 1771-1780, (2015).
15. Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R. & Blencowe, B. J. Dynamic integration of splicing within gene regulatory pathways. Cell **152**, 1252-1269, (2013).
16. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of alternative splicing through coupling with transcription and chromatin structure. Annual review of biochemistry **84**, 165-198, (2015).
17. Paulsen, M. T. et al. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. Methods **67**, 45-54, (2014).
18. Vargas, D. Y. et al. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell **147**, 1054-1065, (2011).
19. Carmo-Fonseca, M. & Kirchhausen, T. The timing of pre-mRNA splicing visualized in real-time. Nucleus **5**, 11-14, (2014).
20. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell **154**, 240-251, (2013).
21. Kornberg, T. in Current Topics in Developmental Biology (ed P Wassarman) 552-567 (Academic Press, 2016).
22. Hammonds, A. S. et al. Spatial expression of transcription factors in *Drosophila* embryonic organ development. Genome biology 14, R140, (2013).
23. Fowlkes, C. C. et al. A quantitative spatiotemporal atlas of gene expression in the *Drosophila blastoderm*. Cell **133**, 364-374, (2008).
24. Li, X. Y. et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila blastoderm*. PLoS Biol **6**, e27, (2008).
25. Fisher, W. W. et al. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **109**, 21330-21335, (2012).
26. Frise, E., Hammonds, A. S. & Celniker, S. E. Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. Molecular systems biology **6**, 345, (2010).
27. Berman, B. P. et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome biology **5**, R61, (2004).
28. Thomas, S. et al. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. Genome biology **12**, R43, (2011).
29. Stoiber, M., Celniker, S., Cherbas, L., Brown, B. & Cherbas, P. Diverse Hormone Response Networks in 41 Independent *Drosophila* Cell Lines. G3 (Bethesda) **6**, 683-694, (2016).
30. Garst, A. D., Edwards, A. L. & Batey, R. T. Riboswitches: structures and mechanisms. Cold Spring Harbor perspectives in biology **3**, (2011).
31. Barbosa, C., Peixeiro, I. & Romao, L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet **9**, e1003529, (2013).
32. Liu, N. et al. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature **518**, 560-564, (2015).
33. Ding, Y. et al. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature **505**, 696-700, (2014).