MODULAR ORGANIZATION

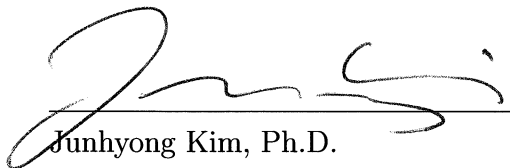# AND COMPOSABILITY OF RNA

Miler T.S. Lee
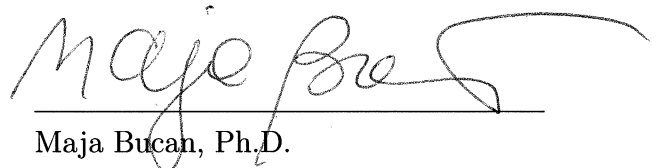
A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2009

Junhyong Kim, Ph.D.

Supervisor of Dissertation

Maja Bucan, Ph.D.

Graduate Group Chair

Harold Riethman, Ph.D.

Dissertation Committee Chair

James Eberwine, Ph.D.

Dissertation Committee Member

Sampath Kannan, Ph.D.

Dissertation Committee Member

R. Scott Poethig, Ph.D.

Dissertation Committee Member

"It is interesting that people try to find meaningful patterns in things that are essentially random."

— Lt. Cmdr. Data, *Star Trek: The Next Generation*, Ep. 5.22

ACKNOWLEDGEMENTS

Inspiration and guidance come in many forms. This work would not have been possible without the help of countless individuals and entities, some of whom I can name below:

My advisor, Dr. Junhyong Kim, who deserves more praise than I am capable of lavishing.

Current and former members of the Kim Lab, who despite working on disparate topics, have helped guide my research and refine my ideas.

Our collaborator Dr. James Eberwine and his lab, with whom we've closely worked on several fruitful projects. In particular, Chapter 4 of this dissertation contains work from a project that was a joint effort between his student Peter Buckley and myself.

The members of my dissertation committee: Dr. Harold Riethman, Dr. James Eberwine, Dr. Sampath Kannan, Dr. Scott Poethig, and formerly Dr. Artemis Hatzigeorgiou, who have each provided me with several years of helpful feedback.

Numerous other mentors, colleagues, and staff at the University of Pennsylvania, particularly the faculty who were instrumental in drawing me to Penn, including Dr. Richard Spielman, Dr. Warren Ewens, Dr. Maja Bucan, and Dr. Lyle Ungar; as well as my fellow students in the Genomics and Computational Biology Program, whose combined talents are the greatest strength of this graduate group; and the many graduate group coordinators who have occupied the office in 1425 Blockley Hall and

have done their part to manage the chaos.

Mentors and colleagues at several other institutions who have contributed to my never-ending education: Sandia National Laboratories; The University of California, Berkeley; Stanford University; The College Preparatory School; and Bentley School.

The various and sundry locations where I wrote the bulk of this dissertation: Good Karma; Brew; the Penn Bookstore; Barnes and Noble; Borders; Starbucks; the Walnut Bridge Coffee House; the Coffee Bar at the Radisson; Last Drop Coffee House; the Green Line Cafe; LaVa Cafe; Chapterhouse; Lovers and Madmen; Sweet Endings; Capo Giro; Rittenhouse, Washington, and Franklin Squares; Clark Park; the Schuylkill Banks; Locust Walk; the steps of the National Constitution Center; the basement of the Mellon Independence Center; Andy's Laundry; the Penn Graduate Student Center; several living rooms, dining rooms, and kitchens; my bed; the shower; the 14th floor of Blockley Hall; the San Leandro Public Library; Delta Airlines flights 3229, 3432, 1643, 1644, 3212, and 2504; the Philadelphia, Minneapolis-St. Paul, Salt Lake City, and Oakland International Airports; SEPTA bus routes 12, 21, 40, 42, and 44 and regional rail route R1; and my office in the Carolyn Lynch Laboratory.

And lastly, in order but not in importance: my parents, who from the day I was born taught me through their words, their actions, and their love how to live a life of meaning; and my friends, without whom I'd be a very dull boy indeed.

ABSTRACT

MODULAR ORGANIZATION
AND COMPOSABILITY OF RNA

Miler T.S. Lee

Supervisor: Junhyong Kim, Ph.D.

Life is organized. Organization is largely achieved via composability – that at some level of abstraction, a system consists of smaller parts that serve as building blocks – and modularity – the tendency for these blocks to be independent units that recombine to form functionally different systems. Here, we explore the organization, composition, and modularity of ribonucleic acid (RNA) molecules, biopolymers that adopt three-dimensional structures according to their specific nucleotide sequence. We address three themes: the efficacy of specific sequences to function as modules or as the context in which modules are inserted; the sources of novel modules in modern genomes; and the resolutions at which functionally relevant modules exist in RNA.

First, we investigate the structural modularity of RNA sequences by developing the Self-Containment Index, a method to quantify *in silico* the degree to which RNA structures deviate in changing genomic contexts. We show that although structural modularity is not a general property of natural RNAs, precursor microRNAs are strongly modular, which we hypothesize is a consequence of their unique biogenesis and evolutionary history.

Next, we consider the role of modularity in the regulation of subcellular localization. We identify a novel module, the ID element retrotransposon, contained in the introns of rat neuronal genes, and demonstrate that it is sufficient to drive localization

of mRNAs to dendrites via regulated retention of intron sequence. This mechanism shows that introns can provide the context for functional module insertion, and that transposable elements can be co-opted as source material for these modules. As a further example, we present evidence that a *Camk2a* localization signal can be mimicked by Alu retrotransposon sequence.

Finally, we propose that RNAs can be conceptually decomposed into sets of basic RNA functions. To identify these, we automatically construct an ontology of RNA function using Wikipedia documents. We show that many of the functions encoded in ontology terms are significantly associated with common structural features, highlighting an underlying structure-function relationship that can be encapsulated in elemental RNA building-block units.

In sum, we show how the phenomena of organization, composition, and modularity can frame RNA research in an evolutionary context.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

xviii

CHAPTER 1

INTRODUCTION

Life is organized. On the macro scale, individuals and communities are organized into ecological niches; on the subcellular level, nucleotides are organized into genes, and genes into chromosomes. To the extent that we as biologists study life, we are in some sense attempting to decode how our system of interest – the human genome [1], the MAPK signal transduction pathway [2], the migratory patterns of passerine birds [3], the cell-cycle-coupled oscillation of yeast genes [4] – is organized in space and in time.

Some of the earliest forays into the study of biology as a scientific discipline centered on describing the components of human bodies as well as contextualizing human beings in the natural world. The endeavor of characterizing and understanding human anatomy, the shapes and structures that form a living person, began with ancient medicinal and ritualistic traditions, and rose to prominence with the work of Andreas Vesalius, a 16th-century practitioner of observation-based medicine. In *De humani corporis fabrica*, Vesalius described a system composed of discrete organs and organ groups, each contributing a specific function to the organism as a whole [5]. By the nineteenth century, a theory emerged, due to Theodor Schwann and Matthias Jacob Schleiden, stating that all living organisms and their parts were composed of elemen-

tal units called cells [6]; subsequently, Rudolf Virchow added a temporal dimension, observing that all cells are produced from pre-existing cells, thus forming the basis for understanding developmental phenomena such as pattern formation, which we now know to be an organized and coordinated process [7].

On the macro scale, the study of the organizing principles of life date back to at least the fourth century B.C., when Aristotle formulated the *scala naturae* (the "Great Chain of Being"), a taxonomy of living things organized into an eleven-tiered hierarchy. Plants occupied the lowest rung of life, while humans were at the top – "infected," as it were, with a maximal degree of "potentiality" as reflected in their birth forms: warm, wet, and live offspring, in contrast to cold eggs or seeds [8]. Versions of this linear hierarchy prevailed for 2000 years, until 1735, when Carl Linnaeus proposed a rank-based classification system in *Systema Naturae* [9], in which natural entities were divided into three parallel kingdoms – animals, plants, and minerals – which each were subdivided into groups of increasing specificity – class, order, genus, species, and variety. For example, lions (Leo) constituted a separate genus from tigers (Tigris) but belonged to the same order (Ferae) and by extension, the same class and kingdom (Quadrupedia, and Animalia, respectively) [10]. Under the Linnaean taxonomy, which became the foundation for modern systematics, the organization of living things began to look more like a tree than a single chain.

As the formal discipline of life science matured, revealing an ever-increasing complexity of natural systems, certain patterns became apparent. To a large extent we observe that organization is achieved via composability – that at some level of abstraction, a system consists of a series of smaller parts that serve as fundamental building blocks. The cells that comprise organisms are themselves composed of smaller units called organelles, and the instruction set from which these components derive consists of particular combinations of four biomolecular building blocks – the DNA nucleotides

adenine, guanine, cytosine, and thymine. Both functionally and conceptually, these units are separate entities that interact and interface with one another, contributing to a system of hierarchical complexity.

Intimately related, we also observe a phenomenon of modularity, the tendency for these building blocks to be independent units that can be reused or recombined to form functionally different systems. The DNA instructions that encode a protein component are contained in a single, discrete gene; changes in that gene result in localized changes to that specific product. Multiple, slightly different copies of genes can occur throughout a genome, the result of duplication events that create new functional products that do not need to evolve *de novo*. The same gene product can be produced in different cell types or in different developmental stages, resulting in subtly or profoundly different effects depending on the other gene products present at that place and time.

Characterizations of biological modularity come predominantly from studies in morphology, development, and evolution [11], recapitulating similar themes of reuse and independence of modules. Early on, studies of variation in biological subfields such as systematics, paleontology, and comparative anatomy revealed the existence of common forms and common components in the body plans of diverse organisms [12, 13]. Mammalian forelimbs, for example, constitute modules in the sense that they are largely operationally independent from the rest of the body. This is manifest both in the maintenance of the global organismal form when a forelimb is removed, and in the diversity of shapes among different species – wings in bats, flippers in whales – that reflect localized evolutionary variation.

Developmental modularity occurs on a more mechanistic level and concerns the realization of localized components of these body plans through the coordinated expression of a discrete set of genes [14]. The classic examples are the developmental

modules that determine body segmentation and axis polarity in arthropods such as Drosophila [15]. Arthropod bodies consist of a series of repeated segments, each of which is patterned using the same regulatory network of genes [16] that establish direction locally specific to the particular segment. In this way, the same regulatory plan causes legs to develop in the thoracic segments but antennae to develop in the head, as a result of different downstream gene effects in different segments. This phenomenon is dramatically illustrated in the induced ectopic formation of legs in place of antennae when the gene *Antennapedia* is expressed in the head segment [17], revealing both the positional independence of segment identity and the developmental homology between different segments.

Evolutionary modularity arises as a consequence of natural selection, the fundamental process driving evolutionary change that was first described by Charles Darwin in 1859 [18]. Natural selection occurs when individual variation of specific traits (phenotypes) causes fitness differentials, such that some individuals are better adapted to surviving to reproduce and pass their genetic material to the next generation. Subsequent generations will preferentially be composed of gene sequences (genotypes) that conferred the fitness advantage in the previous generation.

Richard Lewontin argued that adaptation is possible only if the genotypes that lead to variable reproductive fitness are "quasi-independent," meaning that particular genes should be able to change in response to selection without inducing side effects in many other genes [19]. John Bonner expanded this argument by invoking the existence of modular "gene nets" – groups of genes that are highly interdependent within the group (i.e., pleiotropic) but relatively independent of genes outside of the group [20]. The effects of genetic changes are then localized to a small subset of the entire genome, such that incremental changes, as necessitated by evolution, tend not to have systemic effects that would result in an overall decrease in fitness

4

in the organism. Robert Brandon refines this further by proposing that a gene net must possess "unitary function" in order to be an evolutionary module, such that the genotype of a module maps exclusively onto a single phenotype or set of phenotypes that constitute a trait upon which selection can act [21]. For example, the overall spot patterns on the wing of a Viceroy butterfly constitute a unitary function, as selection acts on the ability of the viceroy to mimic a Monarch butterfly; however, the color of an individual spot does not constitute a unitary function, as the selection advantage has little meaning at that level of resolution [22]. Thus, modularity emerges as a phenomenon that defines discrete units of evolution.

The confluence of these three forms of modularity occurs in the holistic "evo-devo" treatment of development and form in the context of evolution [23], and relies on a mapping of morphological modules onto developmental modules, which in turn map to discrete evolutionary modules [24, 25, 26]. In practice, this mapping is not always clear [27]; however, the close coupling of genotype and phenotype suggests that modularities defined at various levels of abstraction will interact in the form of constraints on the space of possible forms of organization.

This dissertation explores the properties of organization, composition, and modularity as pertaining to ribonucleic acid (RNA) molecules, biopolymers occurring ubiquitously among all living cells and essential for life function. RNA is composed of a linear sequence of basic units called nucleotides, whose pairwise energy-minimizing interactions (base pairs) confer the RNA with a three-dimensional folded structure (for a detailed discussion of RNA biology, see Chapter 2). In turn, the function of an RNA molecule is a direct consequence of the structure that it adopts [28], resulting in a variety of roles as an information carrier, a catalytic species, or a substrate for chemical reactions or biomolecular-complex assembly. As a result, RNA can serve as a basic model for studying the genotype-phenotype relationship [29, 30], since

we can treat RNA structure – which at a basic level is a computationally tractable characteristic – as a proxy for phenotype.

The principles of organization and composition in RNA biology are apparent in many forms. RNA genes, like protein-coding genes, can exist in multi-copy families, in which each version shares a core functionality but is also specialized in some way. For example, most organisms contain hundreds of copies of transfer RNA (tRNA) genes, each containing the same structural elements, but differing only in a few nucleotides according to the specific amino acid substrate they bind [31]. Similarly, combinatorial regulation is a common theme in RNA biology, in which the aggregate effects of different RNA species combine in different ways serves to bring about a specific regulatory plan; this phenomenon is exemplified in microRNA-mediated regulation, in which target transcripts contain multiple copies of different microRNA-recognition sequences [32, 33]. In RNA structures, repeated patterns of base configurations have been well characterized by both biophysicists [34, 35] and computational biologists [36, 37], suggesting the existence of a higher-order code that dictates the composition of natural RNAs.

The modular properties of RNA structures was the topic of a seminal paper by Ancel and Fontana in 2000 [38]. Using the base-paired structure of RNA as a model for phenotype, Ancel and Fontana used computational simulations to characterize the plasticity of individual RNAs, defined as the propensity for an RNA sequence to adopt several different thermodynamically-favorable structures, as opposed to a single, stable structure with high probability. High plasticity means low specificity of shape, which is seen as a negative consequence if the function of the RNA depends on a specific conformation. Thus, there is evolutionary pressure to reduce plasticity and canalize particular configurations, which according to their simulations occurs when the individual structured components in the global RNA structure have a high degree

6

of independent thermodynamic stability – i.e., modularity. Modular components in a single RNA tend to fold autonomously from each other, they remain intact over a broad range of temperatures, and they are structurally insensitive to genetic context.

From a bottom-up perspective, modularity in RNA components has particular relevance under a model of RNA evolution in which primordial RNA fragments with specific catalytic abilities combined together to form molecules of increasing complexity [39, 40]. *In vitro* selection experiments, in which artificial RNA species are evolved to perform a specific biochemical process, have shown the feasibility of constructing catalytic RNAs using random sequence ligation and shuffling [41]. Simulations by Manrubia and Briones [42] showed that selection for large, complex RNA structures is significantly easier when modular subcomponents are first allowed to evolve separately before combining together, as compared to direct evolution of the larger RNA.

Here, we address three themes in the study of RNA composition and modularity: the efficacy of specific sequences to function either as modules or as the context in which a module is inserted; the sources of novel modules in modern genomes; and the resolutions at which functionally relevant modules exist in RNA.

Chapter 3 addresses the first of these themes. We draw inspiration from Ancel and Fontana and investigate the structural modularity of natural RNAs in the context of changing genomic sequence, using the "self-containment index" to measure the degree of intrinsic structural robustness that an RNA possesses. We find that although structural modularity is not a general property of most natural RNAs; precursor microRNAs do exhibit an extremely high degree of modularity, which we hypothesize is a consequence of unique biogenesis constraints.

Chapter 4 looks at modularity in the specific context of transcript localization. We identify a functional module, the ID element retrotransposon, that is contained in the introns of several neuronally expressed rat genes whose messenger RNA transcripts are

transported to the dendrite compartment. We show that this ID element structure is sufficient to drive the localization phenotype, via the regulated retention of intronic sequence. Our findings show that introns can provide the context for functional module insertion and that transposable elements can be co-opted to serve as the source material for functional RNA modules. As a further example of the relationship between transposable elements and localization modules, we show that the *Camk2a* localization motif can be mimicked by Alu retrotransposon sequence, suggesting that Alus may also be co-opted as functional modules.

Finally, in Chapter 5 we investigate the functional components from which single RNA molecules are constructed. We propose that an RNA can be conceptually decomposed into a set of basic RNA functions, each of which is shared by diverse classes of RNAs. The specificity of any one RNA is determined by the particular combination of functions. To identify these functions, we use information-extraction techniques to construct an RNA-function ontology, such that basic RNA functions are represented by individual ontology terms. We show that RNA classes that are annotated with similar terms contain similar structural components, recapitulating the structure-function relationship in RNAs and reflecting the existence of a common repertoire of functionally-relevant building blocks that span a diverse set of natural RNA structures.

We hope to show that the themes of organization, composition, and modularity are useful ways to frame ongoing research in RNA biology and to understand new observations and discoveries in the the context of modular RNA evolution.

References

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

[2] Fanger GR (1999) Regulation of the MAPK family members: role of subcellular localization and architectural organization. Histol Histopathol 14:887–94.

[3] Prez-Tris J, Bensch S, Carbonell R, Helbig AJ, Tellera JL, et al. (2004) Historical diversification of migration patterns in a passerine bird. Evolution 58:1819–1832.

[4] Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science 310:1152–8.

[5] Vesalius A, Park K (1964) De humani corporis fabrica. Culture et Civilisation, Bruxelles.

[6] Magner L (2002) A history of the life sciences. CRC.

[7] Hazen RM (2009) The emergence of patterning in lifes origin and evolution. Int J Dev Biol 53:683–92.

[8] Mayr E (1982) The growth of biological thought. Belknap Press Cambridge, MA.

[9] Linnaeus C (1766) Systema naturae.

[10] McKenna MC, Bell SK (1997) Classification of Mammals above the species level. Columbia University Press.

[11] Callebaut W (2005) The ubiquity of modularity. In: Callebaut W, Rasskin-Gutman D, editors, Modularity: Understanding the Development and Evolution of Natural Complex Systems, Cambridge, MA: The MIT Press, The Vienna Series in Theoretical Biology. pp. 3–28.

[12] Riedl R (1978) Order in living organisms: a systems analysis of evolution. John Wiley & Sons.

[13] Eble GJ (2005) Morphological modularity and macroevolution: Conceptual and empirical aspects. In: Callebaut W, Rasskin-Gutman D, editors, Modularity: Understanding the Development and Evolution of Natural Complex Systems, Cambridge, MA: The MIT Press, The Vienna Series in Theoretical Biology. pp. 221–238.

[14] Raff R (1996) The shape of life: genes, development, and the evolution of animal form. University of Chicago Press.

[15] Hyman L (1940) The invertebrates: protozoa through ctenophora.

[16] Gerhart J, Kirschner M (1997) Cells, embryos, and evolution: Toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Blackwell Science Malden, MA.

[17] Postlethwait JH, Schneiderman HA (1969) A clonal analysis of determination in Antennapedia a homoeotic mutant of Drosophila melanogaster. Proc Natl Acad Sci U S A 64:176–83.

[18] Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. New York: D Appleton .

[19] Lewontin RC (1978) Adaptation. Sci Am 239:212–8, 220, 222 passim.

[20] Bonner J (1988) The evolution of complexity by means of natural selection. Princeton Univ Pr.

[21] N BR (1999) The units of selection revisited: The modules of selection. Biol Philos 14:167–180.

[22] Brandon RN (2005) Evolutionary modules: Conceptual analyses and empirical hypotheses. In: Callebaut W, Rasskin-Gutman D, editors, Modularity: Understanding the Development and Evolution of Natural Complex Systems, Cambridge, MA: The MIT Press, The Vienna Series in Theoretical Biology. pp. 51–60.

[23] Roush W, Pennisi E (1997) Growing pains: evo-devo researchers straddle cultures. Science 277:38–9.

[24] Wagner G (1996) Homologues, natural kinds and the evolution of modularity 1. Integrative and Comparative Biology 36:36–43.

[25] Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–31.

[26] Kuratani S (2009) Modularity, comparative embryology and evo-devo: developmental dissection of evolving body plans. Dev Biol 332:61–9.

[27] Williams TA, Nagy LM (2001) Developmental modularity and the evolutionary diversification of arthropod limbs. J Exp Zool 291:241–57.

[28] Tinoco IJ, Bustamante C (1999) How RNA folds. J Mol Biol 293:271–81.

[29] Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. Science 280:1451–5.

[30] Fontana W (2002) Modelling 'evo-devo' with RNA. Bioessays 24:1164–77.

[31] Beuning PJ, Musier-Forsyth K (1999) Transfer RNA recognition by aminoacyl-tRNA synthetases. Biopolymers 52:1–28.

[32] Doench JG, Sharp PA (2004) Specificity of microRNA target selection in translational repression. Genes Dev 18:504–11.

[33] Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, et al. (2005) Combinatorial microRNA target predictions. Nat Genet 37:495–500.

[34] Leontis NB, Westhof E (2003) Analysis of RNA motifs. Curr Opin Struct Biol 13:300–8.

[35] Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38:221–43.

[36] Kim N, Shiffeldrim N, Gan HH, Schlick T (2004) Candidates for novel RNA topologies. J Mol Biol 341:1129–44.

[37] Pasquali S, Gan HH, Schlick T (2005) Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. Nucleic Acids Res 33:1384–98.

[38] Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. J Exp Zool 288:242–83.

[39] Gilbert W (1986) Origin of life: The RNA world. Nature 319:618.

[40] Joyce GF (2002) The antiquity of RNA-based evolution. Nature 418:214–21.

[41] Burke DH, Willis JH (1998) Recombination, RNA evolution, and bifunctional RNA molecules isolated through chimeric SELEX. RNA 4:1165–75.

[42] Manrubia SC, Briones C (2007) Modular evolution and increase of functional complexity in replicating RNA molecules. RNA 13:97–107.

Chapter 2

RNA biology

There are three fundamental information-carrying molecules in biology: DNA, RNA, and proteins. All three are polymers composed of different combinations of alphabetic monomers. DNA is the genetic blueprint, the instructions that encode all of life's functions. Proteins are effectors: the instructions they carry, which are specified by DNA, explicitly define what they look like and how they interact with other molecules around them. RNA falls somewhere in the middle, literally and figuratively. RNA's traditionally defined role is as a messenger, an intermediate created from DNA genes for the purpose of generating protein products.

Not content to remain relegated to "middle-child" status, RNA proved to be a much more versatile molecule, a fact that biologists slowly became aware of, starting with the discovery of the first non-intermediate-messenger RNAs, transfer RNA (tRNA) and ribosomal RNA (rRNA); through the 1980s when the first example of RNA catalytic activity was discovered [1]; and into the 1990s and 2000s, when the number of characterized functionally distinct RNA classes exploded. The role of RNA as something other than a middleman is increasingly less seen as an exception; taken to the extreme, one might argue that RNA can do everything that DNA and proteins do, albeit in a limited fashion. This is in fact the basis for the "RNA world

hypothesis" formulated by Walter Gilbert [2], which suggests that ancient biological processes were carried out solely by RNA molecules that served both as the information carriers and as the effectors. In this way, RNA is actually the eldest sibling, and over evolutionary time DNA and proteins evolved to adopt and specialize roles previously assumed by RNA alone.

In this chapter, we will explore the role of RNA in all aspects of molecular biology. Section 2.1 summarizes these roles framed in terms of the phenomenon of directional information transfer. Section 2.2 describes the physical properties of RNA molecules and illustrates the importance of sequence and structure. Section 2.3 describes several of the major classes of RNAs that have been characterized. Section 2.4 reviews the experimental and computational techniques commonly used to elucidate RNA structure. Finally, Section 2.5 describes techniques to identify and quantify RNAs.

2.1 The role of RNA in the Central Dogma of molecular biology

In 1958, shortly after he and James Watson presented their model for the structure of the DNA double helix, Francis Crick proposed that the relationship between the three fundamental molecules of life – deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins – conformed to a constrained model of information flow that he called "The Central Dogma" [3]. The theory states that information can be passed from nucleic acid (DNA, RNA) to protein and from nucleic acid to nucleic acid, but never protein to protein or protein to nucleic acid (Figure 2.1).

Under this model, biological information is encoded as sequence, words formed from a four-letter nucleotide alphabet in the case of DNA and RNA and a twenty-letter amino-acid alphabet for proteins (Figure 2.2). Different combinations of letters
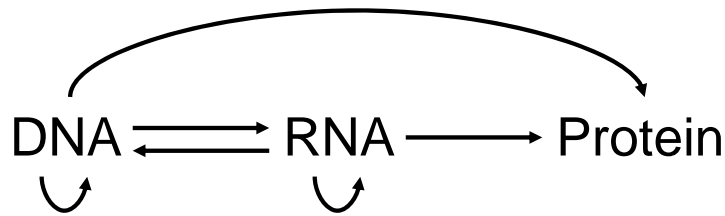
Figure 2.1: Diagram showing flow of information between biomolecules according to Crick's Central Dogma. Arrows indicate possible directions of flow.

encode different instructions, which lead to the production of the various molecular products that carry out life functions.

DNA		
A	Adenine	
C	Cytosine	
G	Guanine	
T	Thymine	

RNA		
A	Adenine	
C	Cytosine	
G	Guanine	
U	Uracil	

Protein				
A	Alanine	M	Methionine
C	Cysteine	N	Asparagine
D	Aspartate	P	Proline
E	Glutamate	Q	Glutamine
F	Phenylalanine	R	Arginine
G	Glycine	S	Serine
H	Histidine	T	Threonine
I	Isoleucine	V	Valine
K	Lysine	W	Tryptophan
L	Leucine	Y	Tyrosine

Figure 2.2: The three genetic alphabets

2.1.1 RNA AS A MESSAGE-CARRYING INTERMEDIATE

The specific nucleotide sequence of DNA gives rise to a specific nucleotide sequence of RNA in a process called transcription. The DNA in a cell is organized into one or several chromosomes as double helices, pairs of long strings of nucleotide sequence that

together constitute an organism's genome. Discrete regions of these chromosomes are defined as genes – individual units of inheritance and functionality, and the sites of transcription. When these genes are bound by an enzyme called RNA polymerase in concert with other associated proteins called transcription factors, transcription initiation is triggered, and a nascent RNA molecule is created. During the elongation phase, the DNA double helix is unwound and RNA polymerase moves forward along one of the DNA strands (the so-called "template" strand) and assembles an RNA molecule by ligating together ribonucleotides according to the sequence specified by the DNA template – i.e., for every nucleotide in the DNA strand, a corresponding complementary nucleotide is added to the RNA. A specific nucleotide signal in the DNA causes termination of this process, allowing the new single-stranded RNA molecule to dissociate.

Proteins are in turn created using the information encoded in an RNA molecule in a process called translation, so named because the nucleotides of the RNA are "translated" into an amino-acid alphabet according to specific rules. As there are 20 different amino acids and only four different nucleotides, a three-nucleotide combination – e.g., ACG or GCC – is required to encode all of the amino acids, albeit redundantly ($4^3 = 64$ different three-nucleotide words). Each of these nucleotide triples is called a codon, and there are 61 different codons that encode amino acids, with three additional "stop" codons specifying a termination signal (Figure 2.3). This so-called genetic code is mostly conserved across all organisms, though slight variations exist.

In eukaryotes (all animals, plants, fungi, and some additional single-cell organisms), transcription can be followed by a post-processing step called splicing in which segments of the RNA called introns are removed. Introns range in length from a few nucleotides to thousands and always occur between exons; thus conceptually an RNA is composed of an [exon, (intron, exon)*] pattern – i.e., an RNA sequence always

AAA	→	K	ACA	→	T	AGA	→	R	AUA	→	I
AAC	→	N	ACC	→	T	AGC	→	S	AUC	→	I
AAG	→	K	ACG	→	T	AGG	→	R	AUG	→	M
AAU	→	N	ACU	→	T	AGU	→	S	AUU	→	I
CAA	→	Q	CCA	→	P	CGA	→	R	CUA	→	L
CAC	→	H	CCC	→	P	CGC	→	R	CUC	→	L
CAG	→	Q	CCG	→	P	CGG	→	R	CUG	→	L
CAU	→	H	CCU	→	P	CGU	→	R	CUU	→	L
GAA	→	E	GCA	→	A	GGA	→	G	GUA	→	V
GAC	→	D	GCC	→	A	GGC	→	G	GUC	→	V
GAG	→	E	GCG	→	A	GGG	→	G	GUG	→	V
GAU	→	D	GCU	→	A	GGU	→	G	GUU	→	V
UAA	→	*	UCA	→	S	UGA	→	*	UUA	→	L
UAC	→	Y	UCC	→	S	UGC	→	C	UUC	→	F
UAG	→	*	UCG	→	S	UGG	→	W	UUG	→	L
UAU	→	Y	UCU	→	S	UGU	→	C	UUU	→	F

Figure 2.3: The genetic code showing the mapping of nucleotide codons to amino acids. (*) indicates a stop codon.

starts with an exon, followed by zero or more intron-exon pairs. The spliceosome, a complex composed of catalytic RNA and protein subunits, carries out the splicing process by binding each intron at opposite ends, cleaving at the intron-exon boundary, and ligating the free exons together.

The RNA that contains the coded protein sequence is called a messenger RNA (mRNA). Prior to translation, the mRNA is bound by the components of the ribosome; in bacteria, this can happen while the mRNA is still being transcribed, but in eukaryotes, transcription and translation do not overlap since transcription occurs in the nucleus where the chromosomes are sequestered from the rest of the cell, while translation occurs outside the nucleus after the mRNA is exported. The ribosome positions the mRNA at the beginning of its sequence code, then catalyzes the decod-

ing by processing one codon (three nucleotides) at a time, recruiting a specific amino acid corresponding to the code in Figure 2.3, and creating bonds between consecutive amino acids, forming a polypeptide chain. This process ends when a stop codon is encountered. The amino acids are carried to the ribosome by tRNAs that are specific to specific codons. Where the ribosome begins reading the mRNA code naturally affects the resulting protein sequence – e.g., shifting the code to the right by one nucleotide would result in a completely different codon sequence. Thus to ensure the mRNA is read "in frame," the ribosome recognizes sequence signals in the mRNA, such that the code always begins with a specific "AUG" codon.

2.1.2   RNA in other roles

The processes of transcription and translation constitute the canonical modes of genetic information processing in the cell and paint a picture of RNA as a sort of molecular middle man. As the field of biology matured, the role of RNA became increasingly more varied and nuanced with the discovery of mechanisms of information transfer that de-emphasize the role of either the canonical starting point (DNA) or end point (protein).

Early on, viruses were identified that contained genomes composed of RNA rather than DNA. For many types of viruses (e.g., the common cold, Hepatitis A, SARS), genomic RNA is used directly to create protein products, bypassing transcription. For other types (e.g., Influenza, measles, Ebola), the RNA genes are transcribed as though they were DNA using an RNA-specific RNA polymerase, and the resulting mRNAs are subject to translation. Retroviruses such as HIV-1 use both the transcription and translation machinery of their host cell by inserting their genetic material directly into the host genome. Using an enzyme called reverse transcriptase that is encoded by the viral genome, RNA genes are converted into DNA and then integrated into one of

the host's chromosomes, where they are transcribed as though they were endogenous host genes. Thus the flow of information is RNA→DNA→RNA→protein.

The existence of RNA molecules whose purpose was not to serve as an intermediate messenger for protein production, but rather to serve as end products, was also recognized shortly after the Central Dogma was proposed, with the first tRNA primary structure characterized in 1965 by Richard Holley [4]. tRNAs play an essential role in protein translation by serving as the decoder between nucleotide sequence and amino acids, a function facilitated by their three-dimensional shape. Each tRNA specifically binds one type of amino acid on one end of the molecule. On the opposite end, the tRNA contains an anticodon, a three-nucleotide sequence that is complementary to a specific codon – this is how specificity of the codon-amino acid mapping is achieved. During translation, the anticodon on the tRNA and the codon on the mRNA form a temporary interaction in the ribosome, allowing the correct amino acid to be presented and added to the growing polypeptide. The ribosome itself is also composed of several ribosomal RNAs (rRNAs). The ribosome consists of two protein-RNA hybrid subunits, with the small subunit containing one rRNA (16S rRNA in bacteria, 18S in eukaryotes), and the large subunit containing two or three different rRNAs (5S and 23S in bacteria; 5S, 5.8S, and 28S in most eukaryotes).

Within the last decade, our understanding of the cellular repertoire of these so called non-coding RNAs has increased dramatically, suggesting that the non-protein-including DNA→RNA flow of information should not be relegated to special-case status.

The flow of information from RNA to RNA in isolation can also occur – i.e., without a protein end product. In most higher organisms, a transcript-silencing mechanism called RNA interference (RNAi) occurs when double-stranded RNA (dsRNA) is found in the cytoplasm. Since endogenous cytoplasmic RNA is normally single

stranded, the presence of dsRNA triggers an inactivation pathway that prevents the RNA from being translated. This process is believed to be similar to an immune response, since invading pathogens can contain dsRNA [5, 6], but endogenously coded dsRNA can also trigger the response, as a means of translation regulation. In higher eukaryotes, endogenously-coded short RNAs called microRNAs (miRNAs) are produced that are complementary to nucleotide sequences in specific "target" mRNA transcripts. Specifically in plants, upon binding of the miRNA to the target sequence, a double-stranded region is formed, which induces RNAi-style cleavage of the transcript at the site of binding. In all cases, cleavage of the transcript inactivates it, thus preventing it from being used for protein translation. However, in some instances, cleavage also initiates a cascade in which a new RNA strand is synthesized using the cleaved transcript as a template, by the RNA-dependent RNA polymerase RDR6. The resulting newly double-stranded RNA is then targeted for additional cleavage, which produces small RNA fragments that themselves can bind to other transcripts, targeting them for cleavage [7, 8].

RNAi is one example of epigenetic control, a mode of heritable phenotypic change that is not encoded in DNA. The short silencing RNA effectors of RNAi can be transcribed in one cell, and then transferred during cell division when the contents of the cytoplasm are divided between daughter cells. This is the basis for the phenomena of maternal effects, in which the phenotype of a zygote is influenced by the cytoplasmic contents of the egg, as opposed to being completely determined by only the genetic material contributed by each parent. Another form of epigenetic control is regulation of how particular mRNA transcripts are processed – for example, how they are spliced. Although most genes exhibit a canonical splicing pattern when they are transcribed, sometimes variants are created in which certain exons are excluded or rare additional exons are included [9]. These alternative splice forms exert different phenotypes, and

inheritance of an alternatively spliced mRNA causes these phenotypic effects to be passed on.

2.2   The biophysics of RNA

An RNA molecule is a polymer composed of nucleotide monomers. A nucleotide is composed of a sugar – ribose for RNA in contrast to deoxyribose for DNA; a phosphate group attached to the 5-carbon of the sugar; and a nitrogen-containing base attached to the 1-carbon (Figure 2.4). Bases are either purines (adenine (A) and guanine(G)) or pyrimidines (cytosine (C), thymine (T), and uracil (U)). Both DNA and RNA molecules use nucleotides containing A, C, and G; T is normally found only in DNA and U only in RNA.

To form a polymer, nucleotides are joined together via a phosphodiester bond between the phosphate and the hydroxyl group on the 3-carbon of the sugar. Thus, a nucleotide chain is directional, with the 5′ end denoting the nucleotide with the free phosphate, and the the 3′ end denoting the nucleotide with the free hydroxyl on the 3-carbon (Figure 2.4).

Interactions between nucleotides occur via hydrogen bonding of their bases. Due to their molecular geometries, the most typical interactions are one adenine paired with one thymine or uracil; or one cytosine paired with one guanine. These are known as Watson-Crick base pairs. Base pairing occurs in an anti-parallel fashion, such that the base-paired nucleotides are in opposite orientations with respect to their 5′ phosphates and 3′ hydroxyls. It is these interactions that allow the formation of the DNA double helix – two separate DNA molecules whose nucleotide sequences are exactly complementary, but due to the orientation rules, are reversed with respect to one another. For example, a DNA sequence of ACTGG would base pair to its reverse
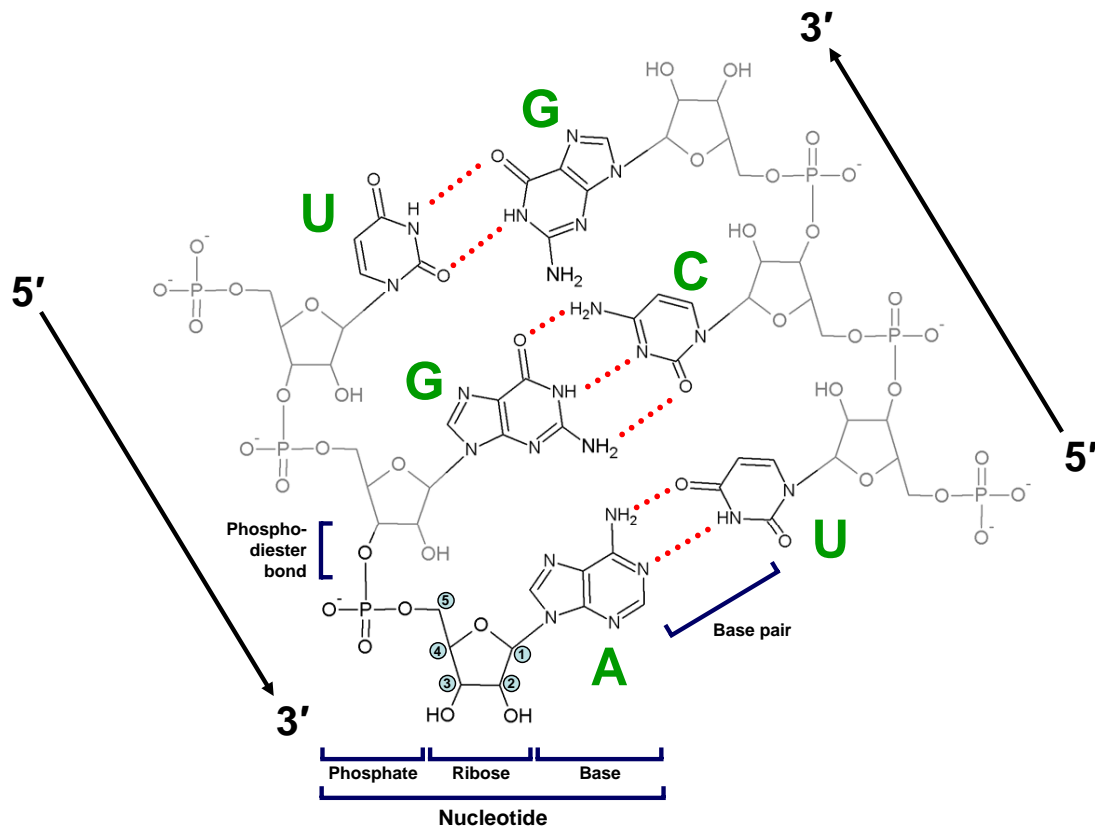
Figure 2.4: Molecular diagram of RNA nucleotides. Two base-paired strands of RNA are shown. Base types are labeled in green, and hydrogen bond interactions are shown as red dotted lines. For a reference ribose sugar, the carbon atoms are labeled 1 through 5 according to standard conventions. Structure drawing was done using ACD/ChemSketch.

complement of CCAGT:

$$5'\text{-ACTGG-}3'$$

$$3'\text{-TGACC-}5'$$

Base pairing causes the molecules involved to be in a lower energy state; thus, it is favorable for nucleotide strands to base pair when possible.

Base pairs can also form between nucleotides on a single strand, which is generally

the case for RNA molecules. Because number of base pairs correlates with energy minimization, a strand of RNA will tend to maximize the number of sterically-favorable base pairs it forms by adopting a compact folded structure to allow bases to interact with each other. The base pairing pattern of an RNA constitutes its secondary structure and is a direct consequence of the RNA nucleotide sequence, which is commonly called the primary structure. Long stretches of base-paired nucleotides confer the RNA with a helical shape, analogous to the DNA double helix.

Secondary structure follows particular biophysical rules. Bases pair according to the Watson-Crick rules listed above – A with U, C with G – but RNAs commonly also have weaker pairs between G and U, the so called "wobble" base pair, along with several additional minor interactions that can involve three or more bases. Bases can only pair if they are separated by a sufficient distance in the sequence, approximately three nucleotides, since it is energetically unfavorable for the RNA sugar backbone to bend sufficiently to allow very close bases to pair. Also, bases must pair in a nested fashion, such that if we imagine that paired bases are connected by a thread, these threads cannot cross (Figure 2.5). Thus secondary structure exists in a two-dimensional plane.

The convention typically used to represent a secondary structure is a dot-parenthesis notation popularized by the Vienna RNA Package [10], which consists of a string formed from an alphabet of three symbols: "(", ")", and "." The left parenthesis represents a nucleotide that is base paired with some downstream nucleotide, while the right parenthesis represents a nucleotide base paired with an upstream nucleotide. The dot indicates an unpaired base. In this way, the secondary structure of a k-length RNA nucleotide sequence can be unambiguously encoded by a k-length dot-parenthesis sequence (Figure 2.6).

The constraints imposed on secondary structure result in a set of commonly oc-
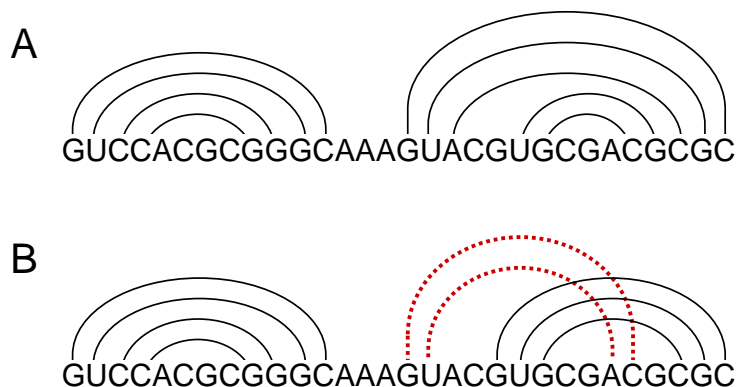
Figure 2.5: Examples of basepairing. Basepairing occurs in a nested fashion. (A) Arcs show basepairing; note that no arcs cross. (B) An illegal secondary structure with crossing arcs.

curring structural elements seen in various RNA structures. These elements are composed of different patterns of base paired nucleotides (stems) and unpaired nucleotides (loops). The most commonly cited element is the hairpin, which consists of a stem terminated on one end by a loop of at least three nucleotides. Loops that occur within a stem are called interior loops, resulting in varying degrees of disruption of the helical shape. Bulges are asymmetric interior loops, resulting in free nucleotides protruding from only one side of the stem. Branch loops occur at the junction of three stems (Figure 2.7).

Higher-ordered structure can exist on top of secondary structure, such that interactions occur outside the plane or between secondary-structure elements; this constitutes the tertiary structure of the RNA, which is a truer three-dimensional representation of the RNA molecule in space than secondary structure. Some tertiary motifs consist of base-pairing interactions that violate the nested base-pairing rules of secondary structure. For example, pseudoknots are formed between nucleotide sequence in a loop and single-strand sequence outside the enclosing stem [11] (Figure 2.8); this is a particularly stable motif due to interactions between the two stem regions [12]. A

```
5'-GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGA-
  ((((((((..(((((........)))).(((((.......

 -UCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA-3'
  )))))....(((((.......))))))))))))))....
```
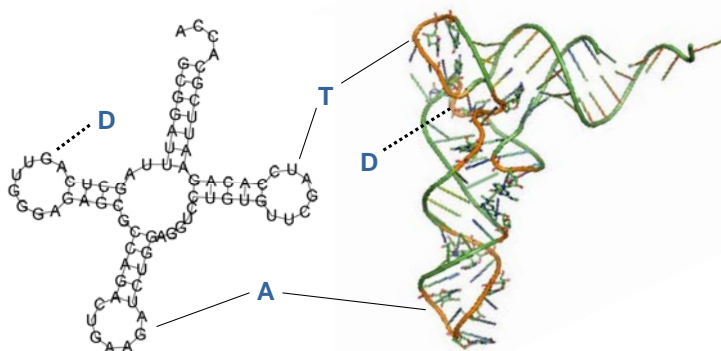
Figure 2.6: Primary, secondary, and tertiary structures of yeast phenylalanine tRNA. Structural conformation was derived from X-ray crystallographic data PDB:1EHZ [15] and visualized using RNAplot and PyMol. Corresponding loops in the secondary and tertiary structures are labeled (D, T, and A).

similar motif is the kissing hairpin, which involves base pairing between two hairpin loops [13]. Many other tertiary motifs do not involve canonical base pairing, such as the G quadruplex, which consists of "Hoogsteen" base interactions between quartets of guanines arranged in a square [14]; and the D-Loop:T-Loop interactions in tRNAs, which are illustrated in the yeast tRNA tertiary structure shown in Figure 2.6.

The net result of these nucleotide interactions is a molecule with specific sequence and shape properties, whose function will directly follow from these characteristics. Some of these functions were alluded to in the previous section; in the next section, we will explicitly delineate the broad classes of functionality and describe some of the major types of RNA.
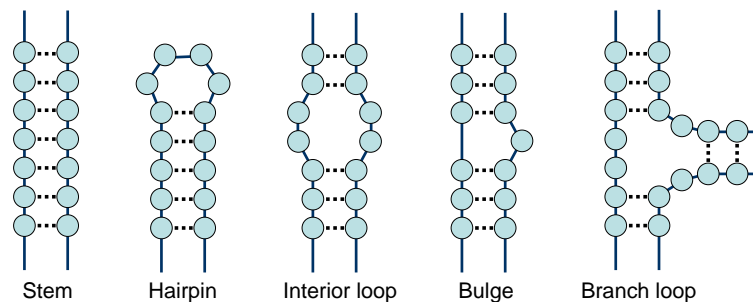
Figure 2.7: RNA structure elements. Nodes represent nucleotides, solid lines are covalent bonds, dotted lines are hydrogen bond interactions.
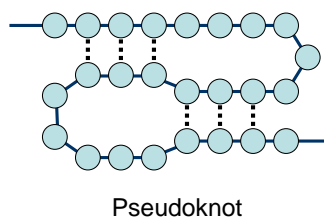
Figure 2.8: RNA pseudoknot example. Nodes represent nucleotides, solid lines are covalent bonds, dotted lines are hydrogen bond interactions.

2.3 FUNCTIONAL CLASSES OF RNA

At any given moment, a living cell contains on the order of 10^6-10^8 individual RNA molecules [16], the sum of which constitutes the cell's transcriptome. A transcriptome is dynamic, with new RNAs being transcribed as others are being degraded. It consists of a variety of different species of RNA, ranging in length from tens of nucleotides to several thousands. Some RNA will exist in single copy, while others will have thousands or more.

As we saw in Section 2.1, we can distinguish between two broad classes of RNA – those that code for protein products (mRNAs); and everything else, which are collectively called noncoding RNA (ncRNA). This is a useful dichotomy conceptually,

but the elevated status it confers to mRNAs compared to the rest of the crowd may be a bit misleading.

We can think of the functionality of an RNA molecule as arising from a unique combination of nucleotide sequence and structured components. Nucleotide sequence is well suited for recognition, by virtue of the base-pairing specificity that is fundamental to nucleotide biochemistry. Nucleotide structure, similar to protein structure, confers a range of biological function, which can be broken down into three general categories: catalysis – i.e., enzymatic facilitation of chemical reactions on biomolecules; interaction with other biomolecules, particularly proteins, to form complexes or to serve as substrates; and scaffolding, providing a platform on which biological processes can occur. More often that not, the sequence and structural components that give rise to these four basic functions are difficult to separate; however, we can think of any RNA molecule as consisting of a mixture of components that contribute to an approximate ratio of these functions. In this sense, RNA classes can be mapped onto a three-dimensional simplex – a tetrahedron, where each of the vertices represents one of the basic functions. In such a subspace, RNAs that map close to one of the vertices are predominantly composed of sequence and structure directed toward that specific function. RNAs that map to the interior are composed of several components that encompass multiple functions. A visualization of this map is given in Figure 2.9.

Family groupings of RNAs consist of individual RNA genes that all share a common mixture of sequence, structure, and functionality, for the purpose of carrying out a specific biological process. RNAs differ in terms of their biogenesis pathways, their cellular location, and whether they exist autonomously or in complex with proteins or other RNAs. Despite such a wide range of structure and function, commonalities exist, some arising from obvious evolutionary relationships, others not. Biologists have yet to fully characterize the RNA repertoire, and assuming this is even possible,
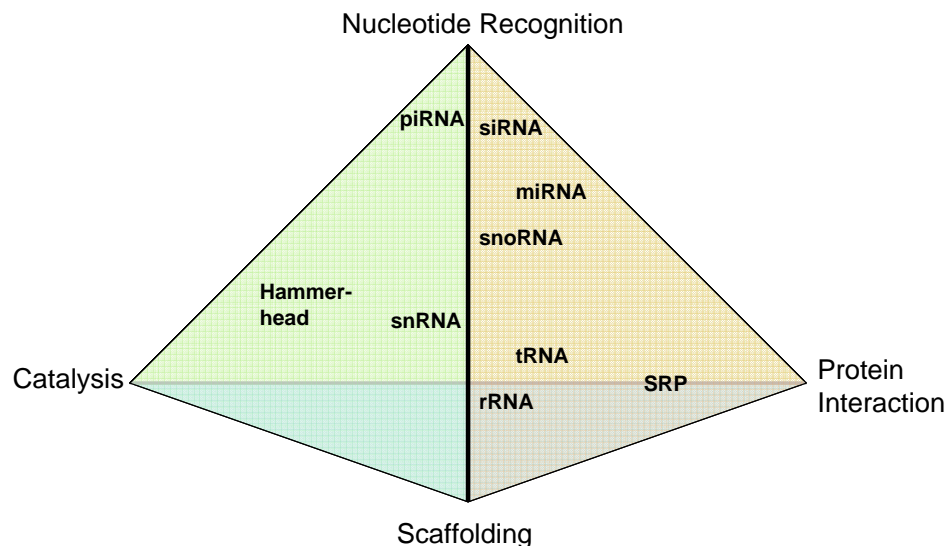
Figure 2.9: A simplex of RNA function. The figure is a tetrahedron with each vertex corresponding to one of four basic RNA functions. Points within the simplex represent classes of RNAs, showing the relative degree to which each RNA carries out these functions. For example, RNAs close to one of the vertices predominantly perform the function defined by that vertex, while RNAs in the interior can be characterized by a mixture of functions.

it will be interesting to reconstruct how such a diversity of function evolved.

2.3.1 RNAs involved in protein translation

Messenger RNA

The primary role of mRNA, as described above, is to serve as an information intermediate between DNA and proteins. Why proteins are not produced directly from DNA does not have a clear answer; one prevailing hypothesis is that the RNA→protein production pathway is more ancient, and that the DNA step evolved subsequently [2]. Regardless, splitting the protein production pathway into two discrete phases does allow for more fine-grained regulation targeting different points.

As is the case for all RNAs, mRNA transcription is regulated, depending on many

Table 2.1: RNA abbreviations found in the literature

Abbrev.	Name
aRNA	antisense RNA
cRNA	complementary RNA
dsRNA	double-stranded RNA
endo-siRNA	endogenous short-interfering RNA
gRNA	guide RNA
lincRNA	large intervening noncoding RNA
lncRNA	large/long noncoding RNA
mRNA	messenger RNA
mtRNA	mitochondrial RNA
nat-RNA	natural antisense transcript RNA (NAT)
ncRNA	non-coding RNA
nmRNA	non-messenger RNA
piRNA	Piwi-associated RNA
rRNA	ribosomal RNA
rasiRNA	repeat-associated small interfering RNA
sRNA	small RNA
scaRNA	small Cajal-body-specific RNA
shRNA	short hairpin RNA
siRNA	small interfering RNA
smRNA	small modulatory RNA
sncRNA	small noncoding RNA
snmRNA	small non-messenger RNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
ssRNA	single-stranded RNA
stRNA	small temporal RNA (= miRNA)
tRNA	transfer RNA
tasiRNA	*trans*-acting small interfering RNA
tmRNA	transfer messenger RNA
vRNA	vault RNA

factors. Foremost is the availability of transcription factors (TFs); these are proteins that bind promoters, specific loci on the chromosome upstream of genes that activate a site for transcription and recruit components of the transcriptional machinery. Production of TFs is itself a regulated process, which can depend on external stimuli to the cell, and generally corresponds to a requirement to produce more (or less) of a given gene product. Accessibility of the binding sites is also regulated, depending on an open chromatin state on the chromosome that is determined by how tightly the DNA molecule is wound around its protein spools (histones) [17].

Transcription of mRNA is carried out by RNA polymerase II (Pol II) in eukaryotes, which produces an RNA strand in the $5'$ to $3'$ direction using the $3'$ to $5'$ DNA template strand of the gene as a reference – i.e., the RNA strand is the reverse complement of the template. The resulting transcript is called a pre-mRNA and is subject to three eukaryote-specific post-transcriptional modification steps, some of which can sometimes occur co-transcriptionally. The first is capping the $5'$ end with a modified guanosine nucleotide, which protects the mRNA from certain forms of degradation. Second, the $3'$ end is trimmed, and a series of ~ 200 adenine nucleotides is attached to form the poly(A) tail. This is a unique characteristic of Pol II-transcribed RNAs and can be used to segregate RNA populations experimentally. Finally, splicing of introns occurs in many genes, catalyzed by the spliceosome (see Section 2.3.2), and the exonic portions of the mRNA are ligated together to form the mature mRNA. In many cases, what constitutes an exon or intron is not static, such that the spliced form of the mRNA may differ for different mRNAs produced from the same gene, a phenomenon called alternative splicing. Since the splice pattern affects the final sequence of the mature mRNA, often the protein-coding instructions are changed, resulting in the production of an alternate protein product, though some alternative splice forms preserve protein sequence and instead contain different regulatory

sequences [18]. Following post-processing, the mature mRNA is exported from the nucleus and directed to protein translation pathways or stored.

The final mature mRNA consists of five distinct regions: the $5'$ cap; the $3'$ poly(A) tail; and the nucleotide sequence, which consists of the protein coding sequence flanked upstream and downstream by the $5'$ untranslated region (UTR) and $3'$ UTR, respectively. The UTRs contain regulatory sequence that affect the manner in which the mRNA is used for protein translation.

In bacteria, none of these processing events occurs, as the mRNA produced from transcription is already in a mature state. However, bacterial genes are commonly transcribed polycistronically in an operon, such that two or more functionally associated genes are transcribed as a single, connected mRNA. This coupling facilitates coordinated regulation, as is the case for the oft cited *lac* operon responsible for regulated lactose metabolism in *Escherichia coli*.

As a message carrier with an encoded set of instructions for protein production, the bulk of the mRNA transcript is used for nucleotide recognition, the mechanism of which is described below. However, protein recognition is also a large component in the function of an mRNA. A large number of proteins interact with the mRNA transcript throughout its biogenesis and functioning, and as discussed in Section 2.3.4, various structural components in the mRNA contribute to its specific regulation.

Transfer RNA

As the first ncRNA to be identified, tRNAs play an essential role in decoding a nucleotide message into an amino acid sequence. A tRNA is small, approximately 75 nts in length, is transcribed by RNA Polymerase III (Pol III), and adopts a three-dimensional cloverleaf structure. In the acceptor stem, the free $3'$ end of the sequence is covalently bonded to a specific amino acid, a reaction catalyzed by an aminoacyl

tRNA synthase enzyme specific to the tRNA subtype [19]. On the opposite end is the anti-codon loop, a single-stranded region containing the three-nucleotide anticodon sequence that is complementary to a codon in an mRNA that encodes the amino acid cargo. This is the nucleotide recognition component of a tRNA.

The shape of the tRNA allows it to fit into one of three binding pockets in the ribosome during translation, allowing the tRNA to deliver correct amino acids to the site of protein synthesis (see below). Thus a tRNA also functions in a protein/RNA-recognition capacity.

Ribosomal RNA

Four rRNAs (three in bacteria) comprise major components of the ribosome, the site at which protein translation takes place; over half of the ribosome is made up of rRNA [20]. In eukaryotes, transcription of rRNA occurs in the nucleolus substructure of the nucleus. The 18S, 5.8S, and 28S rRNAs are transcribed together as a polycistron by RNA Polymerase I (Pol I) [21, 22] from one of hundreds of copies of a primary rRNA gene, which occur in tandem in so-called nucleolar organizing regions on several different chromosomes. The large transcript is processed, involving nucleotide modification and cleavage facilitated by small nucleolar RNAs (snoRNAs, see Section 2.3.2), resulting in the production of mature rRNAs [23, 24]. These, along with 5S rRNA, which is transcribed by Pol III outside of the nucleolus and transported in, and various protein components, are all assembled into the large and small subunits of the ribosome and exported out of the nucleus and into the cytoplasm.

Ribosomes are the molecular machines that drive translation, a complex coordinated process involving rRNAs as catalysts, interaction partners, and scaffolds [25]. During translation, the small subunit (consisting of the 18S rRNA and about 33 proteins) binds a free mRNA at an AUG start codon and recruits the large subunit (5S,

5.8S, 28S, and about 49 proteins) to form the assembled ribosome. The ribosome contains three binding pockets for tRNAs, the A, P, and E sites. During protein synthesis, a tRNA bound to an amino acid binds at the A (aminoacyl) site; this tRNA will have an anticodon sequence complementary to the in-frame codon on the mRNA. A condensation reaction occurs, joining the new amino acid with the existing polypeptide chain. The ribosome shifts over by one codon frame, causing the tRNA to move into the P (peptidyl) pocket, still bound to the amino acid, which has now been incorporated into the protein. The A pocket is now free to accept the next tRNA. The tRNA in the P pocket is deacylated, severing the bond between it and the polypeptide, and as the ribosome shifts to the next codon, the deacylated tRNA enters the E (exit) site, where it is released in the next step.

2.3.2 RNAs that modify other RNAs

As we alluded to in the previous section, the primary function of some types of RNA is to catalyze modifications of other RNA sequences. In some cases, the RNAs are associated with proteins and form functional complexes, while in others the RNAs act in isolation.

Small nuclear RNAs

Small nuclear RNAs (snRNAs) are eukaryotic RNAs found in the nucleus that occur in conjunction with proteins as small ribonucleoprotein complexes (snRNPs). snRNPs are involved in a variety of nuclear regulatory processes, including intron splicing.

The spliceosome is composed of five snRNAs – U1, U2, U4, U5, and U6 – along with approximately 200 proteins (U3 is a snoRNA, see Section 2.3.2). Each intron consists of recognition sequences, located at the 5' and 3' splice sites where the flanking exons meet the intron, and a branch point site (BPS) located just upstream the 3'

splice site. Assembly of the spliceosome begins at the 5′ splice site of the intron, where U1 snRNP binds the intron via base-pairing interactions. Next, the spliceosomal E complex is formed downstream the BPS, consisting of several RNA-binding proteins, which in turn recruit U2 snRNA and additional proteins to form the spliceosomal A complex. Next, a tri-snRNP consisting of U4/U6 snRNP and U5 snRNP subunits is recruited, which joins all of the components of the spliceosome together as the B complex. Conformational changes involving the release of U1 and U4 and the base pairing of U2 and U6 ultimately result in the catalytically active B* complex. The first splicing step occurs, in which cleavage at the 5′ splice site occurs, resulting in a lariat structure formed between the 5′ -most intronic nucleotide and the BPS nucleotide. At this point, the spliceosome exists as the C complex, further rearrangements occur, then the second catalytic step causes excision of the downstream exon and ligation with the already-freed upstream exon [26].

Some species contain an alternate spliceosome that is specific to a rare class of introns called U12 introns. This minor spliceosome is functionally analogous to the major spliceosome, except the corresponding snRNPs are U11, U12, U4atac/U6atac, and U5 (the only snRNP shared between the two spliceosomes).

Besides those involved in splicing, other snRNAs include the mammalian 7SK RNA, which along with HEXIM1 binds and negatively regulates the protein complex elongation factor P-TEFb, whose role is to activate RNA polymerase II; and telomerase RNA, the RNA component of telomerase, which maintains the length of telomeres, the protein-DNA structures that cap and protect the ends of chromosomes [27].

Sometimes classified as a subtype of snRNAs, small nucleolar RNAs (snRNAs) reside in the nucleolus, the site of ribosome assembly, and as the RNA component of snoRNPs, help to catalyze the base modification of other RNAs in the nucleolus. Base modification is a post-transcriptional regulatory step and is often critical to the maturation of the RNA due to the subtle structural changes that result. There are a large number of annotated snoRNAs [28], which belong to one of two large subfamilies based on the particular chemistry they catalyze. The C/D box snoRNAs catalyze methylation of the $2'$ oxygen on the ribose portion of the specific substrate nucleotide, while the H/ACA box snoRNAs catalyze pseudouridylation conversion of uridines – i.e., the isomerization of a normal uridine into a modified pseudouridine base. Each snoRNA has a specific substrate RNA that is determined by base pairing of the snoRNA sequence to its target; catalysis is carried out by the associated proteins in the snoRNP complex.

The majority of characterized snoRNAs have specificity to rRNAs and tRNAs, the major RNA species in the nucleolus. The pre-rRNA, for example, contains approximately 200 modified bases, each catalyzed by a separate snoRNA [29]. All tRNAs also contain modified bases, a large number of which are created by snoRNAs. A subset of snoRNAs do not actually reside in the nucleolus; these small Cajal-body-specific RNAs (scaRNAs) guide the modification of the spliceosomal RNAs, which occurs in the Cajal body subnuclear organelles [30]. Still others (called "orphan" snoRNAs) have unknown targets, and may function on substrates not in the normal repertoire of snoRNAs [31].

RNase P and RNase MRP RNA

RNase P and RNase MRP are endoribonucleases, a class of enzymes that hydrolyze internal phosphodiester bonds in a ribonucleotide sequence, causing cleavage of an RNA strand into two pieces. Most endoribonucleases are composed exclusively of proteins; however, RNase P and RNase MRP are exceptions, consisting of both catalytic RNA and protein components. The RNA-induced silencing complex (RISC) is another exception (see Section 2.3.3).

RNase P is found throughout all lineages and in its primary capacity functions as a post-transcriptional modifier of tRNA molecules – tRNAs are transcribed with a leading 5′ sequence that is removed upon maturation by the RNase P. RNase P might have a general role in the transcription and processing of several other Pol-III transcribed small RNAs including 5S rRNA, U6 snRNA, and 7SL RNA [32].

RNase MRP (mitochondrial RNA processing) is found only in eukaryotes and plays a role in mitochondrial DNA replication by cleaving the RNA primers used for DNA synthesis. It also has been shown in yeast to cleave the internal transcribed spacer 1 between 18S and 5.8S rRNAs in the large primary rRNA transcript [33, 34].

Autonomous ribozymes

A number of RNAs have distinct, independent catalytic ability and function in roles normally associated with protein enzymes. Accordingly, such RNAs have been called ribozymes (RNA enzymes). Technically any RNA with catalytic function can be considered a ribozyme – 23S rRNA for example, despite residing in complex with several proteins, is in fact independently catalytic, and thus is a ribozyme [35]; similarly, RNase P and MRP are both ribozymes that are complexed with proteins.

One class of autonomous ribozymes catalyzes nucleotide sequence cleavage, com-

prising the hammerhead, hairpin, Hepatitis delta virus (HDV), and Varkud satellite (VS) ribozymes. The best studied of these is the Hammerhead ribozyme, which is found in plant viruses and plays a role in viroid genome replication by trimming a newly generated RNA strand to the correct length [36, 37, 38]. The Hammerhead ribozyme is a self-cleaving RNA that consists of a three-stem-loop structure surrounding an autocatalytic core sequence. The stems are numbered from 5′ to 3′ as I, II, and III according to the their position with respect to the site of cleavage, which occurs at an unpaired nucleotide upstream of the 3′ strand of stem I. Hammerhead type I ribozymes are folded such that stem I is formed by the ends of the RNA sequence; type III is oriented such that stem III is formed by the end; type II is not known to exist in nature [39].

In vitro, the Hammerhead cleaving and target domains can be separated into two different RNA molecules, such that the Hammerhead RNA can act in trans and catalyze cleavage of many RNA substrates. One notable application for such a system is the construction of molecule-level biosensors for the detection of specific nucleotide sequences [40].

Another class of ribozymes is the self-splicing introns. Similar to conventional introns, these sequences occur as spacer sequences between exons that are removed in a post-transcriptional modification step. However, the self-splicing introns do not use the canonical spliceosomal machinery to catalyze splicing. Group I introns are found in diverse transcripts and species and adopt a complex 10-hairpin (helices P1 through P10) structure, which contains a catalytic core. The 5′ splice site first undergoes cleavage with a GTP cofactor, causing the upstream exon sequence to be covalently released from the intron, although it still remains associated with the intron through base pairing. Following a conformation change, cleavage and subsequent ligation with the 5′ exon occurs at the 3′ splice site, resulting in release of the intron and the ligated

exons as separate molecules [41].

Group II introns are found in the RNAs of organelles in protists, fungi, and plants; and also in bacteria. Their structure consists of six domains, dI - dVI, which contain the autocatalytic regions for splicing. In an analogous pathway to spliceosome-catalyzed splicing, the group II intron forms a lariat between a 3′ bulged adenine nucleotide and the 5′ splice site, followed by cleavage at the 3′ splice site and ligation of the exon ends [42]. *In vivo*, this process is aided by additional protein factors [43], some of which are encoded in open-reading frames of the introns themselves. Additionally, several examples of nested introns, called twintrons, have been identified, such that an internal intron is spliced prior to the excision of the external intron [44].

2.3.3   Antisense RNAs

The term "antisense RNA" is used to describe regulatory ncRNAs whose primary function is to base pair specific (sense) sequence. The most widely studied mode of antisense regulation is RNA interference, a form of post-transcriptional regulation mediated by endogenously and exogenously encoded small interfering RNAs (siRNAs), microRNAs (miRNAs), and most recently characterized, Piwi-interacting RNAs (piRNAs). However, other antisense RNAs exist that are involved in transcriptional silencing as well. Some antisense RNAs act in *cis*, meaning that they are created from the opposite strand of the gene target that they regulate; others, notably miRNAs, operate in *trans*, such that the antisense RNA gene loci are distinct from the genes that are the targets of regulation.

Small interfering RNAs

siRNAs are small RNAs, ∼21 nts in length, that are created by endonucleolytic cleavage of a double-stranded RNA precursor and are the specificity determinants of

RNAi. The source of the dsRNA is normally external to the cell, as might be the case if the cell were infected by an RNA virus or some other pathogen; however, several instances of endogenously-encoded double-stranded precursors exist [45].

The presence of dsRNA activates Dicer, a cytoplasmic RNase III that binds the ends of the RNA and cleaves through both strands at a distance approximately two helical turns from the ends [46]. The same dsRNA can be cleaved several times in succession, generating multiple ~21 nt duplexes, each of which is a distinct siRNA. The resulting siRNA duplex is loaded into RISC, the RNA-induced silencing complex, consisting primarily of an Argonaute protein (Ago-2 in humans). One of the siRNA strands is cleaved and dissociates, leaving the other siRNA strand to serve as the base-pairing component of RISC. Activated RISC can then bind other single-stranded RNA targets containing near-perfect complementary sequence to the siRNA, which causes subsequent "slicing" of the target – i.e., endonucleolytic cleavage catalyzed by Ago-2 resulting in inactivation of the RNA [47].

These cleaved targets can act as precursors for further siRNA generation by serving as templates for double-stranded RNA synthesis through the action of an RNA-dependent RNA polymerase (RdRP); thus the silencing effect can be amplified. siRNA amplification has been observed in nematodes [48], though is notably absent in insects and vertebrates. One consequence of this amplification is that siRNAs generated from one transcript can cause siRNA-generation on an unrelated transcript due to sequence similarity. As described above (Section 2.1.2), plants can generate *trans*-acting siRNAs (ta-siRNAs) through the involvement of a parallel silencing mechanism driven by miRNAs (see below).

The effects of silencing are not limited to post-transcriptional regulation. siRNAs associated with the RNA-induced transcriptional silencing (RITS) complex target chromosomal loci via base-complementarity of DNA sequence with the siRNA, which

in some species promotes histone methylation [49] as well as DNA methylation [50]. The result is induction of heterochromatin formation, a state of compact chromosomal conformation that inhibits transcriptional activity for genes contained in those chromosomal regions.

miRNAs are endogenously encoded analogs of siRNAs. Most miRNAs are transcribed by Pol II [51, 52] in long primary transcripts called pri-miRNAs, which are often several kilobases long. Embedded within the pri-miRNAs are stem/loop structures that constitute the precursor miRNAs (pre-miRNAs), which in turn contain a mature ~ 22 nt miRNA sequence, analogous to the ~ 21 nt siRNAs. The majority of miRNA genes lie in the introns of protein-coding genes [53], and a large number of primary transcripts contain several different miRNA genes that are transcribed together but individually processed [54].

The pre-miRNAs are cleaved from the primary transcript by the nuclear RNase III Drosha [55] in the Microprocessor complex, which also includes the double-stranded RNA-binding protein Pasha/DGCR8 that is believed to confer substrate specificity [56, 57, 58, 59]. Upon recognition of an appropriately-structured hairpin flanked by single-stranded sequence in the primary transcript [60], Drosha cleaves the stem at a point two helical turns from the stem/loop junction, forming a characteristic 2-nt $3'$ overhang. The resulting hairpin is exported to the cytoplasm by Exportin 5 [61], a RanGTP-dependent dsRNA-binding protein [62, 63] that binds the stem portion of the hairpin.

Cytoplasmic pre-miRNAs are processed by Dicer – in most species, this is the same enzyme that processes siRNAs, though in *Drosophila*, two distinct Dicer proteins have separate miRNA and siRNA functionality [64]. As with siRNAs, Dicer

cleaves from the terminal end of the stem to form an RNA duplex with two 2-nt $3'$ overhangs [65], and the resulting duplex is incorporated into RISC, where one of the strands is lost. Like siRNA-bound RISC, miRNA-bound RISC downregulates gene expression by binding to target transcripts via base complementarity. In plants, this complementarity is strong and can occur anywhere along the transcript [66]. In animals, complementarity is much weaker and occurs almost exclusively in the $3'$ UTR [67]. Downregulation in plants is achieved predominantly via cleavage of the bound mRNA, while in animals miRNA-RISC mediates translational repression. However, there are notable exceptions where complementarity is exact and cleavage occurs in animals [68]. Animal transcripts can have several miRNA binding sites, and in fact combinatorial binding appears to be a paradigm for animal miRNA-mediated regulation [69].

Piwi-interacting RNAs

A third, more recently characterized player in RNA silencing is the Piwi-interacting RNA (piRNA). First identified as repeat-associated small interfering RNAs (rasiRNAs) in *Drosophila* germline cells [70], these short RNAs are 23-26 nts in length – slightly longer than canonical siRNAs – and additionally occur in the testes of *Caenorhabditis elegans*, rodents, and humans. These small RNAs are bound by a class of Argonaute proteins called Piwi, which contain an eponymous domain that possesses nuclease activity.

piRNAs occur in chromosomal clusters in their respective genomes and are hypothesized to be the cleavage products of transcribed transposable elements [71], a class of repetitive sequence elements that occur throughout eukaryotic genomes with high frequency and can be thought of as endogenous parasites that can jump around the genome (i.e., "transpose"), disrupting existing sequence in the process

(see Section 2.3.5). As such, transposition of these elements is a potentially detrimental process, so silencing of active transcripts may be desirable, an activity that appears to fall under the purview of piRNA pathways. Under the proposed "ping-pong" cycle model [72, 73], existing piRNAs (perhaps inherited from cell division or explicitly transcribed) in complex with Piwi target specific transposon transcripts with base complementarity to the piRNA sequence, causing cleavage and inactivation of the transposon. In the process, new "secondary" piRNAs are generated as cleavage products, which are loaded into another Piwi-containing complex (Aubergine in *Drosophila*) and in turn target additional transcripts.

Long antisense RNA

Modern transcriptome sequencing technologies have facilitated the characterization of a more widespread phenomenon of *cis* antisense transcription, in which previously annotated loci that are known to be transcribed in one direction also appear to be transcribed from the opposite strand in the other direction as well [74]. It is estimated that as many as 40 percent of all transcriptional units may have partly or completely overlapping antisense counterparts [75], though it is still unclear what if any portion of these transcripts are artifacts arising from the identification techniques.

Several functions have been hypothesized for these "natural" antisense transcripts (NATs). If the corresponding antisense and sense transcripts base pair to form double-stranded RNA (which is not yet known [76]), some form of silencing may occur, in which the duplex is cleaved to form siRNAs or analogs. Other roles for the duplex include regulation of the sense transcript, either by disrupting regulatory binding sites or signals by virtue of blocking secondary structure formation or sequence accessibility; or by promoting editing or recognition by double-strand-specific enzymes [77].

Another intriguing hypothesis is that antisense transcription regulates transcription in the sense direction. The occupancy of loci with transcription machinery producing an antisense transcript may physically block the binding of factors necessary for transcription in the other direction. Or, transcription may commence from both directions, but then at some midpoint the RNA Pol complexes collide, preventing completion of a full-length sense transcript [78]. Yet another form of transcriptional control would be on the chromatin state of the locus – i.e., the degree to which the DNA is compacted around histones, a family of proteins that provide spools around which the DNA double helix is wound; active transcription is thought to require a non-compacted chromatin state. One model suggests that the antisense transcripts are bound by histone-modifying proteins, and base-pairing specificity causes the complexes to be recruited to the sense loci, where the neighboring histones are modified to induce changes in the chromatin state [79].

Chromatin state plays a role in X chromosome inactivation, a process that occurs in every somatic cell of mammalian females, causing one of the two copies of the X chromosome to become completely transcriptionally silent. In mice, the long sense/antisense pair Xist/Tsix, which are transcribed from the same locus in opposite directions, regulate the recruitment of histone-modifying complexes that exert their effect systemically across the chromosome. Tsix inactivates Xist by a mechanism that may involve small silencing RNA production [80], while in the absence of Tsix expression, Xist activity triggers X inactivation. Humans also contain transcribed Xist and Tsix homologs, though Tsix appears to have no effect on X inactivation [81], suggesting nuanced lineage-specific differences in the regulation pathways.

Finally, it is possible that a large portion of antisense transcription, despite being a real biological phenomenon, has little functional relevance. For instance, overlapping transcripts may not be functionally linked despite sharing common sequence, and

may be co-located in the genome due to chance or the existence of common regulatory sequence in overlapping UTRs. Or transcription may be an inherently noisy process, in which unregulated transcription occurs at many loci in both directions, generating transitory transcripts that are promptly degraded [82]. Given the potential side effects of duplex-forming sequences, however, it would be surprising if such an unregulated process could exist.

2.3.4 RNA components embedded in mRNAs

The UTRs of protein-coding RNAs are known to contain many regulatory elements, some of which are specific sequences that are bound by protein effectors. However, some are locally structured elements that can be considered a form of nested RNA. In every known case, these elements affect the translation of their host transcripts, either directly through interaction with the translational machinery, or indirectly by altering the composition or location of the mRNA.

Internal ribosome entry site

The start codon, AUG, at the 5′ end of an mRNA defines the site of translation initiation, where the protein-coding message begins. The upstream sequence is by definition an untranslated region. AUG also codes for the amino acid methionine, and generally occurrence of that particular three-nucleotide combination is not rare; thus initiating translation from the correct AUG is essential for producing a correct protein sequence. Normally the 5′ cap directs the initiation complex to the correct site. However, the presence of an internal ribosome entry site (IRES) in the 5′ UTR of an mRNA facilitates cap-independent translation initiation. IRESs were first identified in viral transcripts [83] as a way to cause host cells to preferentially translate viral RNA when coupled with inhibition of cap-binding proteins necessary for normal cap-

dependent translation. Subsequently, IRESs were found in cellular mRNAs spanning several species including human, *Drosophila*, and yeast, and may exist as a way to enhance translation of endogenous transcripts under cellular conditions that affect normal translation initiation [84]. There is no consensus IRES sequence or structure, suggesting either difficult-to-detect higher-order structural similarity or a diversity of mechanisms for achieving similar function.

RNA LOCALIZATION MOTIFS

The better-understood mode of regulating protein location in eukaryotes occurs via explicit transport of a protein from the site of translation (the ribosome, usually positioned on the endoplasmic reticulum, an organelle responsible for protein trafficking throughout the cell) to another destination distal to the nucleus – the plasma membrane for example. However, protein localization can also be mediated prior to translation, by localizing the mRNAs themselves to the correct subcellular compartment, where proteins can subsequently be translated. The advantages to this mode of localization include a finer level of regulation, since local stimuli can directly affect protein production, rather than triggering a more time-consuming cascade of signaling; the reduced cost of localizing a single mRNA molecule that can produce multiple proteins on site, compared to localizing several proteins independently; and the effective sequestering of protein products in a particular compartment to prevent off-target effects where the protein activity is not desired [85].

Localization of mRNAs is thought to involve specific sequence and structural signals contained predominantly in the UTRs that are bound by carrier complexes and shuttled to their destination, though few of these signals have been well characterized compared to the number of transcripts believed to be localized. The best known examples are involved in developmental pattern formation in *Drosophila* embryos – the

45

bicoid mRNA contains a large, ~ 600 base pair region in its 3′ UTR that was shown to be necessary for localization of the transcript to the anterior pole of *Drosophila* oocytes [86]. This region consists of several localization elements, many of which fold into hairpin structures that each confer specificity for the different stages of localization. In budding yeast, the mating-type determiner *ASH1* is selectively localized to the daughter cell during cell division via a cluster of four small stem-loop structures, each of which was shown to be independently sufficient to confer localization of the transcript to the bud tip, but to have enhanced efficacy in combination [87]. Localization motifs have also been identified for *Xenopus vg1* transcript, chicken β-actin, and *Camk2a* and *Map2* in rodent neurons, where these (along with potentially hundreds other transcripts [88]) are localized to dendrites.

Selenocysteine insertion sequence

The selenocysteine insertion sequence (SECIS) element is a motif in the 5′ UTR that mediates the introduction of a non-standard amino acid, selenocysteine, into a protein sequence [89]. The presence of a SECIS element in the transcript causes recruitment of a specialized selenocysteine-carrying tRNA, which binds to the UGA codon during translation; UGA is normally read as a stop codon. SECIS elements are common among both eukaryotic and bacterial transcripts that encode a class of proteins called selenoproteins, but despite similarity in size (~ 60 nts) and shape (hairpin), the different SECIS signals have distinct sequence [90]

Iron response element

Cells are often responsive to different concentrations of small molecules and ions. Transcripts that function in iron metabolism pathways contain hairpin motifs called iron response elements (IRE), which are bound by iron-regulatory proteins (IRPs)

that affect the translation and stability of the transcripts. In ferritin transcripts, IREs located in the 5′ UTR mediate translation inhibition when cellular levels of iron are low, and the iron-storage functionality of ferritin protein are not needed. In transferrin receptor transcripts, IREs occur in the 3′ UTR in a cluster, and binding of these by IRPs in low iron states causes stabilization of the transcript, facilitating active translation of receptor proteins for iron uptake [91]. IREs are structurally conserved among many different transcripts, as a bulged hairpin with a specific six-nucleotide loop.

Riboswitches

In contrast to IREs, which respond to iron concentration with the aid of protein complexes, riboswitches are a class of autonomous RNA aptamers, which cause transcriptional modulation upon binding of specific metabolites [92]. The canonical ribozyme is located in the 5′ UTR of an mRNA and consists of two domains. The first is the metabolite-binding aptamer domain, which is highly specialized to bind specifically to one particular metabolite, such as nucleotides or coenzyme B12. The second domain is the expression platform, the effector of transcriptional or translational control. Binding of the metabolite to the aptamer domain induces a conformational change that causes activation of the expression platform. In the case of the bacterial coenzyme-B12 riboswitch, the expression platform has two distinct effects. Upon binding of the coenzyme-B12 molecule to a partly transcribed mRNA, the conformational change induces the formation of a terminator stem that causes premature transcription termination before a functional mRNA can be created; under low coenzyme-B12 concentrations, the unbound aptamer allows the expression platform to form an anti-terminator that inhibits formation of the terminator stem, so that the mRNA is transcribed normally [93]. The second effect is on fully transcribed mRNAs, where the

47

conformational change blocks access to the ribosome binding site and inhibits translation [94]. Another mechanism for control is found in the *glmS* gene in Gram-positive bacteria, which contains a riboswitch that is a self-cleaving ribozyme; activation under conditions of high concentration of glucosamine-6-phosphate, the product whose formation is catalyzed by GlmS, causes cleavage and subsequent degradation of the *glmS* mRNA [95].

Although most riboswitches have been found in bacteria, the thiamine pyrophosphate (TPP) riboswitch, which is sensitive to cellular thiamine levels, is found in plants and fungi as well, though interestingly in plants the riboswitch resides in the 3′ UTR [96].

2.3.5 Transposable elements

Despite the diversity of protein coding genes and ncRNAs, they together constitute only a small fraction of most organismal genomes. Especially in animals and plants, the majority of genome sequence consists of high-copy-number (repetitive) sequence, which is due in large part to the action of transposable elements [97], also known as mobile elements or transposons. Transposons are sequences that at one point had the ability to mobilize and integrate into a host genome. There are three broad classes of transposons: DNA transposons, which employ a "cut and paste" mechanism for integration, such that they are excised from one genomic locus and reinserted into a different site; autonomous retrotransposons; and non-autonomous retrotransposons, both of which rely on a "copy and paste" mechanism, such that the end result is a duplication of the transposon sequence and insertion of the copy in a new genomic locus – essentially replication. Autonomous retrotransposons catalyze their own duplication and insertion, while non-autonomous retrotransposons rely on the machinery encoded by the autonomous retrotransposons.

Retrotransposons of both the autonomous and non-autonomous variety are relevant to a discussion of RNA because during the replication process, an RNA intermediate is involved, arising from active transcription of the retrotransposon sequence. For autonomous retrotransposons, transcription of the transposon results in the production of several protein-coding RNA segments that encode the protein components of the retrotransposition machinery. One class of autonomous retrotransposons, the long-terminal-repeats (LTR), have a gene architecture similar to retroviruses such as HIV, containing genes that encode a viral particle coat (GAG), a reverse transcriptase that catalyzes the creation of RNA sequence from DNA sequence, ribonuclease H for insertion-site strand cleavage, and integrase, which catalyzes the insertion of the copied transposon sequence into the genome. Non-LTR retrotransposons, such as mammalian LINE-1 (long interspersed nucleotide elements 1 or L1), encode a nucleic acid binding protein, reverse transcriptase, and an endonuclease.

The non-LTR-encoded enzymes are capable of operating in *trans*, causing other RNA in the nucleus to be integrated into the genome. This is the replication mechanism of non-autonomous retrotransposons, which include the broad class of elements called SINEs (short interspersed nucleotide elements). SINEs are the product of retrotransposition of endogenous host RNA (the "master" gene) or transposon copies of these RNAs. The ubiquitous Alu SINE element in the human genome, for example, originated from retrotransposition of 7SL RNA, and has since expanded to 1.1 million copies [97]. Similarly, in rodents, the ID element arose from BC1 RNA [98]. Retrotransposition of RNAs without subsequent expansion also occurs, resulting in a class of retrotransposons called processed pseudogenes, which are retrotransposed spliced mRNA or ncRNAs characterized by a high degree of sequence divergence from their functional counterparts [99].

Transposable element frequencies vary greatly between species, in both the com-

position and numbers of individual elements. For example, the rodent ID element appears in a few hundred copies in mouse and guinea pig, but in rats it appears more than 100,000 times [100] due to recent lineage-specific amplification.

Most transposable elements are not known to be active – the estimated novel Alu retrotransposition rate in humans is only about once per 30 individuals [101]; however, active mechanisms exemplified by piRNA-mediated silencing exist to prevent rampant retrotransposition, which can cause genomic instability due to disruption of gene sequence at the sites of insertion. Still, not all retrotransposition activity is deleterious, as transposons have the capacity to drive evolutionary change [97], by providing novel functional sequence at the site of insertion, such as regulatory regions [102] or exons [103].

2.4 RNA structure determination

2.4.1 Experimental determination of RNA structure

Several methods exist for elucidating the three-dimensional structural characteristics of individual RNA molecules that all rely on quantifying the nearness in space of individual atoms or nucleotides in the RNA to each other, or where these elements are positioned with respect to the global RNA structure. Given accurate characterization of all the pairwise atomic distances in an RNA, it is possible to reconstruct the overall three-dimensional configuration to some degree of resolution. However, generating such data using these protocols can be laborious and expensive, and as such tends to be impractical for longer RNAs. To our knowledge, there do not exist high-throughput protocols for experimental structure determination.

To determine which nucleotides in an RNA are participating in base pairs, enzymatic probes can be used that specifically cleave an RNA at either single-stranded

(un-base-paired) or double-stranded (base-paired) nucleotides. If the sequence of the RNA is known, then the resulting fragments can be size analyzed, and the cleavage sites can be mapped back onto the RNA sequence. Different enzymes have different sequence specificity – e.g., RNase V1 will cleave any base-paired region while RNase T1 preferentially cleaves single-stranded RNA at a G nucleotide [104].

Various chemical probes also exist that help determine the solvent accessibility of individual nucleotides – if we model an RNA structure as a crumpled piece of string, portions of the structure will be more exposed while other portions will be buried in the interior, thus less likely to be exposed to water molecules in solution. Low concentrations of diethylpyrocarbonate (DEPC), for example, will selectively chemically modify purine (A and G) bases by carbethoxylating them, rendering them susceptible to cleavage by aniline [105]; as above, position of structural elements (i.e., an A or G on the exterior) can be deduced from the cleavage fragments. In a similar approach called hydroxyl radical mapping, RNA is subjected to strand cleavage by high OH$^\bullet$ concentrations, and regions of the RNA that are protected from cleavage are deduced to be buried in the interior of the structure. Typically these free hydroxyls are generated using the Fenton reaction, in which hydrogen peroxide (H_2O_2) reacts with iron (Fe^{2+}) [106, 107].

The nucleotide analog interference mapping (NAIM) approach analyzes RNA structure by selective replacement of nucleotides with one of several nucleotide analogs tagged with phosphorothioate substitutions that interrupt the normal base interactions at that site. Over a series of different replacements, if a nucleotide is structurally/functionally important, then the modified RNA containing the analog at that position will not function correctly, which can be gauged using a functional assay specific to the RNA under study. The phosphorothioate substitutions serve as cleavage sites for iodine, so the positions of the nucleotide substitutions can be mapped using

a size-fractionation gel on the cleaved fragments [108, 109].

Cross-linking experiments can be used to identify nucleotides in a structure that are in close proximity to each other. Exposing unmodified RNA to short-wave ultraviolet light causes adjacent nucleotides to form covalent bonds, resulting in a permanent nonlinear structure once the RNA is treated to inhibit base pairing interactions. Different cross-linked RNAs will have different fractionation patterns when visualized on a gel [110, 111]. Because the UV is applied non-specifically, it can be difficult to deduce much structural information from the gel patterns; thus, strategies exist to introduce photoaffinity agents to specific sites along the RNA, so that cross linking can be done in a more controlled manner [112].

Fluorescence resonance energy transfer (FRET) relies on the use of paired fluorophores, one acting as an energy donor and one as an acceptor. When the fluorophores are in close proximity, laser-induced activation of the donor will cause energy to be transferred to the acceptor, resulting in a shift in the light wavelength emitted by the system. Thus, fluorophores attached onto individual nucleotides on the RNA can be used to measure distance in space between the nucleotides. FRET was used to generate a three-dimensional structure for the hammerhead ribozyme [113].

Finally, two physical approaches are used to measure atomic distances in the RNA structure, allowing the conformation to be reconstructed. The first, X-ray crystallography, has a long history of use for organic molecules, particularly proteins, but notably was instrumental in helping to shape Watson and Crick's model of the DNA double helix, which was influenced by crystallographic data generated by Rosalind Franklin [114]. In this process, a molecule of interest is formed into a crystal, then over a series of different orientations, the crystal is bombarded with X-rays, producing a two-dimensional pattern of diffraction that is a function of the way the individual atoms are packed in the crystal. These data are assembled into a three-dimensional

model of the electron densities in the molecule, a non-trivial problem. A major limitation lies in the availability of a crystallized form of the molecule of interest; most of the successfully characterized RNA structures tend to be short, on the order of 10s of nucleotides in length [115, 116].

The second method, nuclear magnetic resonance (NMR) spectroscopy, relies on the characteristic resonance patterns generated by atoms in different local chemical environments – e.g., what other atoms they are bonded to – when exposed to a magnetic field. The resulting resonance spectrum contains information about the number and characteristics of different atomic isotopes in the molecule, which can be used to generate constraints on atom position in the molecular structure model. Due to the magnetic indifference of isotopes most commonly found in biological samples, such as ^{12}C, radiolabeled sample is often used to generate more informative spectra [117].

2.4.2 Computational prediction of RNA secondary structure

Because of the low-throughput nature of experimental RNA structure determination, they are unsuitable for most genome-scale applications. Thus, researchers commonly rely on computational structure prediction. Due to relative computational ease, typical applications use secondary structure prediction and presuppose that most of the salient structural characteristics of an RNA are contained in its base-pairing pattern. Whether this is a reasonable biological assumption largely depends on the application and the RNA in question, but from a computational standpoint, considering only secondary structure rather than full tertiary structure reduces the size of the problem considerably.

Secondary structure prediction can be formulated as generating a complete list of pairwise interactions between bases in a linear sequence of RNA. Bases may only pair

with one other base, or they may remain unpaired. The gold standard for a correct secondary structure representation is of course the actual biological configuration of the RNA in three-dimensional space; however, that information is rarely available, so some heuristic measure of correctness is used. Despite the relatively tractability of secondary structure prediction compared to full structure prediction, it is still infeasible to enumerate all possible base-paired configurations, as the number of such structures is exponential in the length of the sequence, approximately 1.8^n for an n-length sequence [118].

There are many different prediction algorithms (see [119] for example) that tend to vary along three axes: how correctness of structure is judged and the source of parameters used to calculate a correctness measure; how the space of different structures is searched; and what the sequence input(s) and structure output(s) are. The most common method, as implemented in Vienna RNAfold [10] and mfold [120], applies a dynamic programming approach (the Nussinov method [121]) to search the space of possible thermodynamically favorable base pairs in a single sequence, which has cubic time complexity with respect to the sequence length. Thermodynamic parameters are in the form of empirically-derived stacking and destabilizing energies associated with particular base-pair combinations [122], which are summed over the optimal path through the dynamic programming matrix to yield a minimum-free energy (mfe) specification of base pairing. This mfe secondary structure represents the optimal configuration for the input sequence; however, it is not always the case that the predicted optimal structure corresponds to the actual structure *in vivo* [123], since higher-order nucleotide interactions or external factors such as chaperone proteins may affect the actual base-pairing pattern. Thus, the basic algorithm is suited also to return any number of sub-optimal structures (the default for mfold) subject to some ranking criterion.

Perhaps a more biophysically accurate way to envision RNA structure is to consider the distribution of possible structures expected in a large pool of sequence copies. Such an ensemble of structures can be modeled using an application of the Boltzmann distribution – the secondary structure partition function [124], which specifies the probability that a sequence adopts any one possible structure based on the structure's free energy and the temperature of the system. Intuitively, a sequence that has two equally minimal-energy conformations will have equal high probability of adopting either conformation. However, in situations where there are many equally-probable low-energy structures, the probability density of any individual structure is low; thus, it becomes useful to aggregate probabilities over individual base pairs in order to identify high-probability substructures that the sequence will adopt. Base-pair probabilities are available using RNAsubopt [125] in the Vienna package. In Sfold, the probabilities of full structures statistically sampled from the ensemble are aggregated into topologically similar clusters defined by a centroid structure that has minimum distance to all structures in the cluster [126].

When the input is a set of sequences, rather than just a single sequence, the task becomes consensus structure prediction – i.e., predicting a secondary structure that is common to all the input sequences. When a multiple sequence alignment is available or can be computed for the input set, patterns of nucleotide substitution should reflect the maintenance of secondary structure features despite sequence change. For example, if an "A" and a "U" occur at two positions in one sequence, and a "C" and "G" occur at aligned positions in another sequence, then a reasonable hypothesis would be that the nucleotides co-evolved to preserve a base-pairing interaction. Such information is leveraged by the covariation modeling methods (e.g., [127]) that produce secondary structures based on consistent patterns of substitution in the input sequences. RNAalifold integrates covariation information into the thermodynamic

framework of RNAfold by optimizing a summed score derived from a modified energy model that averages the possible free energies from base interactions at identical positions in all of the sequences, coupled with strength of covariation at those positions [128]. Other methods use the covariation information to define probabilistic models, such as stochastic context free grammars (SCFGs) (see Section 2.5.4).

The Sankoff algorithm [129] attempts to simultaneously optimize sequence alignment and thermodynamic stability of a consensus structure; however, the time complexity is n^{3m} to calculate a consensus structure for m sequences of length n, which is infeasible for large-scale applications; this has led to heuristic implementations of the algorithm, e.g., Foldalign [130] and Dynalign [131].

2.5 IDENTIFICATION AND QUANTIFICATION OF RNA

The utility of a predicted structure presupposes that the sequence of interest is functionally significant. The main criterion for assessing significance is whether the RNA sequence is present in the transcriptome, which can be determined using various experimental assays on RNA extracts from tissues or cells. However, computational techniques also exist, in which the statistical properties of known RNAs are used to build predictive models that can be applied to genome sequence. These methods are complementary, as the experimental techniques are often used to confirm predictions made by the computational techniques, or computational techniques are used to further annotate novel RNAs identified in experimental screens.

2.5.1 RNA AMPLIFICATION

Since RNA is a relatively unstable molecule, most of these assays start with a cloning step that produces complementary DNA (cDNA) from the RNA sample, using reverse

transcriptase [132]. A supply of primers is required to start the reverse transcription process; these are short nucleotide sequences complementary to the 3′ end of the RNA that serve as nucleation sites for cDNA strand synthesis. In the case of RNAs transcribed by Pol II, the 3′ poly-A tail presents a natural site for the binding of a poly-T primer. For other types of RNAs, in the absence of prior knowledge of the sequence content of the RNAs, random primer sequence is used. Following cDNA first-strand synthesis, a complementary second strand is typically synthesized, followed by one or more rounds of amplification by polymerase chain reaction (PCR) to create multiple copies of each RNA, thus facilitating easier detection.

Special strategies are required to create cDNA from small amounts of RNA. In single-cell applications, transcriptome characterization is possible using an antisense RNA amplification protocol [133]. The first strand synthesis proceeds using a modified poly-T primer that has a T7 RNA polymerase promoter ligated to the 5′ end. Following second strand synthesis, the cDNA can serve as a template for transcription by T7 polymerase, producing ~ 2000 RNA strands per cDNA, oriented antisense to the original source RNA. These RNAs can in turn be used for new cDNA generation, via random primer sequences, for further amplification. For isolating RNAs shorter than a few hundred base pairs, an initial fractionation step can be used to isolate particular RNA length classes, or functional criteria such as binding to a known protein interactor can be used to isolate a specific RNA family.

Following cDNA library creation, either a hybridization-based approach or a sequencing-based approach can be used to identify and quantify the RNA species represented in the library.

2.5.2 Hybridization approaches

Hybridization-based approaches rely on the identification of unique RNA sequences based on whether they are base-complementary to a probe of known sequence. The classic hybridization technique is the Northern blot, which is performed by using gel electrophoresis to separate an RNA sample by length on a gel, then applying a radioactive or fluorescent probe sequence, which will base pair to the target sequence on the gel if it was present in the sample [25]. Northern blots are generally useful only for small-scale analysis, as different probes for the detection of different RNA species must be tested in serial.

In contrast, microarrays are a high-throughput version of a "reverse" Northern blot, in which a sample of interest is fluorescently labeled and applied to a slide containing a matrix of thousands of probe sequences [134]. Each probe or set of probes occupies a discrete coordinate on the slide. Presence or absence of a particular RNA species is determined by whether or not the respective probes are bound by labeled RNA, which is indicated by the presence of a fluorescence signal at the coordinate of interest. The intensity of the signal is roughly proportional to the relative amount of that RNA species present in the sample; thus quantitative statements can be made between sequences in the sample or across different samples.

Two technologies exist to create microarrays; both require *a priori* knowledge of the probe sequences. Spotted arrays use probes that are synthesized prior to placement ("spotting") on glass slides. Due to variability in the amount of probe sequence in each spot, spotted arrays are usually for two-channel experiments, in which two samples are simultaneously applied, each labeled with one of two different fluorophores (e.g., green-fluorescent Cy3 and red-fluorescent Cy5), such that the ratio between the two signals per spot is considered rather than absolute fluorescence level, which can

be artifactually low if there is only a small amount of probe sequence. For oligonu-cleotide arrays such as those produced by Affymetrix, probes are synthesized directly on the slide. This process controls for copy number, so single-channel experiments can be performed and fluorescence intensities directly compared.

There is a large degree of flexibility in selecting probe sequences to include on a microarray. Gene expression arrays contain probe sequences spanning the entire characterized transcriptome for a particular species and are suitable for comparing global transcriptome expression changes in different cell or tissue conditions. To discover novel transcripts, tiling arrays can be used; these contain probes from densely overlapping segments of the genome, over a region of interest that is not necessarily known to be transcribed. Microarrays can also be custom tailored to specific classes of RNAs such as miRNAs [135] or to detect alternative splicing events [136].

To obtain spatial information about transcripts, RNA samples must be obtained from the compartment of interest – e.g., the nucleolus or the dendrite. Alternatively, *in situ* hybridization techniques can be used for fine-grained resolution of small num-bers of transcripts [137]. Radioactive or fluorescent probes are applied *in vivo*, into a single cell or tissue, and the spatial pattern of the signal indicates where the probe binds its target, reflecting the quantity and distribution of the transcript of interest.

2.5.3 Sequencing approaches

Sequencing refers to the direct determination of nucleotide sequence, which was first achieved by the Sanger method [138]. Given a single-stranded DNA (cDNA) template, sequencing proceeds via PCR synthesis using radiolabeled nucleotides, plus a supply of one of four analogous dideoxynucleotide triphosphates (ddNTPs) – ddATP, ddCTP, ddGTP, ddTTP – that serve as chain terminators due to their inability to form a $3'$ phosphodiester bond. As an example, in the presence of ddATP, during strand

synthesis ddATP will be randomly incorporated into the strand in the place of a normal A nucleotide, causing termination of the reaction and the production of a truncated sequence that ends in A. Over many such reactions, truncated sequences ending at every possible A will be created, each of different length. By separating these fragments using gel electrophoresis, the relative position of each A in the full-length sequence can be determined. Combining these positions with those determined by identical reactions with the three other ddNTPs yields a full characterization of the nucleotide sequence. By replacing the radioactive ddNTPs with fluorescent labeled ddNTPs, such that each of the four ddNTPs emits a different wavelength, the sequencing can occur in a single reaction rather than four separate ones, resulting in a more efficient pipeline.

Although the Sanger method was able to scale to produce multiple-fold coverage of various genomes consisting of billions of nucleotides, the process took several years to complete at a high cost. Thus, for transcriptome studies, strategies were developed to capture a maximal amount of information from a relatively small amount of sequence data. Expressed sequence tags (ESTs) are a way to rapidly characterize short fragments (no more than a few hundred bases) of sequences in a cDNA library using a highly error prone one-pass sequencing strategy [139]. Despite the relatively high sequence error rate, ESTs can often be mapped to unique genomic sequence, thus implicating that region as a site of transcription. For quantitative information, various other tag-based strategies can be used. Serial analysis of gene expression (SAGE), for example, uses sequence-specific endonucleases to fragment a cDNA sample into short pieces, ~10-20 nts long [140, 141]. These "tags" are ligated together into long DNAs that are then sequenced. Relative frequencies of each tag sequence correspond to the relative quantities of each transcript the tags identify, assuming they can be identified unambiguously.

Recently, so-called "next-generation" sequencing technologies have become a viable option for high-throughput sequencing applications, including transcriptome profiling. These methods can generate gigabases of sequence data in a matter of days, in the form of short sequence read fragments ranging in length from ~50 to a few hundred nucleotides. The 454 platform uses pyrosequencing chemistry to couple DNA strand synthesis with a chemiluminescent signal specific to each base type that is added to the growing strand [142]. The sequencing reaction takes place on streptavidin-coated beads, with each bead containing several copies of a bound single-stranded DNA template to be sequenced. The specific pattern of light emission from each of the identical synthesis reactions per bead reports the sequence of the template.

Illumina (formerly Solexa) technology uses a modified Sanger reaction on template sequences immobilized on a slide [143]. DNA templates are ligated to adapter sequences on both ends, both of which bind via complementarity to the slide surface, forming a bridge structure. Amplification of the sequence occurs by priming off of the adapter sequence, which generates large clusters ("polonies") of identical template sequences. During sequencing, fluorescent reversible terminator nucleotides are added one at a time, and at each step, the base-specific fluorescent signal for each polony is read. Temporary terminator groups at the end of the added nucleotides prevent multiple nucleotides from incorporating in the same round; these terminator groups are removed at the end of the round to begin the next synthesis reaction.

In contrast to the above technologies, the ABI SOLiD platform of Applied Biosystems uses a technique called sequencing by ligation, in which short fluorescently labeled oligomers of eight to nine nucleotides are successively ligated together by DNA ligase along the length of the template to be sequenced [144]. Oligomers are preferentially ligated so that the two 5′ -most nucleotides of the oligomer are base complementary to the template sequence, so the fluorescence signal of the oligomer that

61

was added corresponds to the dinucleotide sequence at that position in the template. Nucleotide identity is thus determined as di-bases spaced every few nucleotides along the template; varying the starting point of the ligation process ensures that every nucleotide in the template is covered.

For transcriptome analysis, evaluation of the sequence reads entails determining their source, often a non-trivial task given the short length of the sequence reads and the large number of reads per experiment, as well as the somewhat error-prone nature of the sequencing reads. Several read-alignment algorithms (e.g., Eland (part of the Illumina GA Pipeline, unpublished), Maq [145], SOAP [146], Bowtie [147]) are optimized for short-read alignment to large genomes. However, in many cases the length of the reads prevents unambiguous assignment of a genomic locus, if the sequence appears in multiple locations in the genome. One way to alleviate this problem is to use paired-end sequencing, in which the same template is sequenced from both ends, resulting in a mate pair consisting of two sequence reads known to be separated by a short distance (typically about 200 nucleotides). The additional information often is sufficient to anchor an otherwise ambiguous read alignment to one specific locus. The Illumina platform accommodates paired-end experiments by running the sequencing reaction twice in succession, on the two strand orientations in a polony [148].

2.5.4 Computational RNA gene-finding

There are several drawbacks to purely wet-experimental approaches to RNA identification. First, absence of an RNA from a transcriptome sample does not imply that a sequence is never transcribed; it may be expressed at low levels below the detection threshold of the technique used, or it may be conditionally expressed in particular tissues, developmental stages, or environmental contexts, and it is generally not possible

to obtain RNA samples from all possible combinations of these. Second, transcription of a sequence does not necessarily mean it is functionally significant – a somewhat controversial claim based on speculation that transcription may be a noisy mechanism [149] – or more generally, does not always provide information about *how* a sequence is functionally significant. Third, the techniques involved can be expensive and time consuming.

The flexibility and high-throughput nature of many computational approaches to RNA identification make them amenable to addressing some of these issues. All such approaches assume that there is some identifiable set of characteristics that distinguish functionally significant RNAs from background genomic sequence.

In one class of methods, explicit or implicit comparison to known RNA sequences is performed, and similarity to one or more exemplars belonging to a specific RNA family or subfamily generates a hypothesis that the unknown sequence being queried comes from the same family. Direct sequence comparison using algorithms such as BLAST [150] is often useful when there is a high level of sequence similarity over all or part of the query RNA to known RNAs; however, in many cases, sequence is poorly conserved despite structural and functional similarity [151]. Thus, secondary structure comparison is also commonly used. The simplest strategy is to apply string comparison algorithms to the Vienna dot-parenthesis structure representations, as implemented by the RNAdistance program in the Vienna Package [10].

Structure distance often provides a good heuristic for determining coherent sets of RNAs (e.g., [152, 153]) but generally assigns equal weight to all types of structural mutations. In actuality, particular classes of RNAs may vary greatly over one axis, say stem length of one helix, but not over another, such as loop size; in this case, a model that captures this sort of variational information may be a better fit, such as a generative or probabilistic model. The observation that the base-pairing pattern of

an RNA is often stable over evolutionary time, while the base identity in the base pairs can change, motivates the class of covariance models [127, 154]. Such models encode the patterns of base substitutions, obtained from sequence alignments, that are consistent with the maintenance of secondary structure features, and weight these features to form a consensus structure that can be used to classify new instances. Stochastic context free grammars (SCFGs) have also been used to model sequence and structure variation over a set of exemplar RNAs [155]. A grammar is learned where the production rules generate patterns of base pairs and unpaired nucleotides with probabilities derived from the training RNAs. Applying an SCFG to a novel sequence returns the probability that the grammar would be able to produce that sequence, which can be compared to similar probabilities from other SCFGSs or from other sequences. Other custom strategies exist for several classes of RNAs, such as miRNAs (reviewed in [156]) and snoRNAs [157].

In the absence of exemplar information, or as an augment to it, general properties of natural RNA sequences structures can be used to determine how likely an unknown sequence codes for a functional RNA. One such property is the thermodynamic stability of the RNA, as indicated by the free energy of the most stable structure (mfe), which is low for a highly structured RNA. Comparison of the mfe to a background distribution, for example as generated by computationally folding artificially shuffled versions of the query RNA sequence [158, 159], can often distinguish RNA sequences that are more stable than expected by chance, though this is often not a sufficient condition for assessing function significance [160]. Other salient characteristics include nucleotide bias, proportion of bases involved in base pair interactions, or the plasticity of the RNA as measured by the number of different low-energy configurations the sequence adopts [161, 162].

2.6 Conclusions

In this chapter, we reviewed several aspects of RNA biology, simultaneously showing the diversity of function that RNAs can possess and the diversity of methods with which to study RNAs. The subsequent chapters of this dissertation draw heavily from this diversity and attempt to synthesize a holistic view of RNAs, based on a decomposition of their parts.

References

[1] Westheimer FH (1986) Polyribonucleic acids as enzymes. Nature 319:534–5.

[2] Gilbert W (1986) Origin of life: The RNA world. Nature 319:618.

[3] Crick FH (1958) On protein synthesis. Symp Soc Exp Biol 12:138–63.

[4] Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, et al. (1965) Structure of a ribonucleic acid. Science 147:1462–5.

[5] Covey S, Al-Kaff N, Lángara A, Turner D (1997) Plants combat infection by gene silencing. Nature 385:781–782.

[6] Plasterk RHA (2002) RNA silencing: the genome's immune system. Science 296:1263–5.

[7] Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell 121:207–21.

[8] Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev 19:2164–75.

[9] Leff SE, Rosenfeld MG, Evans RM (1986) Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu Rev Biochem 55:1091–117.

[10] Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–31.

[11] Shen LX, Tinoco IJ (1995) The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. J Mol Biol 247:963–78.

[12] Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38:221–43.

[13] Chang KY, Tinoco IJ (1994) Characterization of a "kissing" hairpin complex derived from the human immunodeficiency virus genome. Proc Natl Acad Sci U S A 91:8705–9.

[14] Keniry MA (2000) Quadruplex structures in nucleic acids. Biopolymers 56:123–46.

[15] Shi H, Moore PB (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited. RNA 6:1091–105.

[16] (2001). The qiagen bench guide.

[17] Narlikar GJ, Fan HY, Kingston RE (2002) Cooperation between complexes that regulate chromatin structure and transcription. Cell 108:475–87.

[18] Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res 9:1288–93.

[19] Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64:202–36.

[20] Noller HF (1984) Structure of ribosomal RNA. Annu Rev Biochem 53:119–62.

[21] Bowman LH, Rabin B, Schlessinger D (1981) Multiple ribosomal RNA cleavage pathways in mammalian cells. Nucleic Acids Res 9:4951–66.

[22] Michot B, Bachellerie JP, Raynal F (1983) Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA. Nucleic Acids Res 11:3375–91.

[23] Reichow SL, Hamma T, Ferre-D'Amare AR, Varani G (2007) The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res 35:1452–64.

[24] Filipowicz W, Pogacic V (2002) Biogenesis of small nucleolar ribonucleoproteins. Curr Opin Cell Biol 14:319–27.

[25] Alberts B, Bray D, Lewis J, Raff M, Roberts K, et al. (2002) Molecular biology of the cell. New York: Garland Science, 4 edition.

[26] Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136:701–18.

[27] Greider CW (2006) Telomerase RNA levels limit the telomere length equilibrium. Cold Spring Harb Symp Quant Biol 71:225–9.

[28] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121–4.

[29] Kiss T (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J 20:3617–22.

[30] Darzacq X, Jady BE, Verheggen C, Kiss AM, Bertrand E, et al. (2002) Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. EMBO J 21:2746–56.

[31] Huttenhofer A, Schattner P, Polacek N (2005) Non-coding RNAs: hope or hype? Trends Genet 21:289–97.

[32] Reiner R, Ben-Asouli Y, Krilovetzky I, Jarrous N (2006) A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription. Genes Dev 20:1621–35.

[33] Schmitt ME, Clayton DA (1993) Nuclear RNase MRP is required for correct processing of pre-5.8S rRNA in Saccharomyces cerevisiae. Mol Cell Biol 13:7935–41.

[34] Lygerou Z, Allmang C, Tollervey D, Seraphin B (1996) Accurate processing of a eukaryotic precursor ribosomal RNA by ribonuclease MRP in vitro. Science 272:268–70.

[35] Steitz TA, Moore PB (2003) RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci 28:411–8.

[36] A PG, T BJ, M BJ, R SI, G B (1986) Autolytic Processing of Dimeric Plant Virus Satellite RNA. Science 231:1577–1580.

[37] Hutchins CJ, Rathjen PD, Forster AC, Symons RH (1986) Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. Nucleic Acids Res 14:3627–40.

[38] Forster AC, Symons RH (1987) Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. Cell 50:9–16.

[39] Pley HW, Flaherty KM, McKay DB (1994) Three-dimensional structure of a hammerhead ribozyme. Nature 372:68–74.

[40] Soukup GA, DeRose EC, Koizumi M, Breaker RR (2001) Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. RNA 7:524–36.

[41] Vicens Q, Cech TR (2006) Atomic level architecture of group I introns revealed. Trends Biochem Sci 31:41–51.

[42] Bonen L, Vogel J (2001) The ins and outs of group II introns. Trends Genet 17:322–31.

[43] Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64:435–61.

[44] Copertino DW, Hallick RB (1991) Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. EMBO J 10:433–42.

[45] Golden DE, Gerbasi VR, Sontheimer EJ (2008) An inside job for siRNAs. Mol Cell 31:309–12.

[46] Macrae IJ, Zhou K, Li F, Repic A, Brooks AN, et al. (2006) Structural basis for double-stranded RNA processing by Dicer. Science 311:195–8.

[47] Hammond SM (2005) Dicing and slicing: the core machinery of the RNA interference pathway. FEBS Lett 579:5822–9.

[48] Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. Nature 431:343–9.

[49] Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, et al. (2004) RNAi-mediated targeting of heterochromatin by the RITS complex. Science 303:672–6.

[50] Onodera Y, Haag JR, Ream T, Nunes PC, Pontes O, et al. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell 120:613–22.

[51] Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–60.

[52] Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10:1957–66.

[53] Kim YK, Kim VN (2007) Processing of intronic microRNAs. EMBO J 26:775–83.

[54] Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–70.

[55] Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425:415–9.

[56] Landthaler M, Yalcin A, Tuschl T (2004) The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Curr Biol 14:2162–7.

[57] Han J, Lee Y, Yeom KH, Kim YK, Jin H, et al. (2004) The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 18:3016–27.

[58] Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432:231–5.

[59] Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. Nature 432:235–40.

[60] Zeng Y, Cullen B (2005) Efficient processing of primary microRNA hairpins by drosha requires flanking nonstructured RNA sequences. Journal of Biological Chemistry 280:27595–603.

[61] Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17:3011–6.

[62] Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA 10:185–91.

[63] Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. Science 303:95–8.

[64] Lee YS, Nakahara K, Pham JW, Kim K, He Z, et al. (2004) Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell 117:69–81.

[65] Hutvagner G, Zamore PD (2002) RNAi: nature abhors a double-strand. Curr Opin Genet Dev 12:225–32.

[66] Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002) Prediction of plant microRNA targets. Cell 110:513–20.

[67] Ambros V (2004) The functions of animal microRNAs. Nature 431:350–5.

[68] Yekta S, Shih IH, Bartel DP (2004) MicroRNA-directed cleavage of HOXB8 mRNA. Science 304:594–6.

[69] Doench JG, Sharp PA (2004) Specificity of microRNA target selection in translational repression. Genes Dev 18:504–11.

[70] Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, et al. (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313:320–4.

[71] Klattenhoff C, Theurkauf W (2008) Biogenesis and germline functions of piRNAs. Development 135:3–9.

[72] Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089–103.

[73] Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, et al. (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science 315:1587–90.

[74] Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, et al. (2003) Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21:379–86.

[75] Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, et al. (2006) Complex Loci in human and mouse genomes. PLoS Genet 2:e47.

[76] Timmons JA, Good L (2006) Does everything now make (anti)sense? Biochem Soc Trans 34:1148–50.

[77] Beiter T, Reich E, Williams RW, Simon P (2009) Antisense transcription: a critical look in both directions. Cell Mol Life Sci 66:94–112.

[78] Osato N, Suzuki Y, Ikeo K, Gojobori T (2007) Transcriptional interferences in cis natural antisense transcripts of humans and mice. Genetics 176:1299–306.

[79] Bernstein E, Allis CD (2005) RNA meets chromatin. Genes Dev 19:1635–55.

[80] Ogawa Y, Sun BK, Lee JT (2008) Intersection of the RNA interference and X-inactivation pathways. Science 320:1336–41.

[81] Migeon BR, Lee CH, Chowdhury AK, Carpenter H (2002) Species differences in TSIX/Tsix reveal the roles of these genes in X-chromosome inactivation. Am J Hum Genet 71:286–93.

[82] Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–41.

[83] Jang SK, Krausslich HG, Nicklin MJ, Duke GM, Palmenberg AC, et al. (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. J Virol 62:2636–43.

[84] Baird SD, Turcotte M, Korneluk RG, Holcik M (2006) Searching for IRES. RNA 12:1755–85.

[85] Martin KC, Ephrussi A (2009) mRNA localization: gene expression in the spatial dimension. Cell 136:719–30.

[86] Macdonald PM, Struhl G (1988) cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. Nature 336:595–8.

[87] Chartrand P, Meng XH, Huttelmaier S, Donato D, Singer RH (2002) Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA. Mol Cell 10:1319–30.

[88] Eberwine J, Belt B, Kacharmina JE, Miyashiro K (2002) Analysis of subcellularly localized mRNAs using in situ hybridization, mRNA amplification, and expression profiling. Neurochem Res 27:1065–77.

[89] Walczak R, Westhof E, Carbon P, Krol A (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. RNA 2:367–79.

[90] Krol A (2002) Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis. Biochimie 84:765–74.

[91] Hentze MW, Kuhn LC (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. Proc Natl Acad Sci U S A 93:8175–82.

[92] Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, et al. (2002) Genetic control by a metabolite binding mRNA. Chem Biol 9:1043.

[93] Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS (2003) Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. RNA 9:1084–97.

[94] Lundrigan MD, Koster W, Kadner RJ (1991) Transcribed sequences of the Escherichia coli btuB gene control its expression and regulation by vitamin B12. Proc Natl Acad Sci U S A 88:1479–83.

[95] Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428:281–6.

[96] Sudarsan N, Barrick JE, Breaker RR (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA 9:644–7.

[97] Kazazian HHJ (2004) Mobile elements: drivers of genome evolution. Science 303:1626–32.

[98] Kim J, Martignetti JA, Shen MR, Brosius J, Deininger P (1994) Rodent BC1 RNA gene as a master gene for ID element amplification. Proc Natl Acad Sci U S A 91:3607–11.

[99] Maestre J, Tchenio T, Dhellin O, Heidmann T (1995) mRNA retroposition in human cells: processed pseudogene formation. EMBO J 14:6333–8.

[100] Kass DH, Kim J, Deininger PL (1996) Sporadic amplification of ID elements in rodents. J Mol Evol 42:7–14.

[101] Ostertag EM, Kazazian HHJ (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet 35:501–38.

[102] Santangelo AM, de Souza FSJ, Franchini LF, Bumaschny VF, Low MJ, et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3:1813–26.

[103] Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90.

[104] D'Alessio G, Riordan J, Raines R (1997) Ribonucleases: structures and functions. Academic Press New York.

[105] Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, et al. (1987) Probing the structure of RNAs in solution. Nucleic Acids Res 15:9109–28.

[106] Latham JA, Cech TR (1989) Defining the inside and outside of a catalytic RNA molecule. Science 245:276–82.

[107] Tullius TD, Greenbaum JA (2005) Mapping nucleic acid structure by hydroxyl radical cleavage. Curr Opin Chem Biol 9:127–34.

[108] Gaur RK, Krupp G (1993) Modification interference approach to detect ribose moieties important for the optimal activity of a ribozyme. Nucleic Acids Res 21:21–6.

[109] Strobel SA (1999) A chemogenetic approach to RNA function/structure analysis. Curr Opin Struct Biol 9:346–52.

[110] Branch AD, Benenfeld BJ, Robertson HD (1985) Ultraviolet light-induced crosslinking reveals a unique region of local tertiary structure in potato spindle tuber viroid and HeLa 5S RNA. Proc Natl Acad Sci U S A 82:6590–4.

[111] Stiege W, Atmadja J, Zobawa M, Brimacombe R (1986) Investigation of the tertiary folding of Escherichia coli ribosomal RNA by intra-RNA cross-linking in vivo. J Mol Biol 191:135–8.

[112] Thomas BC, Kazantsev AV, Chen JL, Pace NR (2000) Photoaffinity cross-linking and RNA structure analysis. Methods Enzymol 318:136–47.

[113] Tuschl T, Gohlke C, Jovin TM, Westhof E, Eckstein F (1994) A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. Science 266:785–9.

[114] Klug A (1968) Rosalind Franklin and the discovery of the structure of DNA. Nature 219:808–10 passim.

[115] Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, et al. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 63:751–9.

[116] Berman HM, Zardecki C, Westbrook J (1998) The Nucleic Acid Database: A resource for nucleic acid science. Acta Crystallogr D Biol Crystallogr 54:1095–104.

[117] Furtig B, Richter C, Wohnert J, Schwalbe H (2003) NMR spectroscopy of RNA. Chembiochem 4:936–62.

[118] Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2:919–29.

[119] Machado-Lima A, del Portillo HA, Durham AM (2008) Computational methods in noncoding RNA research. J Math Biol 56:15–49.

[120] Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–15.

[121] Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 77:6309–13.

[122] Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–48.

[123] Higgs PG (2000) RNA secondary structure: physical and computational aspects. Q Rev Biophys 33:199–253.

[124] McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105–19.

[125] Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49:145–65.

[126] Chan CY, Lawrence CE, Ding Y (2005) Structure clustering features on the Sfold Web server. Bioinformatics 21:3926–8.

[127] Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22:2079–88.

[128] Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059–66.

[129] Sankoff D (1985) Simultaneous solution of the rna folding, alignment and protosequence problems. SIAM Journal on Applied Mathematics :810–825.

[130] Gorodkin J, Heyer LJ, Stormo GD (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res 25:3724–32.

[131] Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317:191–203.

[132] Harbers M (2008) The current status of cDNA cloning. Genomics 91:232–42.

[133] Phillips J, Eberwine JH (1996) Antisense RNA Amplification: A Linear Amplification Method for Analyzing the mRNA Population from Single Living Cells. Methods 10:283–8.

[134] Dufva M (2009) Introduction to microarray technology. Methods Mol Biol 529:1–22.

[135] Liang RQ, Li W, Li Y, yan Tan C, Li JX, et al. (2005) An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe. Nucleic Acids Res 33:e17.

[136] Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–4.

[137] Jin L, Lloyd RV (1997) In situ hybridization: methods and applications. J Clin Lab Anal 11:2–9.

[138] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–7.

[139] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–6.

[140] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–7.

[141] Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2:495–502.

[142] Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84–9.

[143] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–9.

[144] Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25:195–203.

[145] Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–8.

[146] Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–7.

[147] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

[148] Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. Genome Res 18:839–46.

[149] Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14:103–5.

[150] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–10.

[151] Athanasius F Bompfnewerer Consortium, Backofen R, Bernhart SH, Flamm C, Fried C, et al. (2007) RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zoolog B Mol Dev Evol 308:1–25.

[152] Leung WS, Lin MCM, Cheung DW, Yiu SM (2008) Filtering of false positive microRNA candidates by a clustering-based approach. BMC Bioinformatics 9 Suppl 12:S3.

[153] Hamilton RS, Hartswood E, Vendra G, Jones C, Bor VVD, et al. (2009) A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in Drosophila retrotransposons. RNA 15:200–7.

[154] Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–7.

[155] Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, et al. (1994) Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res 22:5112–20.

[156] Mendes ND, Freitas AT, Sagot MF (2009) Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 37:2419–33.

[157] Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 24:158–64.

[158] Le SY, Maizel JVJ (1989) A method for assessing the statistical significance of RNA folding. J Theor Biol 138:495–510.

[159] Clote P, Ferre F, Kranakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11:578–591.

[160] Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16:583–605.

[161] Huynen M, Gutell R, Konings D (1997) Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol 267:1104–12.

[162] Higgs PG (1995) Thermodynamic properties of transfer RNA: a computational study. J Chem Soc Faraday Trans 91:2531–2540.

Chapter 3

The modularity of RNA structures

3.1 Introduction

The minimum length of a well-formed RNA secondary structure is about seven nucleotides, consisting of a two-base-pair stem with a three-nucleotide loop. Depending on the nucleotide composition of both the loop region and the base-paired stem, the stability of such a minimal hairpin will vary.

Biologically relevant RNAs tend to be larger. In some cases, such as the SECIS element or precursor miRNAs (pre-miRNAs), they consist of the same basic hairpin structure, with longer loops and stems. In other cases, such as ribosomal RNAs, the RNAs form structures that are essentially combinations of these basic hairpin shapes, linked together with additional structured or unstructured sequence. Conceptually, we can think of these complex RNAs as being composed of a set of structured building blocks, whose specific nucleotide sequence and structure individually combine to confer a specific structure and function to the entire RNA molecule.

In an attempt to better understand the characteristics of these building blocks, we might try to catalog them and look for common patterns. [1] and [2] for example invoke the idea of an RNA building block strongly grounded in the properties of nucleotide-nucleotide interactions. Partly due to scope and partly due to the availability of accurate biophysical data, these building blocks are all small, yet biologically significant fragments of RNA structure – e.g., the GNRA tetraloop, a hairpin motif commonly found in ribosomal RNAs and ribozymes; or the D-loop, one of the domains of tRNA. Using a more computationally-driven approach and a larger scale, [3] defines RNA structures using topological descriptors, such that individual RNA structures are abstracted to simple graphs with edges representing stems and edges representing loops. These can be compared directly or through their graph properties, such as connectivity, and in fact, using this approach the authors show that certain topologies are more "natural" than others by virtue of their patterns of occurrence among known RNAs.

The parts-list enumeration approach in essence defines constraints on RNA structure space and shows that RNA structures draw from a finite set of components and topologies. However, there is an aspect of temporal invariance that is not explicitly captured here, one that is important for understanding these building blocks from a use/reuse perspective.

On the shorter end of biological timescales are the biogenesis processes that many RNAs undergo. Many RNAs, particularly rRNAs, are subject to snoRNA- and snRNA-mediated RNA editing and splicing on the basis of their sequence and shape specificity [4, 5]. Eukaryotic tRNAs are transcribed as longer precursor transcripts, which are recognized and cleaved on both the 5′ and 3′ ends by RNaseP and an uncharacterized endonuclease, respectively [6, 7]; some tRNAs also contain introns, which disrupt the canonical cloverleaf structure and are spliced out before the ma-

ture tRNA is exported out of the nucleus [6, 7]. The eukaryotic 18S, 5.8S, and 28S rRNAs are transcribed as a single unit and subsequently cleaved apart [8, 9]. The hammerhead ribozyme is an example of a self-splicing RNA, such that its three helices mediate cleavage of a motif that occurs on the same RNA molecule [10].

miRNA biogenesis begins with the transcription of long primary transcripts (pri-miRNAs), which fold into large structures that serve as substrates for the endonuclease Drosha [11]. Drosha, in complex with Pasha to form the Microprocessor complex, recognizes specific hairpin substructures in the pri-miRNA and cleaves them at the base of the helical stem region, yielding the pre-miRNA hairpins [12, 13]. These range in size from $\sim 60 - 120$ nucleotides and are subsequently processed by Dicer, which targets the pre-miRNAs on the basis of their hairpin shape [14, 15]. miRNAs are notable in that the sequence of the pre-miRNA hairpin remains a robust structure through these biogenesis steps, regardless of the sequence context: when embedded in the larger primary sequence, the pre-miRNA subsequence folds into a hairpin, and when it is cleaved off to form an independent unit, the sequence folds into the same hairpin [16].

The need for context-independent structural conservation, as exemplified by the miRNA biogenesis pathway, is a hallmark of the broader phenomenon of modular composability that follows from the concept of RNA building blocks, with relevance on the longer timescales of evolutionary change. It is now well recognized that novel proteins can arise from shuffling of structural domains, the most vivid example being circularly permuted proteins [17, 18]. Given the critical role of structural features in RNA function and the already recognized motifs as compiled in databases such as RFAM [19], it is conceivable that many RNAs might also have arisen from evolutionary steps of domain shuffling and domain fusions. Such a process would require that the novel molecule reach a folded state that is a composition of the structural

features of its parts – i.e., the structural features of the combinatorial pieces need to be invariant to composition with other sequences.

On the one hand, structural context robustness may be a product of the specific relationship between each sequence and its genomic context, a property that has been exploited in computational miRNA finders such as in [20]. On the other hand, certain subsequences may have some intrinsic tendency to be structurally indifferent to their surrounding sequence, irrespective of the particular identity of that surrounding sequence – e.g., a pre-miRNA would still be structurally robust if it were inserted into a different context. We call this property of intrinsic structural invariance "self containment." A self-contained structural RNA (or protein) has the potential to be a modular building block in a larger structure, carry out consistent function through biochemical modifications of surrounding sequences, and potentially maintain function when inserted into novel contexts, as might occur with viral elements.

Previously, while studying the general mutational robustness of 170 structural elements of RNA viral genomes, Wagner and Stadler found that there was a trend toward higher structural robustness in conserved elements than in non-conserved elements when placed in short non-genomic contexts [21]. Using a similar approach, Ancel and Fontana studied the intrinsic context insensitivity of a set of canalized artificial RNAs, selected to have reduced environmental plasticity, and found a positive relationship between environmental canalization and modularity [22]. Other work in RNA (e.g., [23, 24]) and proteins (e.g., [25]) suggests that there is an intimate relationship between mutational robustness and domain modularity with folding kinetics, thermodynamic stability, as well as other biogenerative processes.

In this chapter, we analyze self containment over a broad range of biological RNAs using an intuitive scoring method to quantify different degrees of context robustness. We show that in fact pre-miRNAs do exhibit a high degree of intrinsic self contain-

ment, while most other biologically relevant RNAs tend not to show such self containment. We relate self containment to other sequence and structural features of RNA and find that no simple combination of features can completely explain self containment. Finally, we show that variation among miRNAs in degree of self containment is correlated with genomic location and miRNA-family membership, as well as their biogenerative process, as illustrated by miRNAs produced by the alternate mirtron pathway. We propose that high self containment is an intrinsic property of particular RNA sequences and may be an evolutionarily selected characteristic in molecules that need to maintain structural robustness for some aspect of their function in the face of genetic perturbations, generative perturbations, and modular composition in combinatorial contexts.

3.2 The Self-Containment Index measures RNA structural modularity

3.2.1 Measuring self containment

Given a sequence of nucleotides xwy, where w is a sequence of interest and x and y are arbitrary upstream and downstream sequences, w is structurally invariant if the substructure of the w portion is identical to the structure of w in isolation. In this scenario, the paired bases in w are paired exclusively with other bases in w and do not involve the nucleotides in x and y. If w is structurally invariant regardless of the nucleotide identity of x and y, we call w self contained. We formulate self containment as a quantitative trait of w that varies with the degree of structural invariance vis-a-vis the pool of possible x and y sequences.

We developed a scoring method to measure the degree of self containment of an RNA molecule, similar to the methods used in [21] and [22] but better encapsulating

the severity of structure change over a varied number of contexts. The score is calculated as follows: for each RNA sequence w of length L folding into a particular minimum free energy (mfe) secondary structure, we create a larger sequence of length $3L$ by embedding the original sequence in between randomly generated sequences x and y of equal length, forming a concatenated molecule xwy. We fold the resulting larger sequence and measure the proportion of the original structure preserved in the larger structure (Figure 3.1). We repeat the process using 1000 different random embeddings and average the proportions to generate a single value that ranges from 0.0 to 1.0, with 1.0 indicating a maximal degree of self containment. We call this score the self-containment index (SC).
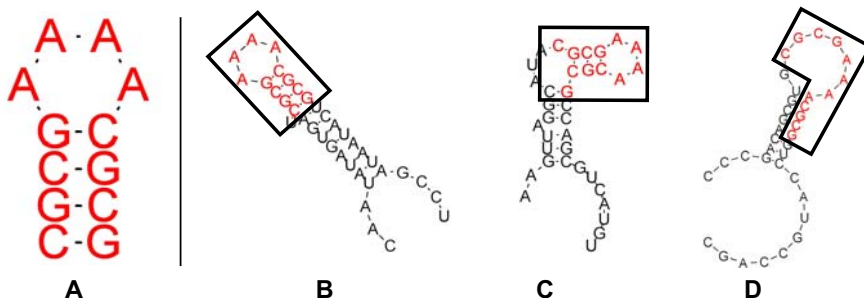
Figure 3.1: Example of varying degrees of structure preservation. (A) An RNA sequence that folds into a hairpin in isolation. (B-D) Embedding the original sequence in different surrounding sequence contexts causes varying degrees of preservation of the hairpin in the global mfe structure: complete preservation (B); loss of one base pair (C); and complete disruption of the original hairpin (D).

When applied to a set of 493 human miRNA stem loops downloaded from miRBase [26, 27], filtered to exclude sequences of > 90% sequence identity using the greedy sequence clustering algorithm Cd-hit [28], we found that the SC index produced a heavily right-shifted distribution, with an average SC value of 0.88 (Figure 3.2). We

repeated the analysis on the stem-loop sequences after trimming the 5′ and 3′ ends to align with the mature miRNA sequence while including the characteristic 2-nt 3′ overhang [11, 16], thus yielding true pre-miRNA stem loops as would be produced by Drosha processing, and found the same right-shifted distribution, again with an average SC of 0.88, though true pre-miRNA SC values tend to be slightly higher than the corresponding foldback values (p = 0.021, Wilcoxon signed rank test) (Figure 3.2). In contrast, when applied to a set of 500 randomly-generated structured RNAs, generated to approximately match the length and degree of base pairing of human miRNA foldbacks (see Materials and Methods), the SC index produced a roughly normal distribution of values centered around 0.54 (Figure 3.2). Thus, the miRNAs exhibit a significantly higher degree of self containment than random (p < 2.2×10^{-16}, Wilcoxon rank sum test).
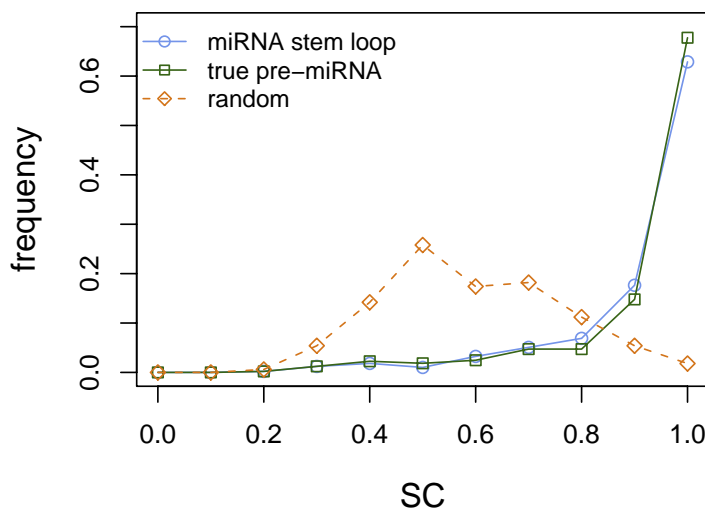
Figure 3.2: Self containment values for human pre-miRNA foldbacks versus random structures. Histograms of self-containment index values are shown for the 493 human miRNA stem loops, the stem loops trimmed to represent true pre-miRNAs, and 500 random structured RNAs.

Table 3.1: Effects of varying the number of random contexts used to calculate the self-containment index

Num. contexts used	RNA	Slope[a]	r² [b]
100	miRNA	1.00	0.98
	rand	0.97	0.98
5000	miRNA	1.00	1.00
	rand	0.99	0.99

[a]Slope of the linear regression line for the modified score as a function of the normal formulation of SC (using 1000 random contexts). [b] Correlation coefficient between the modified score and the normal formulation of SC.

We tested the robustness of the SC index by varying the number of random embeddings used and found that the index gave consistent results using as few as 100 embeddings when applied to random 100-sequence subsets of the miRNA stem loops and random structures. A Pearson correlation between SC values using 100 random embeddings versus 1000 random embeddings yielded an average slope of 0.99 with an average r^2 of 0.98, indicating that the SC index can be reliably estimated with a small sample of randomizations (Table 3.1). Similarly, increasing the number of random embeddings to 5000 also did not affect the scores (Table 3.1).

We also tested the effect of varying the length of the random context by comparing SC values obtained using the normal formulation – left and right random contexts of length L – with values obtained using context lengths ranging from $0.1L$ to $2L$. Longer contexts produced comparable SC values to the original formulation over both miRNAs and random structures, with Pearson correlations ranging from 0.98 to 0.99 and slopes from 0.98 to 1.08. SC values were slightly but significantly lower with longer context lengths, with an average difference of 0.01 for the miRNAs and 0.04 for the random structures between the L- and $2L$-derived values (p $< 1 \times 10^{-9}$, Wilcoxon signed rank test). Conversely, shorter contexts produced lower correlations and inflated SC values, with the context length of $0.1L$ yielding Pearson correlations

Table 3.2: Effects of varying the length of the random contexts used to calculate the self-containment index

Length of context[a]	RNA	Slope[b]	r² [c]	SC diff.[d]	p-value[e]
0.1L	miRNA	0.46	0.65	0.06	2.20E-16
	rand	0.78	0.61	0.21	2.20E-16
0.5L	miRNA	0.90	0.99	0.01	2.22E-16
	rand	0.98	0.97	0.05	2.20E-16
1.5L	miRNA	1.06	0.99	-0.01	9.76E-10
	rand	0.98	0.99	-0.02	2.20E-16
2L	miRNA	1.08	0.99	-0.01	6.55E-15
	rand	0.99	0.98	-0.04	2.20E-16

[a]Length of the random context appended to each end of the query sequence, with respect to L, the length of the query sequence. Under the normal formulation of SC, the context length is equal to L. [b]Slope of the linear regression line for the modified score as a function of the normal formulation of SC (using a context length of L). [c]Correlation coefficient between the modified score and the normal formulation of SC. [d]Average difference between the average SC value obtained using contexts of length L and the average SC value using the modified length context. [e]By a Wilcoxon signed rank test.

of 0.61 to 0.65 and an average increase in SC value ranging from 0.06 to 0.21 (p $< 2.2 \times 10^{-16}$, Wilcoxon signed rank test) (Table 3.2). These data indicate that a context length of L is sufficient to model the effects of large sequence surroundings, but lengths much shorter than L may be insufficient.

Finally, we tested the degree to which the base composition of the random contexts affected the SC values and found that substituting random contexts with coding sequence, intronic sequence, or versions of these with shuffled dinucleotides (i.e., the nucleotide sequences were randomly permuted in a way that preserves both the mononucleotide and dinucleotide frequencies of the original [29, 30]) had little effect on SC values. Pearson correlations between SC values produced by the original formulation compared to each of these variants, for each of the RNA classes, yielded slopes ranging from 0.91 to 1.08 with r^2 values from 0.86 to 0.98 (Table 3.3), again suggesting that the SC index can be well estimated using randomization experiments.

Table 3.3: Effects of varying the source of the random contexts used to calculate the self-containment index

Context source	RNA	Slope[a]	r^2 [b]
coding sequence	miRNA	1.04	0.98
	rand	1.02	0.98
intron	miRNA	0.92	0.97
	rand	0.99	0.96
shuffled coding	miRNA	1.01	0.98
	rand	1.01	0.98
shuffled intron	miRNA	0.89	0.96
	rand	0.99	0.96

[a]Slope of the linear regression line for the modified score as a function of the normal formulation of SC (using random contexts). [b]Correlation coefficient between the modified score and the normal formulation of SC.

3.2.2 RNA CLASSES HAVE VARYING DEGREES OF SELF CONTAINMENT

Using the SC index, we measured the self containment of several other classes of structural RNAs that have been compared previously using other measures (e.g., [31, 29, 32]): tRNAs, U1 and U2 spliceosomal RNAs, Hammerhead type III ribozymes, and 5S rRNAs (Table 3.4). All of these yielded SC ranges much lower than for the miRNAs (Figure 3.3a). The Hammerhead III ribozymes exhibited the highest average degree of self containment at 0.69, which is still significantly lower than those for the miRNAs (p = 3.95×10^{-8}, Wilcoxon rank sum test), while the remaining classes had average SC values ranging from 0.38 for U1 to 0.54 for the 5S rRNA (Figure 3.3a).

To determine whether high self containment is a product of a strong hairpin shape, which these other RNA classes lack, we additionally analyzed selenocysteine insertion sequences (SECIS) and bacterial signal recognition particle (SRP) RNAs from RFAM [19], both of which exhibit strong hairpin secondary structures. We also tested a set of hairpins derived from the protein-coding regions of mRNA transcripts,

Table 3.4: Average self-containment index values for RNA classes analyzed

RNA Class	Num. Sequences	Average SC
miRNA (all species)	4429	0.90
miRNA (human)	493	0.88
Hammerhead III ribozyme	19	0.69
Bacterial SRP	47	0.69
RFAM-extracted hairpins	9572	0.65
SECIS elements	47	0.60
5S rRNA	290	0.54
Random structures	500	0.54
tRNA	751	0.51
U2 spliceosomal	30	0.46
CD hairpins	168	0.43
U1 spliceosomal	31	0.38

originally curated to serve as a negative training set for pre-miRNA detection (CD hairpins) [33]. Both the SECIS and SRP RNAs exhibited higher SC values than all the other structural RNAs except for the Hammerhead ribozymes, yielding average values of 0.60 and 0.69, respectively; however, this was still significantly lower than for the miRNAs (p = 2.2×10^{-16} for SECIS, p = 7.24×10^{-12} for SRP, Wilcoxon rank sum test) (Figure 3.3b). The CD hairpins, despite their structural similarity to pre-miRNAs, turned out to have very low self containment, with an average SC value of 0.43, greater only than that of the U1 RNAs (Figure 3.3b, Table 3.4).

3.2.3 Two additional groups of hairpins show high self containment

In a further attempt to find groups of RNAs with similar SC distributions to the miRNAs, we considered the entire set of RFAM sequences [19, 34], filtered to > 90% sequence identity. We extracted all unbranched hairpins greater than 50 nucleotides in length, with at least half of the nucleotides involved in base pairs; these hairpins were either full-length RNAs, or they were structurally decomposable portions of full
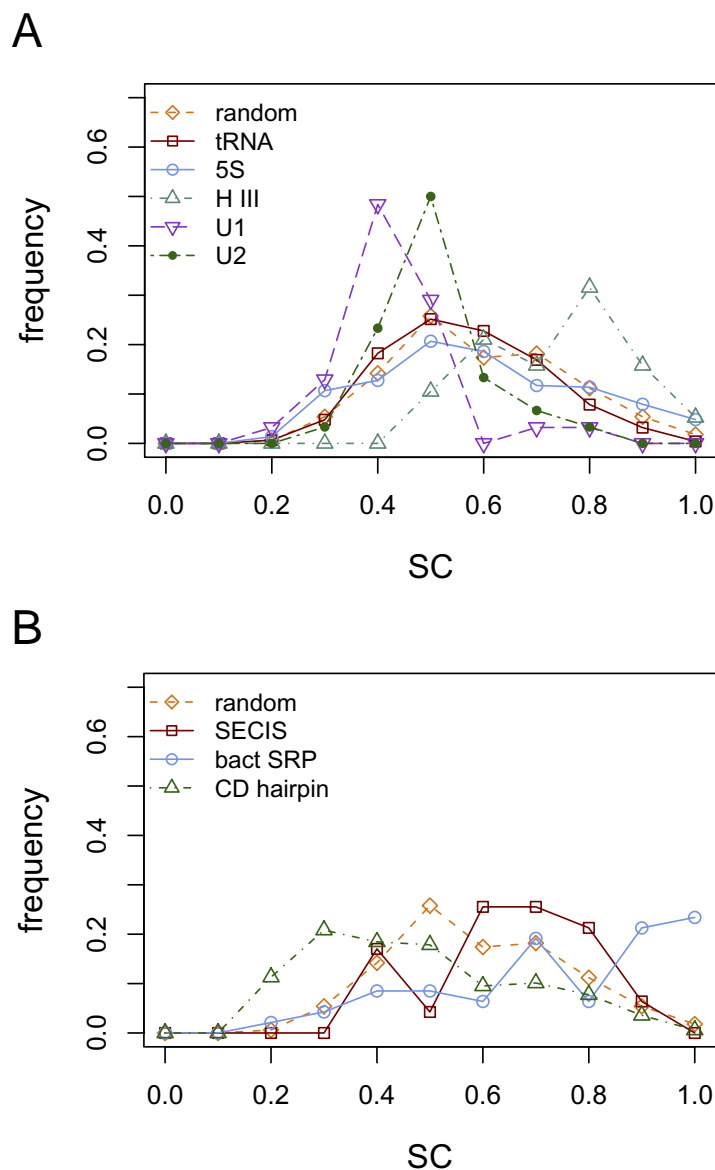
Figure 3.3: Self containment values for RNA classes. Histograms of self-containment index values are shown for (A) tRNAs, 5S rRNAs (5S), Hammerhead type III ribozymes (H III), U1 spliceosomal RNAs, and U2 spliceosomal RNAs, as compared to random structures; and (B) SECIS elements, bacterial SRP RNAs (bact SRP), and hairpins derived from protein-coding regions of mRNAs (CD hairpin), as compared to random structures.

RNAs. In all, we obtained 9572 hairpins, of which 335 were miRNAs.

We computed SC values for each hairpin. As a whole, there exists a bias toward higher SC values, though the distribution is roughly uniform among the SC values greater than 0.5 (Figure 3.4). We extracted the top 15% scoring hairpins, which corresponds to having an SC value greater than 0.900, and looked for overrepresentation of hairpins deriving from particular RFAM families. Nineteen classes show significant enrichment with $p < 0.001$ according to a Fisher's exact test, of which 12 are miRNA families (Table 3.5). Of the remaining classes, the eukaryotic SRP RNA and the hepatitis C virus stem-loop VII show the most significant skews toward high self containment, with the majority of the individuals having SC values greater than 0.9, as was observed among the miRNA stem loops. The next most significant non-miRNA class are hairpins derived from U2, which do not show as pronounced a skew.
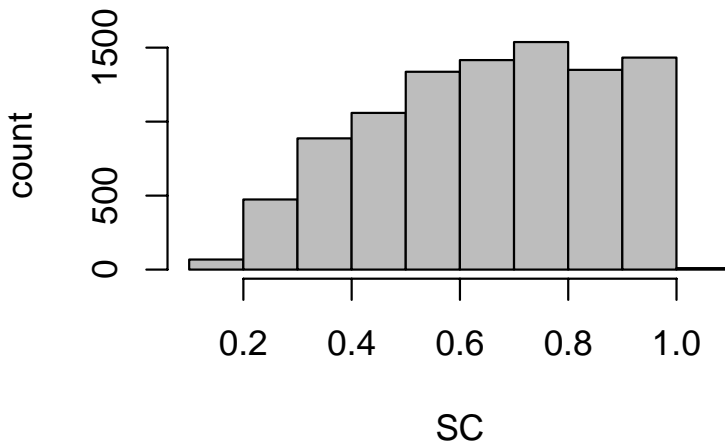
Figure 3.4: Histogram of self-containment index values for the 9572 hairpins extracted from RNAs annotated in RFAM.

Table 3.5: RFAM families whose hairpin structures are significantly enriched for high self containment

Class	Total Num. Hairpins	Observed in Top 15% SC	Expected by Chance	p-value[a]
MIR (combined)[b]	335	285	50.3	3.70×10^{-82}
RF00017 SRP_euk_arch	171	105	25.7	9.76×10^{-25}
RF00468 HCV_SLVII	41	31	6.2	4.11×10^{-10}
RF00451 mir-395	31	27	4.7	6.32×10^{-10}
RF00075 mir-166	21	20	3.2	3.74×10^{-8}
RF00445 mir-399	17	16	2.6	9.49×10^{-7}
RF00073 mir-156	15	15	2.3	1.26×10^{-6}
RF00004 U2	113	43	17.0	1.93×10^{-6}
RF00169 SRP_bact	110	37	16.5	7.17×10^{-5}
RF00247 mir-160	10	10	1.5	7.70×10^{-5}
RF00074 mir-29	9	9	1.4	1.78×10^{-4}
RF00238 ctRNA_pND324	10	9	1.5	2.98×10^{-4}
RF00103 mir-1	10	9	1.5	2.98×10^{-4}
RF00551 bicoid_3	19	12	2.9	3.15×10^{-4}
RF00256 mir-196	13	10	2.0	3.28×10^{-4}
RF00027 let-7	13	10	2.0	3.28×10^{-4}
RF00053 mir-7	8	8	1.2	4.12×10^{-4}
RF00047 mir-2	8	8	1.2	4.12×10^{-4}
RF00042 CopA	12	9	1.8	7.41×10^{-4}
RF00244 mir-26	7	7	1.1	9.62×10^{-4}

[a]By Fisher's exact test.
[b]All miRNA families combined.

3.2.4 Self-containment index correlates with other RNA measures

Having characterized the extent to which self containment varies among different RNAs, we next sought to understand the biophysical basis of SC by comparing it to other measures on structured RNAs. We compared SC values with 14 other measures drawn in part from [31] and [32]: sequence length; %GC nucleotide content; mfe and mfe normalized by length [31, 35] and GC content [35, 36]; normalized Shannon entropy of base-pair probabilities among all the structures in the thermodynamic ensemble (Q) [37]; base-pairing proportion overall (P) and the proportion of those pairs that are AU, GC, and GU pairs; z-scores of mfe, Q, and R when compared to a set of shuffled sequences preserving dinucleotide frequencies [29, 30]; and the stability of the mfe structure with respect to competing alternate structures, which is approximated by the number of structures in the thermodynamic ensemble within $2\,\mathrm{kcal/mol}$ of the mfe [23, 38] (see Materials and Methods). To test whether self containment is related to the complexity of an RNA sequence, we also compared SC to the Shannon entropy of nucleotide, dinucleotide, and trinucleotide probabilities across the sequence. Finally, we tested whether self containment depends more on the strength of base interactions in the $5'$ and $3'$ ends of the sequence rather than in the interior of the structure, using the base-pairing proportion measure limited to the distal portions of the sequence (see Materials and Methods).

We used four RNA classes for comparison: human miRNA stem loops, random structured RNAs, 5S rRNAs, and tRNAs. The correlations between variance-stabilized SC values – using an arcsin square-root transform (see Materials and Methods) – and values obtained from each of these measures are presented in Table 3.6, and scatter plots for length, GC content, mfe, mfe z-score, Q, Q z-score, P, and end-restricted P are presented in Figure 3.5.

Table 3.6: Correlation coefficients (r^2) between self-containment index and other RNA measures

Measure	miRNA	Random	5S rRNA	tRNA
length	0.04	0.01^b	0.12	0.00^b
GC proportion[a]	0.15	0.18	0.02	0.01^b
mfe	0.07	0.00^b	0.46	0.04
length-normalized mfe	0.21	0.03	0.44	0.05
GC-normalized mfe	0.31	0.06	0.63	0.27
mfe z-score	0.58	0.35	0.72	0.48
base pair entropy (Q)	0.56	0.35	0.59	0.25
base pair entropy z-score	0.58	0.37	0.56	0.28
base pair proportion (P)[a]	0.25	0.00^b	0.26	0.01
base pair proportion z-score	0.30	0.04	0.29	0.05
AU base pair proportion[a]	0.21	0.14	0.00^b	0.01
GC base pair proportion[a]	0.13	0.06	0.03	0.00
GU base pair proportion[a]	0.05	0.02	0.03	0.09
end base pair proportion[a]	0.33	0.04	0.27	0.01
end AU base pair proportion[a]	0.16	0.08	0.00^b	0.01
end GC base pair proportion[a]	0.09	0.02	0.02	0.00^b
end GU base pair proportion[a]	0.03	0.03	0.03	0.09
num. alternate structures	0.12	0.04	0.14	0.09
nucleotide entropy	0.02	0.12	0.02	0.00^b
dinucleotide entropy	0.01	0.05	0.01^b	0.01
trinucleotide entropy	0.00^b	0.01^b	0.00^b	0.01

[a]Proportion metrics were variance stabilized by performing an arcsin-square root transform before correlation was calculated.
[b]Correlation was not significant (p > 0.05).
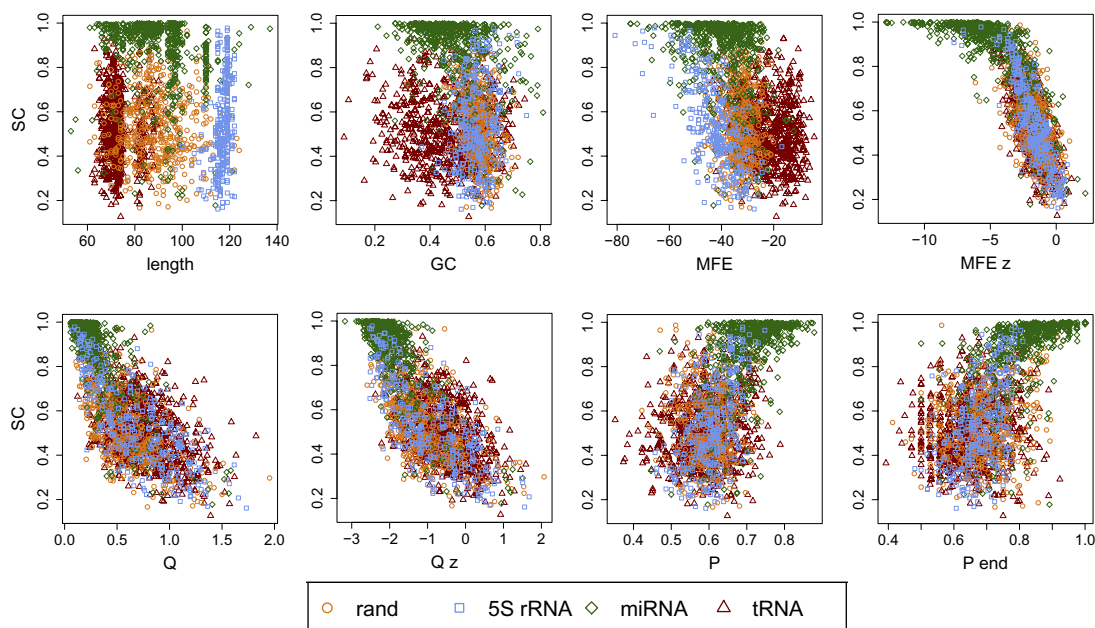
Figure 3.5: Comparison of self containment with other RNA measures. Scatter plots showing self-containment index plotted against eight other RNA measures: sequence length (length); proportion of G and C nucleotides (GC); minimum free energy of the structure (MFE); z-score of the mfe compared to 1000 dinucleotide-shuffled sequences (MFE z); normalized Shannon entropy of base-pair probabilities among all the structures in the thermodynamic ensemble (Q); z-score of Q compared to 1000 dinucleotide-shuffled sequences (Q z); proportion of bases involved in base pairs over the entire structure (P); and proportion of bases involved in base pairs, limited to the 5′ and 3′ ends of the sequence. Four sets of RNAs are overlaid in each plot: tRNAs, random structures, 5S rRNAs, and human pre-miRNAs.

For many of these measures, the relationship with SC varies depending on the class of RNA considered. Minimum free energy, for example, is moderately correlated with SC in the 5S rRNAs, but this is not the case for the other classes. Similarly, base-pairing proportion – overall, partitioned into base-pair type, or limited to particular regions of the structure – is moderately predictive for miRNAs and 5S, but not for tR-NAs. Sequence complexity, as described by the nucleotide entropy measures, appears to have little to no relationship on self containment. The strongest correlations are

between SC and mfe z-score, as well as with base pair entropy and the corresponding z-score, which themselves have all been shown to have strong correlations with one another [31].

We performed a multiple regression using all 21 variables, to assess how SC relates to a linear combination of the various RNA measures. The linear model yielded an r^2 of 0.52 for the random structures, 0.65 for tRNAs, 0.76 for miRNAs, and 0.81 for the 5S rRNAs. However, the significantly predictive variables for the regression model differed between the RNA classes, suggesting that self containment reflects a subtler sequence-structure relationship that is not captured in a common model across these factors and RNA classes.

3.2.5   RNA SEQUENCES HAVE ENHANCED SELF CONTAINMENT GIVEN THEIR STRUCTURE

To further characterize the relationship between structure and sequence in determining degree of self containment, we generated an ensemble of 100 inverse-folded sequences for each human miRNA stem loop using RNAinverse from the Vienna RNA Package [39]; each inverse-folded sequence is predicted to adopt the respective miRNA structure with minimum free energy. We then measured self containment for each set of sequences to produce a distribution of SC values for each miRNA structure and compared these distributions.

Some of the structures have very narrow ranges of admissible SC values, particularly on the high end where it appears that there are structures that are context-robust regardless of the sequence. However, most of the structures admit a wide range of possible SC values, even among structures whose real miRNA sequences exhibit very high self containment, indicating that self containment is not simply determined by structure but is an evolved feature of the sequence given a particular structure (Fig-

96
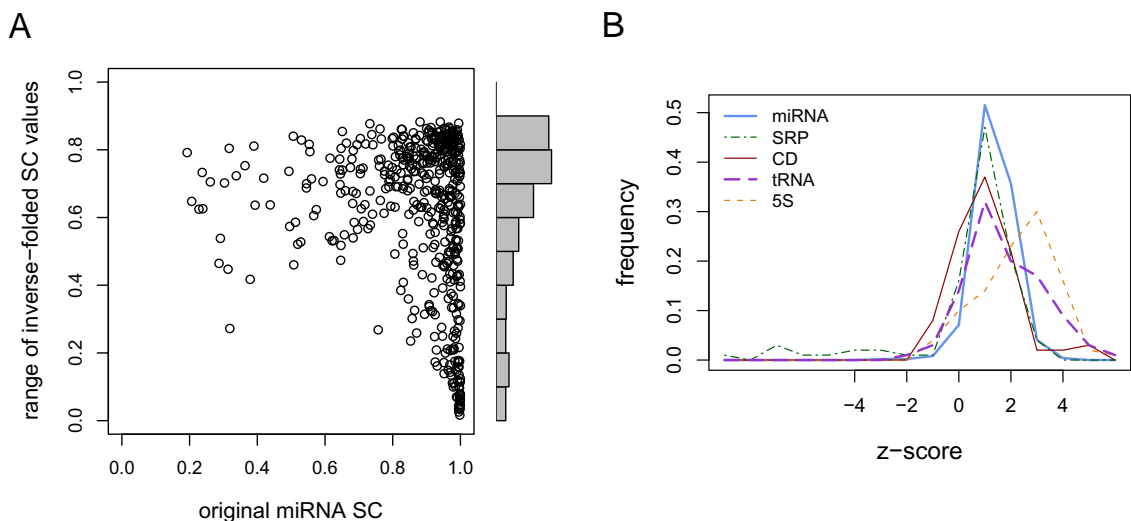
Figure 3.6: Self containment values for natural RNAs versus inverse-folded sequences. (A) Scatter plot showing self-containment index values for each original pre-miRNA versus the range of SC values observed among 100 inverse-folded sequences with the same structure as that miRNA. A range value of 0 indicates homogeneity among the SC values obtained over all 100 inverse-folded sequences, while higher values indicate higher diversity. The marginal histogram of range values is also shown. (B) Histograms showing the RNA class distributions of z-scores calculated from the self-containment index values of each RNA compared to the SC values of its 100 inverse-folded sequence ensemble. Classes shown are human pre-miRNAs (miRNA), hairpins derived from protein-coding transcripts (CD), hairpins derived from eukaryotic signal recognition particle RNAs (SRP), 5S rRNAs (5S), and tRNAs.

ure 3.6a). The same trend was observed when other types of RNA were considered (data not shown).

Using the ensemble of 100 inverse-folded sequences per miRNA stem-loop structure, we calculated the average SC value and standard deviation and compared this to the SC value of the true miRNA sequence by computing a z-score. We found a strong tendency for the real sequences to have higher self containment than average, though few of them had z-scores greater than 2 (Figure 3.6b). We performed the same analysis on random 100-sequence subsets of the 5S rRNAs, tRNAs, CD hairpins, and

the eukaryotic SRP RNA-derived hairpins we previously extracted, and found that all classes displayed right-shifted z-score distributions, indicating that the biological RNA sequences tend to be more self contained than artificial sequences that fold into the same structure (Figure 3.6b).

3.3 Precursor microRNA hairpins exhibit a high degree of modularity

3.3.1 microRNA self containment is prevalent across diverse species

To confirm that high self containment is not particular to miRNAs in humans, we measured the self containment of the miRNA stem loops spanning 56 other species represented in miRBase [26, 27]. We found that among species with at least five annotated miRNAs in miRBase, the average SC was between 0.85 and 0.98 (Table 3.7, 3.8), and that the distributions of scores when grouped into larger taxonomic classes were all heavily right shifted, as was the case for the human miRNAs (Figure 3.7).

3.3.2 Mirtrons are less self contained than canonical miRNAs

The high self containment that distinguishes miRNAs is hypothesized to be partly a function of their unique biogenesis mechanism; therefore, we tested whether enhanced self containment would still be present in the absence of the biogenesis constraint. Recently, several intronic miRNAs were characterized in *Drosophila melanogaster* [40, 41] and *Caenorhabditis elegans* [41] that bypass the Drosha cleavage pathway. Instead, these "mirtrons" are full-length intronic sequences that are spliced from protein-coding transcripts through the normal splicing pathway, giving rise to pre-miRNA foldbacks that are subsequently processed by Dicer to yield mature miRNAs.

Table 3.7: Average self-containment index values for metazoan miRNAs

Abbrev.	Species	Clade[a]	n[b]	Average SC
xla	Xenopus laevis	Amphibia	7	0.86
xtr	Xenopus tropicalis	Amphibia	177	0.92
aga	Anopheles gambiae	Arthropoda	38	0.89
ame	Apis mellifera	Arthropoda	54	0.89
bmo	Bombyx mori	Arthropoda	20	0.93
dme	Drosophila melanogaster	Arthropoda	93	0.89
dps	Drosophila pseudoobscura	Arthropoda	26	0.91
gga	Gallus gallus	Aves	154	0.91
age	Ateles geoffroyi	Mammalia	45	0.90
bta	Bos taurus	Mammalia	105	0.91
cfa	Canis familiaris	Mammalia	5	0.90
cgr	Cricetulus griseus	Mammalia	1	0.98
ggo	Gorilla gorilla	Mammalia	86	0.88
lla	Lagothrix lagotricha	Mammalia	48	0.88
lca	Lemur catta	Mammalia	16	0.86
mml	Macaca mulatta	Mammalia	71	0.89
mne	Macaca nemestrina	Mammalia	75	0.90
mdo	Monodelphis domestica	Mammalia	100	0.91
mmu	Mus musculus	Mammalia	432	0.87
oar	Ovis aries	Mammalia	3	0.74
ppa	Pan paniscus	Mammalia	89	0.89
ptr	Pan troglodytes	Mammalia	83	0.89
ppy	Pongo pygmaeus	Mammalia	84	0.89
rno	Rattus norvegicus	Mammalia	290	0.90
sla	Saguinus labiatus	Mammalia	42	0.88
ssc	Sus scrofa	Mammalia	53	0.92
cbr	Caenorhabditis briggsae	Nematoda	90	0.90
cel	Caenorhabditis elegans	Nematoda	134	0.88
dre	Danio rerio	Osteichthyes	337	0.89
fru	Fugu rubripes	Osteichthyes	131	0.93
tni	Tetraodon nigroviridis	Osteichthyes	78	0.92
sme	Schmidtea mediterranea	Platyhelminthes	63	0.91

[a]Taxonomic group by phylum or by class for vertebrates. [b]Number of miRBase-annotated miRNAs for the species, after filtering to remove sequences with >90% similarity.

Table 3.8: Average self-containment index values for non-metazoan miRNAs

Abbrev.	Species	Clade[a]	n[b]	Average SC
cre	Chlamydomonas reinhardtii	Protistae	39	0.96
ath	Arabidopsis thaliana	Viridiplantae	174	0.92
bna	Brassica napus	Viridiplantae	3	0.99
gma	Glycine max	Viridiplantae	21	0.91
mtr	Medicago truncatula	Viridiplantae	17	0.98
osa	Oryza sativa	Viridiplantae	189	0.94
ppt	Physcomitrella patens	Viridiplantae	211	0.86
pta	Pinus taeda	Viridiplantae	22	0.93
ptc	Populus trichocarpa	Viridiplantae	151	0.92
sof	Saccharum officinarum	Viridiplantae	8	0.94
smo	Selaginella moellendorffii	Viridiplantae	54	0.96
sbi	Sorghum bicolor	Viridiplantae	60	0.94
tae	Triticum aestivum	Viridiplantae	29	0.85
zma	Zea mays	Viridiplantae	79	0.95
ebv	Epstein Barr virus	Viruses	22	0.89
hsv	Herpes Simplex Virus 1	Viruses	2	0.94
hcm	Human cytomegalovirus	Viruses	11	0.91
hiv	Human immunodeficiency virus 1	Viruses	2	0.48
ksh	Kaposi sarcoma-assoc. herpesvirus	Viruses	12	0.87
mdv	Mareks disease virus	Viruses	25	0.89
mgh	Mouse gammaherpesvirus 68	Viruses	9	0.91
rlc	Rhesus lymphocryptovirus	Viruses	16	0.92
rrv	Rhesus monkey rhadinovirus	Viruses	7	0.96
sv4	Simian virus 40	Viruses	1	0.90

[a]Taxonomic group. [b]Number of miRBase-annotated miRNAs for the species, after filtering to remove sequences with >90% similarity.
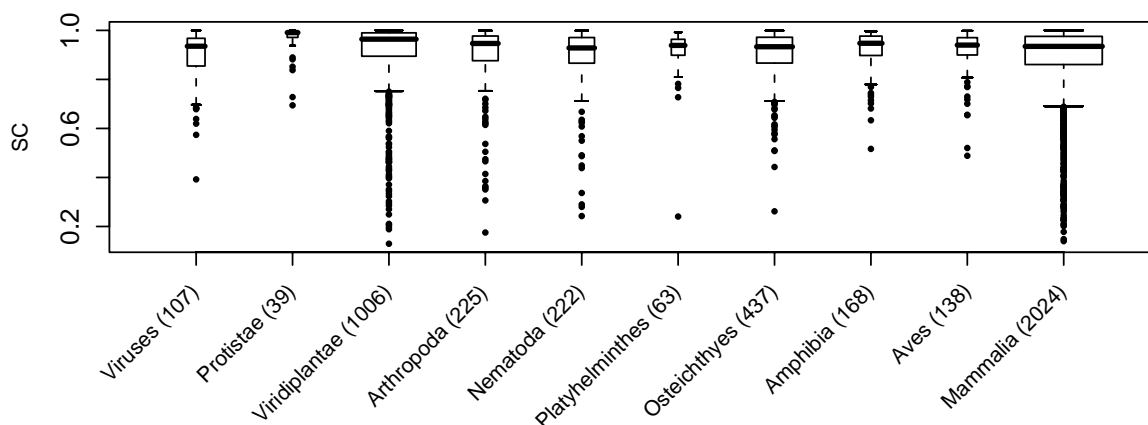
Figure 3.7: Self containment values for pre-miRNAs from various lineages. Box-and-whisker plots showing the self-containment index distribution among pre-miRNAs found in miRBase, indicating the median in bold, the interquartile range enclosed by the box, the smallest and largest non-outliers indicated by the whiskers, and outliers represented as individual points. The lineages displayed are, from left to right: viruses; protists; plants; and animals divided into the phyla arthropods, nematodes, flatworms, and chordates, which are further subdivided into classes/superclasses of fish, amphibians, birds, and mammals. Number of miRNAs for each lineage is shown in parentheses, and box width is proportional to the square root of this number.

Since mirtrons are processed as introns, structural robustness of the hairpin shape is not as critical to biogenesis as it is for pre-miRNAs that need to be excised by Drosha. We hypothesized that this effect would be reflected in lower SC values for mirtrons as compared to canonical pre-miRNAs.

For the mirtrons identified in *Drosophila* [40, 41], this does appear to be the case. We compared the SC values of the 14 mirtrons *dme-mir-1003–1016* against the remaining 76 *Drosophila* miRNAs (filtered to exclude sequences $> 90\%$ similar) and found that mirtrons have lower SC values on average – 0.83 for mirtrons versus 0.91 for canonical miRNAs; this difference achieves a significance level of $p = 0.062$ according to a t test on logit-transformed SC values. An additional degenerate *Drosophila* mirtron was characterized, *dme-mir-1017*, that is aligned to only the 5′ splice site

Table 3.9: Average self-containment index differences between mirtrons and canonical pre-miRNAs

Species	Num. Mirtrons	Avg. Mirtron SC	Avg. miRNA SC	p-value[a]
D. melanogaster	15	0.83	0.91	4.83×10^{-2}
C. elegans	4	0.98	0.88	7.06×10^{-3}
H. sapiens	13	0.50	0.88	4.96×10^{-6}
M. mulatta	11	0.67	0.89	2.39×10^{-5}

[a]By a Wilcoxon rank sum text (*C. elegans*) or by a *t* test (all others).

and has a long 3′ overhang, which presumably is cleaved subsequent to intron splicing [41]. Including *dme-mir-1017* in the analysis, after trimming the sequence from the 3′ end to yield a canonical hairpin, achieves a 5% significance level (p = 0.0483) (Table 3.9).

Among mammalian mirtrons that have recently been characterized [42], the effect is much stronger. Thirteen human and 11 *Macaque mulatta* mirtrons were identified with strong cloning evidence and sequence conservation, including one previously annotated miRNA, *mir-877*. When we compared SC values between the human mirtrons and the set of canonical miRNA stem loops excluding *hsa-mir-877*, we found that human mirtrons had an average SC of 0.50 compared to the canonical 0.88 with p = 4.96×10^{-6}, using a Wilcoxon rank sum test due to the non-normality of the data (Table 3.9). Similarly, macaque mirtrons also had a significantly lower average SC of 0.67, compared to 0.89 for the canonical miRNAs (p = 2.39×10^{-5}, *t* test) (Table 3.9).

In contrast, this trend was not observed in *C. elegans* – all four of the mirtrons identified in *C. elegans* [41] were found to be more highly self-contained than the average *C. elegans* miRNA (p = 7.06×10^{-3}, *t* test) (Table 3.9). Since mirtrons in different lineages may not have a common ancestry [42], perhaps this trend reflects a different biogenesis mechanism or evolutionary history.

3.3.3 Self containment distinguishes miRNA subclasses

Although high self containment seems to be a distinguishing characteristic for Drosha-processed miRNAs, there is still variability in the degrees of self containment among these miRNAs. We sought to account for some of this variability by measuring mean differences in SC along several functional partitions of the set of human miRNAs.

Among the full set of 533 unfiltered human miRNAs, we tested the tendency for self containment to differ among miRNAs depending on their family membership. The miRNAs belonging to a miRNA family as annotated in miRBase [26, 27] – i.e., possessing at least one ortholog or paralog – were found to be significantly more self contained, with an average SC of 0.91, than the non-conserved miRNAs, which had an average SC of 0.78 (p $= 1.32 \times 10^{-7}$, Wilcoxon rank sum test) (Table 3.10). This significance is possibly inflated by the fact that, by definition, miRNAs in a family share nucleotide sequence, which would cause some correlation in SC values among individuals in the same family. Using a more stringent formulation, obtained by averaging the human SC values per family and performing a rank sum test on family averages versus the SC values of the non-conserved miRNAs, we were still able to see the significant difference (p $= 1.37 \times 10^{-4}$). Additionally, we confirmed the result by performing a randomization test (see Materials and Methods), which is robust to sampling bias and distribution shape (p $< 10^{-5}$). Restricting the analysis to only the miRNAs with human paralogs, we again found a significantly higher degree of self containment when compared to the human miRNAs lacking human relatives (p $= 1.05 \times 10^{-4}$, Wilcoxon rank sum test; p $< 10^{-5}$, randomization test).

A large proportion of human miRNAs occur in genomic clusters [43] as part of the same primary transcript [16, 44, 45]. Using a liberal definition of clustering proposed by [20], such that a miRNA is part of a cluster if it is <10,000 nucleotides from an-

Table 3.10: Average self-containment index differences across different human pre-miRNA groups

miRNA group	In-Group Count	In-Group Avg. SC	Out-of-Group Count	Out-of-Group Avg. SC	p-value[b]
In miRNA family	404	0.91	129	0.78	1.00×10^{-5}
In human miRNA family[a]	251	0.92	282	0.84	1.00×10^{-5}
Intergenic	225	0.91	303	0.86	1.54×10^{-3}
Exon overlapping	53	0.81	475	0.89	7.69×10^{-3}
Clustered	241	0.91	287	0.86	1.20×10^{-4}

[a]Belonging to a miRNA family with multiple human members.
[b]By a randomization t test (see Materials and Methods).

other miRNA on the same strand, we found that miRNAs occurring in clusters are significantly more self contained than isolated miRNAs ($p = 1.48 \times 10^{-4}$, Wilcoxon rank test) (Table 3.10). Since clustering turns out to be correlated with family membership ($p < 2.2 \times 10^{-16}$, χ^2 test, 1 degree of freedom), we again used a randomization test to confirm significance ($p = 1.2 \times 10^{-4}$).

Finally, we tested whether miRNAs overlapping genes had differing self containment from intergenic miRNAs. Using miRBase annotations [26, 27], miRNAs classified as intergenic were significantly more self contained than gene-overlapping miRNAs ($p = 0.0195$, Wilcoxon rank sum test) (Table 3.10). When broken down into intron- versus exon-overlapping miRNAs, the effect is stronger, with exon-overlapping miRNAs significantly less self contained than non exon-overlapping miRNAs ($p = 1.5 \times 10^{-4}$, Wilcoxon rank sum test). Again, among human miRNAs there is an association between family membership and genomic location – intergenic miRNAs are overrepresented in families ($p = 2.86 \times 10^{-10}$, χ^2 test, 1 degree of freedom) and exon-overlapping miRNAs are underrepresented in families ($p = 4.84 \times 10^{-3}$, χ^2 test, 1 degree of freedom). Randomization tests again confirmed significance of the SC

differences ($p = 1.54 \times 10^{-3}$ for intergenic versus gene-overlapping, $p = 7.69 \times 10^{-3}$ for exon-overlapping versus non).

3.4 DISCUSSION AND CONCLUSIONS

In the previous sections we showed that there exist RNA sequences that have an intrinsic tendency to maintain their specific folded structure regardless of their embedded sequence context. We developed a way to measure this tendency, the self-containment index, and we used the index to show that degree of self containment varies among functional classes of RNA. miRNAs, with their need to maintain structural invariance through two cleavage steps during biogenesis, exhibit an enhanced degree of self containment, in contrast to other classes of RNAs without such a restriction. When we considered a subset of miRNAs, mirtrons, that bypass one of these cleavage steps, we found a significantly lower average self containment in three species. Among human miRNAs, we found a positive association of high self containment with membership in human-specific or cross-species miRNA families and putative transcription in a polycistronic cluster; as well as with location of the miRNAs in genomic regions not overlapping protein-coding genes. We postulate that self containment is potentially an evolved feature of particular RNA classes rather than a characteristic purely determined by the physicochemical characteristics of folded RNA.

It is possible that possessing some degree of self containment is simply an inherent property of biological RNAs. For example, small RNA subsequences that are also thermodynamically stable may be fast-folding in the kinetic folding pathway (P. Higgs, pers. comm.). Such elements would obtain their base pairing first, which would inhibit their interaction with larger sequence elements. Thus, a certain degree of self containment may be posited to be an epiphenomenon of the folding kinetics.

105

We did observe a strong relationship between SC and other measures that typically denote structurally relevant RNAs, particularly measures for structural saturation (base pair proportion), sequence-conditional structural stability (mfe z-score), and structural specificity (base-pair entropy) (Table 3.6). And, the fact that biological RNA sequences appear to have enhanced self containment given their structure (Figure 3.6b) reflects this trend as well. However, the extreme degree of self containment exhibited by the miRNAs and not by many other similarly shaped and stable RNAs seems to suggest that there is functional relevance to self containment that goes beyond being just a byproduct of structural relevance. And, as pointed out in Hartling and Kim [25] as well as Ancel and Fontana [22], there may be an inherent coupling between the modularity of biopolymer structures and both the equilibrium distribution and kinetic pathways of the folding process. Thus, selection for self containment may be mediated through fast-folding and vice versa.

The decreased self containment of mirtrons as compared to miRNAs that are processed by Drosha (Table 3.9) is evidence that the structural requirements of miRNA biogenesis at least partly explain the tendency toward high self containment. The current model for mirtron biogenesis suggests that mirtrons are spliced from mRNAs as conventional introns, with the formation of a lariat structure covalently linking the 5′ splice junction with the 3′ branch point, effectively isolating the mirtron sequence from the surrounding exonic sequence; it is only after splicing and subsequent debranching that the characteristic pre-miRNA hairpin shape is fully realized [40, 41]. Thus, mirtrons do not need to be "presented" as a context-insensitive substructure the way canonical miRNA hairpins are in the context of the primary transcript. As a result, mirtrons may be more free to accumulate nucleotide changes that lead to lower self containment, provided that the final spliced hairpin structure is not affected, whereas changes in a canonical pre-miRNA might affect recognition by Drosha due to struc-

ture disruption in the context of the primary transcript. Or, a novel proto-mirtron with lower self containment might more easily enter the miRNA processing pathway than a corresponding proto-canonical miRNA, which would additionally have to be structurally compatible with its surrounding sequence.

Still, the biogenesis mechanism may not provide sufficient *a priori* reason why pre-miRNAs should exhibit high *intrinsic* structural robustness, as opposed to structural invariance given their specific genomic contexts. Perhaps the ability to remain robust over many different genomic contexts reflects an explicit mechanism to buffer against change. At the local level, genomic instability of the surrounding primary transcript would be unlikely to affect the structure of a highly self-contained precursor stem loop, and hence would be less likely to disrupt Drosha recognition. Primary transcript sequence immediately surrounding the stem-loop sequence has been shown to be poorly conserved [43, 46], suggesting that miRNA precursor sequences do experience a high degree of instability of surrounding sequence. On a more global scale, high self containment would allow for reinsertion of a pre-existing miRNA or a copy into a novel genomic context, again with a high probability that the embedded stem-loop structure would be preserved. The trend for conserved and clustered miRNAs to exhibit higher self containment (Table 3.10) supports the idea that functional miRNAs arising from genomic modifications such as duplications and rearrangements [47] were better buffered against context change and thus were maintained. Conversely, a miRNA with low self containment would be less likely to give rise to functional paralogs – the duplicated sequence would tend not to fold correctly in the new context, making preservation of the duplicate miRNA sequence less likely due to significant loss of function.

If high self containment allows miRNA stem loops to be modular units, potentially able to function in different genomic contexts, then we might ask why selection for

modularity would exist for miRNAs. In fact, the organization of miRNAs into primary polycistronic transcripts would seem to be facilitated by modularity of the stem loops, especially given that there are several clusters that contain unrelated miRNAs [43] that may have resulted from several insertion events. The role of the primary transcript appears to be to facilitate the expression of several miRNAs at once [16], which would allow easy neofunctionalization of a duplicated miRNA if it is inserted into a primary transcript under different regulation from the source miRNA. But we might also imagine a situation where the release of individual pre-miRNAs from the primary transcript can be modulated, perhaps through RNA binding elements that block access by Drosha. This suggests a model of the primary transcript as a way to organize functionally related miRNAs while simultaneously allowing for fine-tuned control of their individual activities. Furthermore, if miRNA hairpins can be easily inserted or moved around, we can then envision the primary transcript as a collection of miRNA building blocks that can be combined and swapped over evolutionary time according to the evolving regulatory needs of the cell, a mechanism that would be difficult to attain if miRNAs were not as highly self contained.

The high self containment of miRNAs is also interesting given that they have additional sequence constraints that are ostensibly unrelated to the hairpin structure. Among miRNAs that overlap functional regions of another gene, we observed a significant decrease in average self containment (Table 3.10), indicating that these miRNAs are not as free to evolve high self containment, since any nucleotide changes leading to higher self containment might adversely affect the function of the overlapping gene. miRNAs are also constrained to maintain target specificity – loss of complementarity of the mature sequence with the target inhibits miRNA-driven regulation [48], so in a sense, miRNA hairpins are not as freely able to evolve toward highly self-contained sequences, unless compensatory changes occur in the target sequence as well. However,

given that the majority of miRNAs do have high self containment, it is also possible that there are constraints on the space of possible target sequences, such that some classes of sequences are disfavored as targets if the resulting complementary miRNA hairpins would all have low self containment. Further work is necessary to determine whether this is a quantifiable effect that can be exploited for target prediction.

As a strong indicator for miRNAs, the property of self containment can be used in future computational miRNA search strategies, as evidenced by the ability of SC to discriminate between pre-miRNAs and pseudo-hairpins (Figure 3.3b, Table 3.4), which have been repeatedly used as negative training data for miRNA prediction (e.g., [33, 49, 50]). For *de novo* design applications, ensuring high self containment among candidate structures would serve as an effective filter for hairpins that can be robustly inserted into different genetic contexts.

Beyond its potential role in miRNAs, self containment is to a certain degree a requisite property of biopolymers that form through combinatorial elaboration of modular parts. A functional fusion biopolymer cannot be generated if the fused sequences do not retain their original substructures. Recently, Rigoutsos et al. [51] have described the existence of an extensive collection of repeated nucleotide elements in the human genome that have combinatorial arrangements, potentially suggesting that combinatorial generation might be an important feature of novel RNA elements. We propose that understanding the self-containment properties of RNAs and their structural components is fundamental to understanding the extent to which RNAs are modular molecules, such that large RNAs can be decomposed into a set of structurally robust building blocks that can potentially be swapped out or rearranged.

3.5 Materials and methods

Software and implementation We used the default settings of the standalone RNAfold and RNAinverse programs bundled in the Vienna RNA Secondary Structure Package [39] for RNA secondary structure prediction and inverse folding respectively. We used a Python implementation of the Altschul-Erikson algorithm [52] for dinucleotide shuffling written by P. Clote [53]. All other code was custom written using Python and run on Linux machines. High-volume computation, including calculating SC and other structural measures on RNAs, was performed using approximately 40-60 nodes of a Linux cluster. Sequence filtering to exclude highly similar sequences was done using Cd-hit, which implements a greedy clustering algorithm [28]. RNA structure drawings were produced using RNAViz [54]. Graphs were produced using R [55].

RNA sequence sets All miRNA foldback sequences were obtained from miRBase release 10.0 [26, 27]. To obtain the "true" pre-miRNA set, we trimmed these sequences according to the structure annotation found on miRBase such that the hairpin was truncated on the 5′ end to align with the mature sequence in the case of 5′-derived mature miRNAs or the miR* sequence in the case of 3′-derived mature miRNAs; and similarly truncated on the 3′ end, creating a 2-nt 3′ overhang. CD hairpin sequences were obtained from [33]. All other RNA sequences were obtained from RFAM 8.0 seed and full sequence lists [34, 19]. Any wildcard IUPAC nucleotide characters found in the RFAM sequences were replaced with a random consistent RNA nucleotide (e.g., 'B' would be replaced with either 'C', 'G', or 'U' with equal probability).

Random RNA sequences were generated to approximately match the statistics of human miRNA foldbacks. For each candidate sequence, a random length was

chosen from a normal distribution with mean 89 and standard deviation 12.6 (the approximate average length and standard deviation of human miRNA foldbacks), and an RNA sequence was generated using uniform nucleotide probabilities; sequences shorter than 61 or longer than 137 nucleotides (again based on human miRNA shortest and longest lengths) were discarded. Candidates were folded using RNAfold, and only candidates with mfe values within one standard deviation of the average mfe for a miRNA foldback of that length were retained. The resulting set of 500 random sequences had an average length of 88.9 bases and an average minimum free energy of -32.8 kcal/mol.

Genomic coordinates, gene overlap, and family membership for the human miRNAs were also obtained from miRBase [26, 27]. Of the 533 human miRNAs in the database, five lacked genomic location information (*hsa-mir-672*, *hsa-mir-674*, *hsa-mir-871*, *hsa-mir-872*, and *hsa-mir-941-4*) and were thus left out of any analysis that depended on these features.

CALCULATING THE SELF-CONTAINMENT INDEX For each sequence of interest w with length L, a set of $2n$ random sequences of length L are generated, where n is a user-defined parameter determining the number of random contexts to test – typically 1000. The sequence w is folded using RNAfold and the structure stored in Vienna RNA parenthesis-dot notation, $struct(w)$. For each pair of random sequences x and y, a concatenated sequence xwy is created and folded using RNAfold, then the portion of the Vienna structure corresponding to the index positions of w is extracted, $struct'(w)$. $struct'(w)$ is modified to create a legal RNA structure by replacing inconsistent parentheses (indicating bases paired with bases outside of w) with dots (indicating unpaired bases). Hamming distance is calculated between $struct(w)$ and $struct'(w)$ and divided by L, and the resulting proportion is subtracted from 1 to

obtain p_i for the ith random context. All of the p_i's are averaged to obtain the final self containment index value.

For the runs using biological sequence contexts rather than random contexts, we generated a set of one thousand coding and intronic segments from randomly selected human NCBI Reference Sequence genes [56] downloaded from the UCSC Genome Bioinformatics Site [57]. Segments were extracted from a random interval at least 20 nucleotides from either end of the spliced transcript sequence for the coding sequence, or of the concatenated introns with any repetitive sequence removed using RepeatMasker [58] for the intronic sequence. Dinucleotide-shuffled sets were created from these sets as well.

RFAM HAIRPIN EXTRACTION   We started with the entire RFAM full RNA set and filtered it using Cd-hit to exclude 90% similar sequences, resulting in 26,239 sequences. We folded all of the sequences using RNAfold, then extracted all hairpin substructures. We discarded all substructures of length less than 50 nucleotides, substructures where fewer than half the bases were involved in base pairs, and any hairpins with branching, defined in terms of the Vienna representation as containing a left parenthesis in the string to the right of the first right parenthesis. We calculated SC on the resulting set of 9572 hairpins, using $n = 100$ random contexts.

RNA SEQUENCE AND STRUCTURAL MEASURES   All measures were calculated based on previous descriptions (e.g., [31, 32]). Base pairing entropy (Q) was calculated using the formulation in [37]. End base pairing proportion was calculated by summing the number of paired bases contained in the first (5′) one-fourth and the last (3′) one-fourth of the sequence and dividing by half the sequence length. Sequence entropies were calculated using single base probabilities (i.e., the number of A, C, G, and U bases occurring in the sequence each divided by the length of the sequence) in the

112

Shannon entropy equation $H = -\Sigma p_i \log_2(p_i)$ for the mononucleotide case; using probabilities of each of the possible 16 consecutive nucleotide combinations (e.g., AA, AC, ..., UU) in the dinucleotide case; and using the 64 three-consecutive nucleotide combinations in the trinucleotide case.

We reimplemented the algorithm described in [23] to characterize the number of alternate suboptimal structures of a sequence. For each sequence, all suboptimal structures within 2 kcal/mol of the mfe were obtained using RNAsubopt in the Vienna RNA Package. We filtered the results and kept only local minimum structures, defined to be structures such that removal or addition of a single base pair increases the global free energy.

Correlations were calculated using arcsin-square-root ($\sin^{-1}\sqrt{x}$) transformed values for the proportion measures such as SC (i.e., with values on [0,1]) to normalize the variances – the arcsin transformation spreads out values near 0 and 1, reducing the impact of low variance at these boundaries on the statistical analysis [59]. Values from non-proportion measures were used directly.

STATISTICAL TESTS For the randomization tests, we randomly shuffled the assignment of arcsin-square-root transformed SC values to labels (miRNA names, belonging to group A versus group B) N=100,000 times and calculated a two-sided p-value as the number of times the absolute t statistic was greater than the original absolute t statistic, divided by N. We used the Welch t statistic for unequal sample variances, $(\bar{x}_A - \bar{x}_B)/\sqrt{(s_A^2/n_A + s_B^2/n_B)}$ where $\bar{x}_A$ is the average of the group A values, s_A^2 the sample A variance, and n_A the number of members in group A; and similarly for group B.

For parametric hypothesis testing, SC values were logit transformed ($\ln(x/(1-x))$) to normalize the data – similar to the arcsin transform, the logit transform spreads out

values near 0 and 1, though in a more extreme manner to shape the data to assume a more normal-like distribution [59]. Normality was verified using the Shapiro-Wilk test, and similarity of variance was assessed using an F test. Mean differences were tested using a two-sample, two-sided independent t test, with null hypothesis that the mean difference is 0. Data that did not exhibit normality were subjected to a two-sided Wilcoxon rank sum test, or signed rank test if paired.

AVAILABILITY A Python implementation of the self-containment index calculation, as well as a web interface for direct sequence queries, is available at http://kim.bio.upenn.edu/software/.

REFERENCES

[1] Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38:221–43.

[2] Leontis NB, Lescoute A, Westhof E (2006) The building blocks and motifs of RNA architecture. Curr Opin Struct Biol 16:279–87.

[3] Kim N, Shiffeldrim N, Gan HH, Schlick T (2004) Candidates for novel RNA topologies. J Mol Biol 341:1129–44.

[4] Reichow SL, Hamma T, Ferre-D'Amare AR, Varani G (2007) The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res 35:1452–64.

[5] Filipowicz W, Pogacic V (2002) Biogenesis of small nucleolar ribonucleoproteins. Curr Opin Cell Biol 14:319–27.

[6] Wolin SL, Matera AG (1999) The trials and travels of tRNA. Genes Dev 13:1–10.

[7] O'Connor JP, Peebles CL (1991) In vivo pre-tRNA processing in Saccharomyces cerevisiae. Mol Cell Biol 11:425–39.

[8] Bowman LH, Rabin B, Schlessinger D (1981) Multiple ribosomal RNA cleavage pathways in mammalian cells. Nucleic Acids Res 9:4951–66.

[9] Michot B, Bachellerie JP, Raynal F (1983) Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA. Nucleic Acids Res 11:3375–91.

[10] Murray JB, Terwey DP, Maloney L, Karpeisky A, Usman N, et al. (1998) The structural basis of hammerhead ribozyme self-cleavage. Cell 92:665–73.

[11] Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425:415–9.

[12] Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432:231–5.

[13] Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. Nature 432:235–40.

[14] Vermeulen A, Behlen L, Reynolds A, Wolfson A, Marshall WS, et al. (2005) The contributions of dsRNA structure to Dicer specificity and efficiency. RNA 11:674–82.

[15] Ritchie W, Legendre M, Gautheret D (2007) RNA stem-loops: to be or not to be cleaved by RNAse III . RNA 13:457–62.

[16] Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–70.

[17] Lindqvist Y, Schneider G (1997) Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 7:422–7.

[18] Russell RB, Ponting CP (1998) Protein fold irregularities that hinder sequence analysis. Curr Opin Struct Biol 8:364–71.

[19] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121–4.

[20] Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, et al. (2005) Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 6:267.

[21] Wagner A, Stadler PF (1999) Viral RNA and evolved mutational robustness. J Exp Zool 285:119–27.

[22] Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. J Exp Zool 288:242–83.

[23] Higgs PG (1993) RNA secondary structure: a comparison of real and random sequences. J Phys I France 3:43–59.

[24] Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49:145–65.

[25] Harling J, Kim J (2007) Mutational robustness and geometrical form in protein structures. J Exp Zoolog B Mol Dev Evol 308B:DOI:10.1002/jez.b.21203.

[26] Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32:D109–11.

[27] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–4.

[28] Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–9.

[29] Clote P, Ferre F, Kranakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11:578–591.

[30] Le SY, Maizel JVJ (1989) A method for assessing the statistical significance of RNA folding. J Theor Biol 138:495–510.

[31] Freyhult E, Gardner PP, Moulton V (2005) A comparison of RNA folding measures. BMC Bioinformatics 6:241.

[32] Loong SNK, Mishra SK (2007) Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. RNA 13:170–87.

[33] Xue C, Li F, He T, Liu GP, Li Y, et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6:310.

[34] Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31:439–41.

[35] Seffens W, Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27:1578–84.

[36] Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA (2006) Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci 63:246–54.

[37] Huynen M, Gutell R, Konings D (1997) Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol 267:1104–12.

[38] Higgs PG (1995) Thermodynamic properties of transfer RNA: a computational study. J Chem Soc Faraday Trans 91:2531–2540.

[39] Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–31.

[40] Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 130:89–100.

[41] Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. Nature 448:83–6.

[42] Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian mirtron genes. Mol Cell 28:328–36.

[43] Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, et al. (2005) Clustering and conservation patterns of human microRNAs. Nucl Acids Res 33:2697–2706.

[44] Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, et al. (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev 16:720–8.

[45] Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T (2003) New microRNAs from mouse and human. RNA 9:175–9.

[46] Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120:21–4.

[47] Tanzer A, Stadler PF (2004) Molecular evolution of a microRNA cluster. J Mol Biol 339:327–35.

[48] Doench JG, Sharp PA (2004) Specificity of microRNA target selection in translational repression. Genes Dev 18:504–11.

[49] Jiang P, Wu H, Wang W, Ma W, Sun X, et al. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 35:W339–44.

[50] Ng KLS, Mishra SK (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23:1321–30.

[51] Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, et al. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Proc Natl Acad Sci U S A 103:6605–10.

[52] Altschul SF, Erickson BW (1985) Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol 2:526–38.

[53] Clote P (2003). Available: `http://clavius.bc.edu/∼clotelab/RNAdinucleotideShuffle/dinucleotideShuffle.html`.

[54] Rijk PD, Wuyts J, Wachter RD (2003) RnaViz 2: an improved representation of RNA secondary structure. Bioinformatics 19:299–300.

118

[55] The R Project for Statistical Computing. Available:
`http://www.r-project.org/`.

[56] Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences
(RefSeq): a curated non-redundant sequence database of genomes, transcripts
and proteins. Nucleic Acids Res 35:D61–5.

[57] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The
UCSC Genome Browser Database. Nucleic Acids Res 31:51–4.

[58] Smit AFA, Hubley R, Green P. Repeatmasker Web Server. Available:
`http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker`.

[59] Yandell BS (1997) Practical Data Analysis for Designed Experiments. Boca
Raton, Fla.: Chapman & Hall/CRC, 440 pp.

CHAPTER 4

INTRONIC RNA MODULES AND THE
CO-OPTION OF TRANSPOSABLE ELEMENTS

4.1 INTRODUCTION

THE CONTEXT FOR MODULE INSERTION In the previous chapter we explored the role of structural robustness in facilitating modularity, as a necessary condition to ensure that a structured module does not change shape (and by extension, function) upon insertion into a novel context. Of course, there is also a reciprocal question – what is the structural/functional effect on the *context* when a module is inserted? E.g., if a precursor miRNA module is introduced into a primary transcript as a result of in-place or *trans* duplication, would the overall structure of the primary transcript be disrupted in some negative way?

There is some evidence that RNA structures can be phenotypically robust to

insertions. Ribosomal RNAs, although highly conserved in their core functional regions across long evolutionary distances, do differ between lineages in terms of sequence length and corresponding structure size [1]. The *Escherichia coli* 23S rRNA component of the large prokaryotic ribosomal subunit is 2094 nucleotides, while the eukaryotic homologs are hundreds (e.g., yeast 26S) to thousands (e.g., human 28S) of nucleotides longer, due in large part to inserted "expansion segments," implying that rRNAs have some amount of structural flexibility. In fact, recent experimental work has shown that *E. coli* 23S rRNA is tolerant to *de novo* short insertions across multiple loci in the RNA sequence [2]. On a smaller scale, pre-miRNAs consist of a base-paired stem terminated by an unstructured loop. Drosha activity requires a minimum-sized loop for efficient miRNA processing [3], but even among family members, miRNAs have variable sized loops, suggesting that insertion of sequence in the loop region should have minimal effect on miRNA function. But in general, because RNA function is largely a product of structure, we would *a priori* expect structural changes to cause phenotypic changes.

Specific sequence constraints would tend to magnify the issue. Whereas an RNA structure might be robust to sequence insertions occurring in a loop region, disruption of a recognition sequence, or at the extreme, a protein-coding sequence in the case of messenger RNAs, would likely affect phenotype, possibly negatively. There are many well-characterized disease-causing insertional mutations (reviewed in [4, 5, 6]) in which protein-coding or regulatory gene sequence is interrupted by intervening nucleotide sequence. Tay-Sachs disease, a neurological disorder severely affecting mental and physical function, is caused by a four-nucleotide insertion in the gene coding for the alpha chain of beta-hexosaminidase in a majority of affected individuals in the Ashkenazi Jewish population [7]. The insertion changes the coding sequence by introducing a premature termination signal, thus resulting in truncation of the protein product.

Trinucleotide repeat disorder, a class of insertion-causing diseases including Huntington's disease, is characterized by expansion of an in-frame nucleotide triple, resulting in an abnormal number of consecutive codons and a correspondingly abnormal tract of repeated amino acids in the protein product (glutamine for Huntington's [8, 9]). Retroviruses (e.g., Human Immunodeficiency Virus, Hepatitis B), whose replication depends on integration of their genetic material into the host genome, are associated with oncogenesis (reviewed in [10]) due to their propensity to insert into tumor suppressor or proto-oncogene loci and cause transcriptional modulation.

Of course, evolution is mediated by genetic change, so it is not the case that every mutation is deleterious; however, in most cases there is strong selective pressure to maintain the integrity of genetic instructions. The need for an mRNA to robustly encode a primary sequence is often accompanied by a need to encode higher-order information – not only how to make a protein product, but also in what manner that product is expressed. Many mRNAs encode sequence and structure elements aiding in the regulation of translation, but in order not to disrupt coding sequence, these elements are often located outside the protein-coding region of the transcript – i.e., in the upstream (5′) and downstream (3′) untranslated regions (UTRs). Examples of such elements include the internal ribosome entry site (IRES) in viral 5′ UTRs, which guide the ribosome to use specific non-canonical translation initiation sites [11], and microRNA target sites, which for animals exist predominantly in the 3′ UTRs of genes and serve as recognition motifs for miRNAs to mediate translation inhibition [12]. UTRs also can contain spatial-control elements. The mRNA of *bicoid*, a *Drosophila melanogaster* body pattern-specifying gene, is localized to the anterior pole of the developing oocyte by means of a structural motif in the 3′ UTR [13]; the spatial organization of the *bicoid* mRNA facilitates localized protein translation of the Bicoid protein and the formation of a concentration gradient that

determines developmental patterning along the anterior-posterior axis. Other developmentally important *Drosophila* transcripts including *nanos*, *oskar*, and *gurken* have similar mechanisms for localization ([14]). Similarly, in neurons, localization of Ca^{2+}/calmodulin-dependent protein kinase II (*Camk2a*) and microtubule-associated protein 2 (*Map2*) transcripts from the cell soma to dendrites is mediated by distinct sequence elements in their 3′ UTRs [15, 16].

Although many examples of such regulatory elements have been characterized, there remains an excess of regulatory phenomena that do not have an identified associated UTR element. In the case of dendritic localization, the *Camk2a* and *Map2* elements remain the only examples [17], despite there being potentially hundreds of transcripts that are hypothesized to be actively transported to the dendritic compartment [18]. The difficulty may lie in the fact that a common or evolutionarily conserved sequence/structure element simply does not exist, which would complicate computational motif-finding approaches. Alternatively, the elements may lie elsewhere on the transcript.

It is possible that regions of the protein-coding portion of mRNAs might also encode secondary information. For example, plant miRNA target sites are in fact predominantly located in coding regions [19, 20] rather than the UTRs. Recently a set of localization elements was identified for glutelin RNAs in rice endosperm cells that overlap the protein-coding region [21]. The redundancy of the protein code, such that most of the 20 amino acids can be specified by multiple different nucleotide codons, suggests that selection of particular codons could allow for higher-order structure or information to be encoded. Codon bias, the phenomenon of non-uniform codon frequency, has been characterized in many organisms and can vary between organisms or even between genes in the same organism [22]. Although the role of codon bias on the secondary structure of the transcript has been widely studied on a whole-

transcript level (e.g., [23]), no specific function has been associated with localized codon usage.

Signals may also be encoded in sequence regions previously believed to be phenotypically unnecessary. A canonical eukaryotic mRNA is transcribed as a long primary transcript containing alternating regions of exons and introns. Through the act of splicing, intronic sequence is removed and the exons are ligated together to form the mature mRNA before export from the nucleus. What was once believed to be an invariant code is now known to be a differentially regulated process, in which skipping certain exons or inclusion of non-canonical exonic sequence can occur. This process, called alternative splicing, is well documented (reviewed in [24]) and evidence suggests that at least 75 percent of human genes have alternative splice forms [25]. Alternative splice forms generally encode different protein products, since the included or excluded sequence is exonic and thus affects coding sequence. In contrast, Bell et al. ([26]) report a fundamentally different phenomenon in which intronic sequence is retained. In rat hippocampal neurons, a small proportion of BK_{Ca} α-subunit mRNAs retain a specific intron, whose inclusion was demonstrated to have phenotypic effects on the distribution of the BK_{Ca} protein in the dendrite as well as firing properties of the neuron [26]. Further evidence suggests the intron-retaining transcript undergoes extranuclear splicing [27] prior to translation. The specific function of the retained intron remains unclear, but perhaps it is in these retained introns that RNA regulatory modules can exist without affecting coding sequence integrity.

THE MECHANISMS DRIVING MODULAR INSERTION Given the potential for RNA modules to exist, how do they come to be inserted at the site of need? Genomic instability in the form of chromosomal rearrangements, insertions, and deletions has been well documented [28]. These arise from double-stranded breaks in the DNA

molecule, resulting from endogenous processes such as homologous recombination during chromosome replication, or external forces such as mutagens or endonucleases. Duplications of large chromosomal regions are evident in specific loci such as the developmental *Hox* genes, which in vertebrates occur in four paralogous clusters that resulted from two separate cluster duplication events of ancestral genes [29]. In closely related species, such as mouse and rat or chimpanzee and human, although there is a high degree of gene conservation, the architecture of the chromosomes is vastly different, resulting from reorientation and recombination of chromosome segments to yield the gene order found in the modern lineages [30, 31]. On a smaller scale, there are many examples of segmental duplications facilitated by mutagenesis or errors in the replicatory machinery, which create in-place paralogous sequence [32], as well as duplications that result in sequence insertion into distant parts of the genome [33] and possibly the formation of novel genes.

Mobile transposable elements play a large role in effecting genome architecture change. Retrotransposition machinery encoded by active autonomous retrotransposons, such as LINE-1 (L1) in mammals, catalyzes the cleavage of genomic DNA and the insertion of novel DNA sequence at the break site that is created from RNA templates by reverse transcription. Processed pseudogenes arise when the RNA template is a functional protein-coding mRNA or ncRNA, causing the introduction of pseudo-genic sequence back into the genome in a location unrelated to the original gene [34]. Since these sequences tend to lack promoters at their site of insertion, the pseudogenes are not transcribed and thus accumulate neutral mutations over evolutionary time that are characteristic of non-functional genomic sequence. However, there is evidence that re-functionalization of pseudogenes can occur [35], notably the *Drosophila* alcohol dehydrogenase pseudogene, which upon retrotransposition incorporated several exons and introns from an upstream gene to form a new chimeric

gene, *jingwei* [36].

Transposon sequence itself can also become functionalized, in what tends to be a lineage-specific process due to the variability in transposon activity and composition in different species. L1, in addition to making retrotransposed copies of itself, also mobilizes SINE retrotransposons [37], including human Alu elements [38] and rodent B1 [39]. "Domestication" of such elements has led to the creation of new regulatory sequence [40] or protein-coding exons [41] at the site of insertion. Several miRNAs appear to have derived from repetitive elements, including rodent-specific *mir-327* and *mir-341* [42], as well as other ncRNAs such as primate *BC200*, a neuronal RNA that was formed from Alu sequence [43]. A general role of transposon element-driven neofunctionalization has been proposed [44, 45], and it is an appealing hypothesis that transposable elements can provide a source of mobile RNA building blocks that can become functional components of larger RNAs.

CHAPTER OVERVIEW   This chapter explores these questions concerning the context surrounding the exaption of modular RNA building blocks, as pertaining to a specific mechanism, the active transport of mRNA transcripts to the dendritic compartments of rat neurons. A large number of mRNA transcripts are detectable within neuronal dendrites, and many of these are translated locally [46, 47, 48, 49, 50, 51, 52, 53, 54], though the mechanism of targeting specific mRNAs to neuronal projections has proven to be difficult to define (reviewed in [17, 52, 55]). It is assumed that multiple RNA-binding proteins (RBPs) are involved, as well as a variety of RNA-containing granules; however, a single consensus sequence or structural motif responsible for targeting has yet to be identified within dendritically localized transcripts.

Only two separate RNA localization elements have been found – one for *Map2* [16] and one for *Camk2a* [15], both of which reside in the 3′ UTR of their respective

mRNAs. The lack of similar elements for other targeted RNAs suggests that the targeting element may be transiently associated with the RNA, perhaps as a secondary structure or a primary sequence that can be removed. In Section 4.2, we describe the phenomenon of intron retention among neuronally expressed transcripts, which we believe is functionally coupled with their dendritic localization. Contained within these retained introns are a class of retrotransposons called Identifier (ID) elements, which occur in high copy number in the rat genome and are capable of driving dendritic localization of the transcripts in which they occur. Based on a genome-wide analysis, there is evidence that ID-mediated localization is widespread among many different transcripts despite being an evolutionarily young innovation.

In Section 4.3 we focus on the previously-characterized localization element responsible for the dendritic localization of *Camk2a* mRNA and show that potentially target-competent versions of this sequence occur throughout the genome and preferentially occur overlapping Alu retroelements, suggesting that co-option of Alu-derived sequence may be a way for a transcript to obtain a *Camk2a*-style localization phenotype. These results indicate that a closer examination of repetitive elements for possible localization motifs is warranted.

These examples highlight the potential for transposable elements to mediate lineage-specific broad evolutionary change in processes that *a priori* might appear to be fundamental and strongly evolutionarily conserved.

4.2 ID elements in introns effect rat neuronal transcript localization

Intronic sequences are often only considered to play a strong role in mRNA metabolism through splicing and non-sense mediated decay (reviewed in [56, 57, 58, 59]). Recent

studies indicate that the retention of specific intronic sequences within cytoplasmic mRNA in both mammalian neurons and platelets plays an important role in producing functional proteins. Intronic retention in neuronal *Kcnma1* mRNA contributes to the firing properties of the hippocampal neurons and has a role in proper channel localization in hippocampal dendrites [26]. Specific intron retention in cytoplasmic oxytocin transcripts has also been shown in the rat supraoptic nuclei [60]. Intronic retention within *IL1-β* mRNA in anucleate platelets has been implicated in governing activity-dependent splicing and translation of the transcript upon activation of the cell [61].

Introns contain a number of known regulatory sequence elements, many of which are presumed to be involved in the control of pre-mRNA splicing (reviewed in [62]). Previously it was demonstrated that rat hippocampal neurons contain spliceosome components localized outside of the nuclear compartment in the soma and dendrites, and isolated dendrites have the capacity to splice pre-mRNA reporter constructs [63]. Additionally, non-coding sequences of intron-retaining transcripts may also serve as RBP binding sites, making regulatory elements found in these retained introns important to the cellular function of intron-retaining transcripts.

Here we report that the retention of introns is a mediator of dendritic localization for a number of neuronal transcripts. Using a candidate group of genes whose mRNAs are targeted to dendrites, we identified a large and diverse group of retained introns within the dendritically localized mRNAs. Candidates were initially identified by microarray analysis of dendritic mRNA and dendritic localization was confirmed by *in situ* hybridization. A computational analysis of a sub-group of these intron candidates for possible regulatory RNA sequences revealed the enrichment of BC1-derived SINE elements, called ID elements, across positive candidates. We hypothesize that these elements play a role in the dendritic localization of their host genes.

We performed Illumina sequencing on soma and isolated dendrite RNA and were able to confirm the intron retention patterns observed in the microarray. Additionally, we confirmed the presence of a large number of individual ID element loci in the transcriptome samples. Individual intronic ID elements from different genes were cloned, exogenously expressed in primary neurons, and evaluated by *in situ* hybridization for their ability to target mRNA to dendrites. ID elements that were targeting-competent by transgene expression were also shown to compete with endogenous transcripts for dendritic targeting machinery, thereby selectively disrupting the transcripts' normal distribution patterns. Beyond these genes of interest, we found a genome-wide pattern of ID element insertion into genes that have neuronal function, suggesting that the phenomenon of ID element-driven localization may be widespread in rat.

4.2.1 INTRON-RETAINING SEQUENCES ARE DETECTABLE IN DENDRITIC MRNAS BY MICROARRAY AND IN SITU

Based on previous results, we hypothesized that intron retention is a wide phenomenon among rat dendritic transcripts. To test this hypothesis, we built a custom microarray using probe intronic sequence from 33 candidate genes whose RNAs can be dendritically localized [18] (Table 4.1). These probes were designed to contain 30 bp of the 3′ exonic sequence followed by approximately 300-500bp of intronic sequence, for up to three introns per gene: the first intron following the initiator methionine codon, the last intron preceding the termination codon, and an intron located roughly midway through the gene sequence.

We obtained microarray data from three independent rat dendritic samples. High Spearman's correlation between the arrays was found ($\rho > 0.94$, $p < 2.2$ x 10^{-16}), indicating a consistent rank ordering among intron signal intensities. A wide varying range of signal was found across the arrayed intronic sequences, with 33 of 92 introns

Table 4.1: Genes with introns represented on the microarray

RefSeq ID	Symbol	Description	Coordinate		Introns[a]
NM_031007	ADCY2	Adenylate cyclase 2	chr17:4543490-5040433	+	1, 4, 20 (23)
NM_130779	ADCY3	Adenylate cyclase 3	chr6:27118400-27202275	+	2, 3, 21 (21)
NM_019285	ADCY4	Adenylate cyclase 4	chr15:33930534-33946315	-	2, 11, 24 (25)
NM_022600	ADCY5	Adenylate cyclase 5	chr11:67290968-67437468	-	1, 3, 20 (20)
NM_012821	ADCY6	Adenylate cyclase 6	chr7:137339933-137360020	-	1, 2, 21 (21)
NM_134326	ALB	Albumin	chr14:19126965-19142199	-	1, 6, 14 (14)
NM_019288	APP	Amyloid beta (A4) precursor protein	chr11:24457855-24693851	-	1, 6, 17 (17)
NM_147141	CACNA1B	Calcium channel, voltage-dependent, N type, alpha 1B subunit	chr3:2873391-3039747	-	1, 18, 45 (45)
NM_153814	CACNA1H	Calcium channel, voltage-dependent, T type, alpha 1H subunit	chr10:14621372-14679051	-	1, 5, 33 (33)
NM_012920	CAMK2A	Calcium/Calmodulin-dependent protein kinase II alpha subunit	chr18:56879247-56948537	+	1, 3, 10 (11)
NM_021739	CAMK2B	Calcium/Calmodulin-dependent protein kinase II beta subunit	chr14:86634690-86721261	-	1, 3, 15 (20)
NM_012519	CAMK2D	Calcium/Calmodulin-dependent protein kinase II, delta	chr2:223840650-224108082	+	1, 4, 18 (19)
NM_133605	CAMK2G	Calcium/Calmodulin-dependent protein kinase II gamma	chr15:3729433-3786057	+	2, 9, 18 (19)
NM_031334	CDH1	Cadherin 1	chr19:36442693-36512091	+	1, 3, 15 (15)
NM_031017	CREB1	CAMP responsive element binding protein 1	chr9:63170785-63234725	+	2, 6, 8 (8)
NM_052804	FMR1	Fragile X mental retardation syndrome 1 homolog	chrX:154756031-154793782	+	1, 7, 15 (15)
NM_031028	GABBR1	Gamma-aminobutyric acid B receptor 1	chr20:1553313-1582398	-	6, 11, 21 (22)
NM_080587	GABRA4	Gamma-aminobutyric acid A receptor, subunit alpha 4	chr14:39047461-39122526	+	1, 7, 8 (8)
NM_017289	GABRD	Gamma-aminobutyric acid A receptor, delta	chr5:172203065-172214960	-	1, 2, 8 (8)
NM_024370	GABRG3	Gamma-aminobutyric acid A receptor, subunit gamma 3	chr1:108189311-108821051	-	2, 5, 9 (9)
NM_032990	GRIA3	Glutamate receptor, ionotropic, AMPA3 (alpha 3)	chrX:3454606-3719276	-	1, 4, 14 (15)
NM_017263	GRIA4	Glutamate receptor, ionotropic, 4	chr8:957190-1438021	-	2, 4, 16 (16)
NM_017241	GRIK1	Glutamate receptor, ionotropic, kainate 1	chr11:27703875-28106450	-	1, 3, 16 (16)
NM_017010	GRIN1	Glutamate receptor, ionotropic, N-methyl D-aspartate 1	chr3:3453784-3480381	-	1, 8, 19 (19)
NM_031040	GRM7	Glu. receptor, metabotropic 7	chr4:146332578-147270224	+	1 (6)
NM_012970	KCNA2	Potassium voltage-gated channel, shaker-related subfamily, memb. 2	chr2:202560175-202564305	+	1 (2)
NM_031730	KCND2	Potassium voltage-gated channel, SHAL-related family, memb. 2	chr4:47541787-48047906	+	1, 2, 4 (5)
NM_013066	MAP2	Microtubule-associated protein 2	chr9:65174379-65255995	+	3, 4, 12 (12)
NM_019169	SNCA	Synuclein, alpha	chr4:89613731-89722807	-	2, 4 (5)
NM_080777	SNCB	Synuclein, beta	chr17:15907598-15915704	+	2, 5 (5)
NM_031688	SNCG	Synuclein, gamma	chr16:10025979-10030513	-	1, 4 (4)
NM_053788	STX1A	Syntaxin 1A (brain)	chr12:22737113-22765064	-	1, 3, 9 (9)
NM_012700	STX1B2	Syntaxin 1B2	chr1:187089182-187108643	-	1, 4, 7 (9)

[a]Introns spotted on the array (total number of introns in the canonical splice form).

showing high signal intensity (>75th percentile expression on at least one array), ranging from 1.5x to 22x normalized median intensity; an additional 27 introns with above-median intensity are also reported (Table 4.2).

We performed *in situ* hybridization experiments for some of these transcripts to visualize their subcellular location in the neurons (Figure 4.1), using probes synthesized from the intronic PCR products represented on the intron microarray. The PCR products were subcloned into pCRII TOPO (Invitrogen Corporation, Carlsbad, CA) vectors and sequenced. Labeled antisense riboprobes were then generated for both positive and negative microarray sequences and used for *in situ* hybridization to rat neurons in primary cell culture (Figure 4.1). Hippocampi were harvested from embryonic day 18 rat pups and dispersed cells were grown in culture for 14 days before paraformaldehyde fixing in all *in situ* experiments. Cells were co-stained for MAP2 protein to indicate dendrito-somatic regions of neurons (data not shown) and to assess healthy morphology of the cells. All sequences tested reflected microarray results and showed detectable signal in at least the proximal dendrites for microarray positives, while microarray negatives were restricted to the cell soma or not detectable in neurons.

Multiple dendritic distribution patterns can be seen for the intronic probes from any given RNA and across probe sets. In the case of *Stx1b2*, all three intronic regions identified as present from the microarray data are readily detectable in the dendrites by *in situ* hybridization showing a punctate pattern for each probe. In contrast, though FMR1i1 shows a similar punctate pattern, FMR1i7 is far more diffuse with an even intensity throughout the projections as seen for other mRNAs [64, 65, 66]. ALBi6 shows a diffuse pattern similar to FMR1i7, although the intensity is much greater (Figure 4.1). Such diffuse patterns may result from these particular forms of these mRNAs not yet assembling into granules.

Table 4.2: Intron sequences detected by microarray and Illumina sequencing[a]

		Intron A				Intron B				Intron C			
RefSeq ID	Symbol	i	M	SD	SS	i	M	SD	SS	i	M	SD	SS
NM_031007	ADCY2	1	++	+	+	4	-	-	-	20	-	-	-
NM_130779	ADCY3	2	++	++	++	3	++	++	+	21	-	-	-
NM_019285	ADCY4	2	++	+	-	11	++	++	-	24	+	-	-
NM_022600	ADCY5	1	+	++	++	3	++	+	+	20	++	-	-
NM_012821	ADCY6	1	++	++	++	2	++	-	-	21	-	-	-
NM_134326	ALB	1	+	-	-	6	++	-	++	14	-	-	-
NM_019288	APP	1	-	++	++	6	+	++	++	17	+	+	+
NM_147141	CACNA1B	1	++	-	-	18	+	-	-	45	++	+	-
NM_153814	CACNA1H	1	-	++	++	5	+	-	-	33	-	-	-
NM_012920	CAMK2A	1	-	+	+	3	-	+	++	10	-	-	-
NM_021739	CAMK2B	1	+	++	++	3	+	+	++	15	++	-	-
NM_012519	CAMK2D	1	-	-	-	4	-	++	++	18	-	++	-
NM_133605	CAMK2G	2	-	-	-	9	++	++	-	18	-	++	++
NM_031334	CDH1	1	++	-	-	3	+	+	++	15	-	++	-
NM_031017	CREB1	2	-	+	-	6	++	+	+	8	-	++	++
NM_052804	FMR1	1	++	++	+	7	+	+	-	15	-	-	-
NM_031028	GABBR1	6	+	++	+	11	-	+	++	21	+	-	-
NM_080587	GABRA4	1	++	-	-	7	-	-	+	8	++	++	++
NM_017289	GABRD	1	+	-	+	2	-	-	-	8	++	-	-
NM_024370	GABRG3	2	-	++	+	5	+	++	++	9	-	++	+
NM_032990	GRIA3	1	++	-	-	4	-	++	++	14	+	++	-
NM_017263	GRIA4	2	-	-	-	4	-	++	++	16	+	-	-
NM_017241	GRIK1	1	+	++	++	3	+	++	++	16	-	++	++
NM_017010	GRIN1	1	++	-	-	8	++	++	++	19	++	-	-
NM_031040	GRM7	1	+	++	++								
NM_012970	KCNA2	1	+	-	-								
NM_031730	KCND2	1	+	++	++	2	-	-	+	4	+	-	++
NM_013066	MAP2	3	++	-	-	4	+	++	++	12	-	+	-
NM_019169	SNCA	2	++	++	++					4	-	++	++
NM_080777	SNCB	2	++	++	-					5	++	-	-
NM_031688	SNCG	1	+	-	-					4	+	-	-
NM_053788	STX1A	1	++	++	+	3	++	+	+	9	++	-	-
NM_012700	STX1B2	1	++	-	-	4	+	++	++	7	++	+	+

[a]Introns by number (i) marked as present with high confidence (++), moderate confidence (+), or absent (-) on microarray (M), Illumina sequencing on dendrite samples (SD), and Illumina sequencing on soma samples (SS).
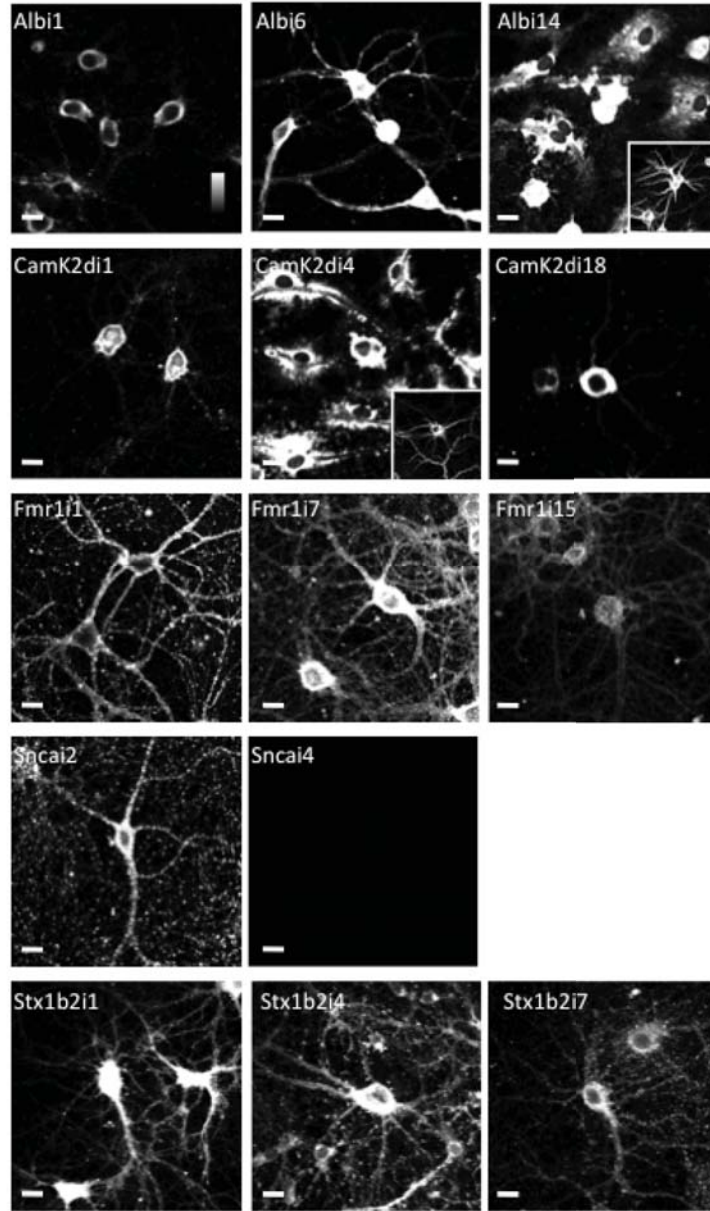
Figure 4.1: Intronic sequences are detectable in dendritic mRNA using *in situ* hybridization. *In situ* hybridization results are shown for 14 intronic riboprobes on paraformaldehyde fixed 14d cultured rat hippocampal neurons. Panels are labeled according to intronic probe detected. Insets represent MAP2 immunostaining. Signal range indicator displayed in top left panel. Scale bars = 20μm.

Some probes (CAMK2Di4, ALBi14) had detectable signals in the cytoplasm of non-neuronal cells that are present in cultures as evidenced by hybridization to MAP2-negative cells (astrocytes), showing that intron retention in the central nervous system is not restricted to neurons and may have functional relevance in other cell types (Figure 4.1).

We also performed additional *in situ* hybridization for exons and retained-introns of dendritically targeted mRNAs. For *Camk2b*, *Fmr1*, *Gabrg3*, and *Grik1*, riboprobes were synthesized corresponding to the exon immediately 5' to a retained-intron and were then used for *in situ* hybridization to rat hippocampal neurons (Figure 4.2). These probes were unique and did not contain any repetitive sequences. In all cases, exon and intron probes are detectable in dendrites, showing that sequences from both coding and non-coding regions within each of these transcripts are localized to dendrites. The distribution patterns of our exon and intron probes show aspects of commonality along with some of distinctions. The exon probes for these targets appear to be discretely localized in puncta compared to their intronic counterparts, which are more diffusely distributed along dendrites. The intron and exon probes for *Fmr1* show the most similarity in distribution pattern to each other among this subset of targets, while the intronic probes for GRIK1i1 are much more intense and dendritically localized than the GRIK1e1 exon probe. These data suggest that intron-retaining transcripts may have different mechanisms of regulation when compared to mature transcripts and when assessed across different mRNAs.

As controls for the *in situ* hybridization studies, we have performed the same types of controls as recommended by the Allen Mouse Brain Atlas (http://mouse.brain-map.org/documentation/index.html). We have used different detection systems (DAB and Quantum Dot), tested for background signal in ISHs performed without probes, and repeated the ISH studies on distinct cultures from different dates of harvesting
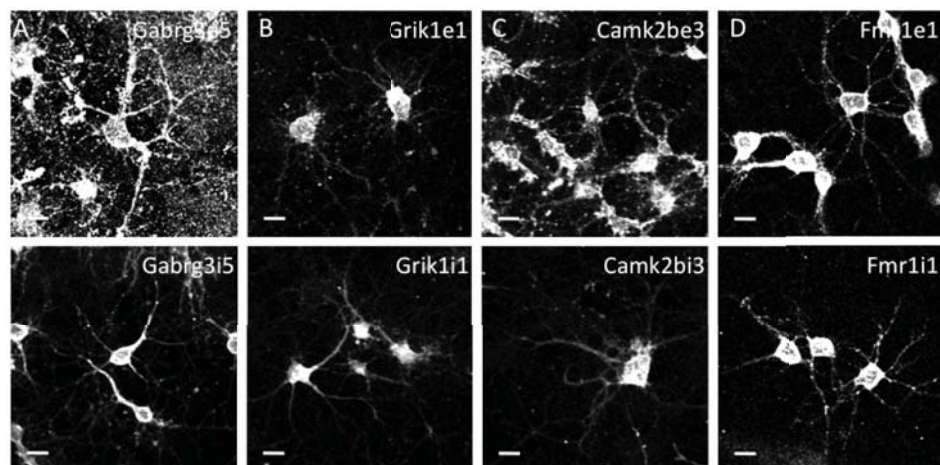
Figure 4.2: *In situ* hybridization of intron and exon riboprobes reveals both distinct and common patterns in neurons. Confocal evaluation of paraformaldehyde fixed 14d cultured primary rat hippocampal neurons hybridized with biotin-labeled riboprobes detected with streptadivin-Qdot605. Panels are labeled according to intronic probe detected. Sequences used are (A) *Gabrg3* exon5 (top), intron 5 (bottom), (B) *Grik1* exon1 (top) intron1 (bottom), (C) *Camk2b* exon3 (top) intron3 (bottom), (D) *Fmr1* exon1 (top) intron1 (bottom). Scale bars = 20μm.

and different litters. Additionally, two controls that directly address the specificity of the ISH signal were performed using other types of probes – short oligo probes and probes to different regions of the RNA – and then assessing the similarity of expression patterns. ISH that is specific should show similar hybridization patterns for each of the sequence-distinct probes. For a subset of the introns probes, we used three different oligonucleotide probes that corresponded to regions of the selected introns and would hybridize to the same area as the longer riboprobe, and obtained similar hybridization patterns (data not shown). Further, for some of the retained introns, we additionally controlled for specificity by using probes corresponding to the exon immediately 5′ to the detected intron (Figure 4.2). The similarity in hybridization signal localization highlights the specificity of the *in situ* hybridization.

4.2.2 Hypothesized retained intron sequence shows an abundance of ID elements

We analyzed the array-positive introns to find sequences that could contain potential regulatory elements. In an initial attempt to find large regions of high sequence similarity, we performed pairwise BLAST between the full genomic sequence for each of the 60 introns with above-median array intensity. We clustered the results using an agglomerative single-linkage method, grouping together alignments on overlapping genomic regions, and obtained 36 sequence clusters, which we number R1 through R36. We annotated these clusters using RepeatMasker [67] and found that all of the clusters except for one were comprised of repetitive sequence (Table 4.3).

Upon further inspection, we noticed that the sequences contained in cluster R4 folded into strong hairpin secondary structures, using a computational structure prediction program [68]. These sequences are all annotated by RepeatMasker as identifier (ID) elements. Although the ID element is not exclusive to the set of retained introns, we became interested in it due to its evolutionary history. ID elements are short interspersed repetitive sequence elements (SINE) originally derived from the noncoding RNA BC1 [69]. They are greatly expanded in the rat genome as a result of active retrotransposition of both the master gene BC1 RNA as well as a number of early progenitor ID elements, which are presumed to have been transcriptionally-active [69]. ID elements share structural similarities with BC1, the $5'$ domain of which has been implicated in dendritic targeting *in vitro* via the presumed folding of its primary sequence into functional secondary structure motifs [70] (Figure 4.3). If ID elements are present in introns that remain intact in the mature transcript, and the ID elements retain the essential structural characteristics of BC1, then perhaps the machinery responsible for targeting BC1 to dendrites can also bind these ID-

Table 4.3: Sequence clusters found in array-positive introns

Cluster	N. Seqs	N. Introns[a]	N. Genes[b]	Repetitive Elements[c]
R1	307	8	8	Alu, B2, B4, ID, L1, L2, ERV1, ERVK, ERVL, MaLR, DNA, MER1, MER2, Low complexity, Simple repeat, Unknown, Other
R2	94	6	6	Alu, B2, ID, MIR, CR1, L1, ERVK, ERVL, MaLR, MER1, Low complexity, Simple repeat, Unknown
R3	35	15	13	Alu, B4, ID, L1, MaLR, scRNA
R4	35	14	12	B2, B4, ID, L1, ERVL, MaLR, scRNA, Simple repeat
R5	21	4	4	L1, MaLR, Unknown
R6	18	11	11	B2, B4, L1
R7	13	3	3	L1, MaLR, Simple repeat
R8	13	9	8	Alu, B2, L1, MaLR, Low complexity, Simple repeat
R9	11	7	7	Alu, B4, ERVL, MaLR, MER1
R10	11	8	7	Alu, B4, L1
R11	7	4	4	L1, MaLR
R12	5	3	3	L1
R13	5	2	2	ID, L1
R14	4	3	3	L1, Simple repeat
R15	4	3	3	B4, L1, ERV1, MaLR, Low complexity, Simple repeat
R16	3	3	3	ERVK
R17	3	2	1	scRNA, Satellite
R18	3	2	2	MaLR
R19	2	2	2	L1, ERVK, Simple repeat
R20	2	2	2	L1
R21	2	2	2	ERVK
R22	2	2	2	ERVK
R23	2	2	2	L1
R24	2	2	2	*none*
R25	2	2	2	ID, Low complexity
R26	2	1	1	L1
R27	2	2	2	B4, ERV1
R28	2	2	2	ERVK, Simple repeat
R29	2	2	2	ERVK
R30	2	1	1	L1
R31	2	2	2	L1
R32	2	2	2	MER1
R33	2	2	2	L1
R34	2	2	2	L1
R35	2	2	2	L1
R36	2	2	2	ERVK

[a]Number of unique introns in the cluster. [b]Number of unique genes in the cluster. [c]Repetitive element families found in the cluster sequences; elements may overlap. Classes include SINE (Alu, B2, B4, ID, MIR), LINE (CR1, L1, L2), LTR (ERV1, ERVK, ERVL, MaLR), and DNA (DNA, MER1 type, MER2 type) elements.

containing transcripts, causing them to be localized to dendrites as well. However, in order for an ID-element mediated dendrite-localization mechanism to be plausible, we needed to verify that ID elements are in fact retained, and that ID elements possess targeting competency.
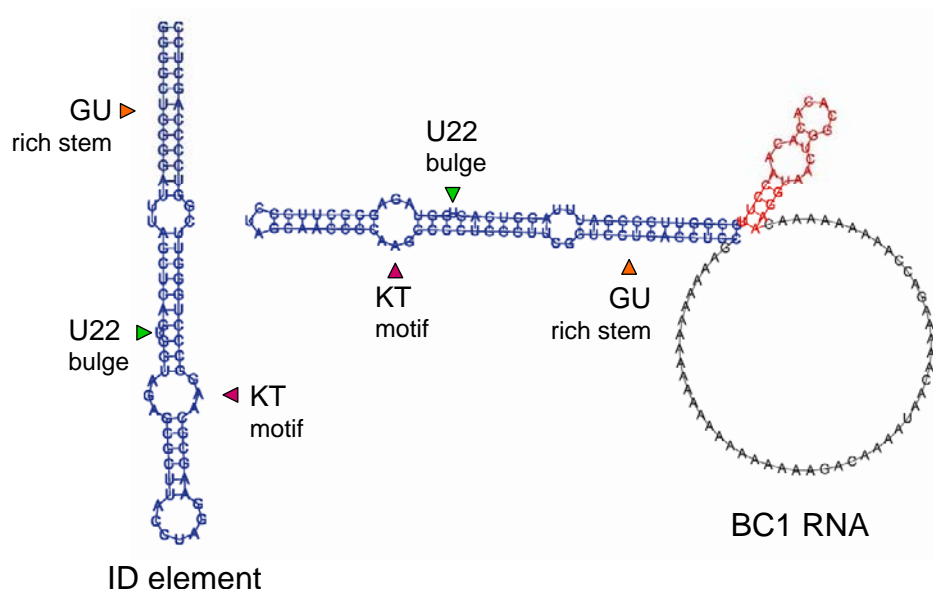
Figure 4.3: Secondary structures of the ID element and BC1 RNA. BC1 5′ domain is colored blue, 3′ domain is colored red. Corresponding motifs on the ID element and BC1 5′ domain are labeled.

4.2.3 SHORT READ SEQUENCING CONFIRMS EXTENSIVE INTRON RETENTION

To verify intron retention and to determine whether specific ID-containing loci are retained, we performed Illumina (formerly Solexa) short-read sequencing on RNA material isolated from primary rat hippocampal neurons. We performed sequencing runs on five single cell soma and four groups of pooled RNA each from 150-300 individually-dissected dendrites using paired-end technology (dendrite samples D1-D3 and soma samples S1-S3) and single-read sequencing (dendrite sample D4 and soma

Table 4.4: Summary of short read sequencing results

| | Soma samples | | | | |
	S1	S2	S3	S4	S5
Read length	50	50	50	36	42
Technology	Paired	Paired	Paired	Single	Single
Total number error-free reads	24,054,234	28,360,642	27,723,528	11,177,563	11,256,670
Genome-wide uniquely matching reads	5,410,480	7,422,480	7,427,020	2,223,767	2,685,506
Gene-overlapping reads	2,753,892	3,389,382	3,772,206	908,405	1,118,723
Num. genes with read coverage	6075	5745	5865	10494	10534
Num. genes with intron read coverage	3219	2880	2430	8262	8260

| | Dendrite samples | | | |
	D1	D2	D3	D4
Read length	50	50	50	42
Technology	Paired	Paired	Paired	Single
Total number error-free reads	25,923,420	25,428,726	21,647,526	11,463,613
Genome-wide uniquely matching reads	9,830,588	4,642,310	8,701,334	2,808,693
Gene-overlapping reads	5,208,384	2,348,138	4,762,266	1,044,267
Num. genes with read coverage	8584	6351	9242	12280
Num. genes with intron read coverage	4925	2563	5370	10678

samples S4-S5) (Table 4.4). Each sample underwent three rounds of amplification using the aRNA protocol [71].

For the 33 genes of interest, we performed specific read alignment using Bowtie [72] (see Materials and Methods). All of these genes were detectable based on the presence of at least one read uniquely matching the exonic region, with a median coverage of eight reads per detectable exon. Additionally, we found a large number of reads uniquely aligning to intronic sequence (Table 4.2). Of these, a large number of microarray positives were independently verified. Over half (31 of 60) of the introns with detectable sequences by microarray show a high level of sequencing coverage. An additional four introns are also detectable using less stringent criteria (see Materials and Methods) (Table 4.2).

The difference in array versus sequencing results is likely due to the coverage difference of the two platforms. The spotted microarray elements cover ~ 500 bases of intronic sequence and can anneal to any intronic sequence within that region whereas

the Illumina reads cover only ~ 50 bases of sequence information, and in our Illumina analysis we conservatively considered only uniquely aligning reads, ignoring non-unique genomic sequences, which are highly prevalent in intronic sequences.

Reads were evaluated across our dendrite and soma samples and scored as absent, present, or highly represented based on the number of reads found for a given intron. Forty-three introns represented on our microarray were scored as absent, showing no detectable uniquely-aligning read coverage within that intron in any dendrite sample. Eighteen introns were scored as present having from one to eight uniquely-aligning reads in at least one sample, while 33 introns were scored as highly represented based on the presence of greater than eight uniquely-aligning reads from that intron in at least one sample (Table 4.2)

We considered whether these results are due to the presence of actual intronic sequence in the transcriptome, or whether non-transcribed genetic material was included with our samples; however, several factors in our assays and experimental design suggest that genomic DNA contamination is highly unlikely. Dendrites are mechanically isolated from cell bodies, thereby preventing any genomic DNA from contaminating the sample. If this were not the case, the microarray results would show uniform signal across all probes, indicating non-specific intron sequence detection. Since there is heterogeneity in microarray signal across different introns of the same gene, it is likely that these signals are specific.

We also analyzed the Illumina sequence data for evidence of non-specific read coverage. For each of the 33 genes of interest, we calculated the cumulative intron length and then obtained roughly equal amounts of upstream and downstream intergenic sequence to serve as a background sequence set. Intergenic sequences were chosen to minimize overlap of known and predicted transcribed units, as well as repetitive sequence. When we compared unique read alignment in six of the Illumina sequenc-

Table 4.5: Sequence coverage in exonic, intronic, and intergenic regions

| | Dendrite samples | | | Soma samples | | |
	D1	D2	D3	S1	S2	S3
N. reads per 1000 nts						
Exonic	29.15	29.51	66.28	45.58	38.48	54.62
Intronic	0.34	0.09	0.25	0.84	2.76	0.83
Intergenic	0.30	0.08	0.08	0.03	0.47	0.17
p-value (Binomial test)						
Exonic vs intergenic	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$
Intronic vs intergenic	$2\mathrm{x}10^{-7}$	$1\mathrm{x}10^{-2}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$	$< 2\mathrm{x}10^{-16}$

ing runs (three dendrite, three soma) onto the intergenic regions compared to the genic regions, we found enrichment in both intronic ($p \leq 0.01$ by the Binomial Proportion Test) and exonic regions ($p < 2$ x 10^{-16}), indicating that read alignments within the gene boundaries are unlikely to arise from non-transcribed sequence (Table 4.5). Although we took efforts to ensure the intergenic sequences are taken from transcription-free regions, it is not possible to guarantee that previously uncharacterized transcripts are not present in those loci, so we treat this statistic to be a conservative measure of background read coverage. In fact, inspection of individual instances of intergenic read coverage reveals large amounts of localized coverage, strongly suggesting the presence of previously uncharacterized transcription units (data not shown).

We additionally found evidence for extensive genome-wide intron retention. On average, about 60 percent of detectable genes throughout the sequencing experiments had unique intronic read coverage, with comparable proportions in both the dendrite and soma samples (Table 4.4).

4.2.4 SPECIFIC ID ELEMENT-CONTAINING LOCI HAVE SEQUENCING SUPPORT

We further analyzed the read alignments to determine whether there was evidence for transcription spanning any of the ID-containing loci in our introns of interest. ID-

element sequence occurs in high copy number throughout the rat genome; therefore, we would not expect to be able to uniquely align reads to every potentially present ID locus.

Using RepeatMasker and BLAST search algorithms, we were able to identify a total of 308 blocks of ID-derived sequence across our focused set of 33 genes. Based on prior research that characterized the functionally significant components of the BC1 RNA localization domain [73], we defined a subset of 136 of these ID elements to be targeting-competent. Each of these ID elements has at least 90 percent alignable nucleotide sequence to the 5′ BC1 domain; is computationally predicted to form a hairpin in its minimum-free-energy secondary structure configuration; and contains an unpaired uracil at nucleotide position 22 in a basal-medial unbranched helix, which is necessary for BC1 localization. Of the 136 targeting-competent ID elements, 70 are found in the sense direction relative to the direction of the gene, and 37 of the 70 are found in our introns of interest. All 37 ID elements have extremely high self containment [74] with an average SC index of 0.9, indicating that the ID elements are robust substructures in the intron.

Of these 37 ID elements, all but one (CAMK2Gi2ID1) are contained in predicted retained introns that have unique short-read sequence coverage; 31 additionally have microarray support (Table 4.6). Nineteen of these ID elements have *cis* sequencing evidence, with reads uniquely aligning to positions within one read length (50 nucleotides) of the element. Eight of these are spanned by uniquely-aligning mate pairs, providing the most direct evidence that these specific ID-element-containing loci are in fact present in the RNA samples (Table 4.6).

Beyond our introns of interest, there is sequencing evidence for an additional 16 of the 70 ID elements in our gene set, six of which are spanned by mate pairs. Genome-wide we were able to find a total of 3658 unique ID loci spanned by a total

Table 4.6: ID elements found in candidate introns

RefSeq	Intron	Coordinate	ID element	M^a	S-I^b	S-50^c	S-MP^d
NM_031007	1	chr17:4557075-4557148	ADCY2i1ID1	++	+		
NM_022600	1	chr11:67398142-67398213	ADCY5i1ID1	+	++	+	+
NM_022600	3	chr11:67338681-67338754	ADCY5i3ID1	++	+	+	
NM_021739	1	chr14:86712517-86712590	CAMK2Bi1ID1	+	++	+	+
NM_021739	3	chr14:86667060-86667132	CAMK2Bi3ID1	+	++		
NM_012519	4	chr2:224012159-224012231	CAMK2Di4ID1	-	++	+	
NM_133605	2	chr15:3748334-3748403	CAMK2Gi2ID1	-			
NM_133605	18	chr15:3784885-3784955	CAMK2Gi18ID1	-	++	+	+
NM_052804	1	chrX:154759680-154759747	FMR1i1ID1	++	++	+	
NM_024370	5	chr1:108287957-108288025	GABRG3i5ID1	+	++		
NM_024370	5	chr1:108285219-108285291	GABRG3i5ID2	+	++	+	
NM_024370	5	chr1:108273028-108273101	GABRG3i5ID3	+	++		
NM_024370	5	chr1:108394161-108394234	GABRG3i5ID4	+	++	+	
NM_032990	4	chrX:3569721-3569788	GRIA3i4ID1	-	++	+	+
NM_017263	4	chr8:1236057-1236129	GRIA4i4ID1	-	++		
NM_017241	1	chr11:27943131-27943204	GRIK1i1ID1	+	++		
NM_017241	1	chr11:27913623-27913697	GRIK1i1ID2	+	++	+	
NM_017241	1	chr11:28081118-28081190	GRIK1i1ID3	+	++	+	
NM_017241	1	chr11:28044738-28044811	GRIK1i1ID4	+	++		
NM_017241	1	chr11:28098620-28098693	GRIK1i1ID5	+	++		
NM_017241	16	chr11:27709224-27709296	GRIK1i16ID1	-	++		
NM_017010	8	chr3:3462386-3462457	GRIN1i8ID1	++	++	+	
NM_017010	8	chr3:3461694-3461767	GRIN1i8ID2	++	++	+	+
NM_031040	1	chr4:146624192-146624265	GRM7i1ID1	+	++	+	
NM_031040	1	chr4:146736661-146736734	GRM7i1ID2	+	++		
NM_031040	1	chr4:146746191-146746264	GRM7i1ID3	+	++		
NM_031040	1	chr4:146760671-146760744	GRM7i1ID4	+	++		
NM_031040	1	chr4:146937318-146937390	GRM7i1ID5	+	++	+	+
NM_031040	1	chr4:146951304-146951377	GRM7i1ID6	+	++		
NM_031040	1	chr4:146433109-146433182	GRM7i1ID7	+	++		
NM_031730	1	chr4:47841144-47841217	KCND2i1ID1	+	++	+	+
NM_031730	1	chr4:47972539-47972610	KCND2i1ID2	+	++		
NM_031730	1	chr4:47998363-47998436	KCND2i1ID3	+	++	+	
NM_031730	1	chr4:48009814-48009886	KCND2i1ID4	+	++		
NM_031730	1	chr4:47642184-47642253	KCND2i1ID6	+	++	+	+
NM_019169	2	chr4:89713754-89713825	SNCAi2ID1	++	++		
NM_012700	4	chr1:187098603-187098671	STX1B1i4ID1	+	++	+	

^{a}ID elements marked as present with high confidence (++), moderate confidence (+), or absent (-) on microarray (M); ^{b}Illumina sequencing read coverage in the containing intron, ^{c}within 50 nucleotides of the ID locus, and ^{d}spanning the ID locus with mate pairs.

Table 4.7: Short-read sequence coverage for intronic ID element loci

	Dendrite samples			Soma samples			
	D1	D2	D3	S1	S2	S3	Overall
Num. uniquely aligning reads spanning intronic ID elements							
Sense[a]							
Targeting-competent	12,440	3,169	3,919	26,287	9,317	8,652	63,784
Non-competent	3,302	396	692	3,660	1,652	440	10,142
Antisense	4,242	3,190	2,371	3,780	4,064	2,371	20,018
Num. individual ID loci with read coverage							
Sense							
Targeting-competent	770	144	533	584	275	163	2046
Non-competent	137	26	99	103	40	23	365
Antisense	396	148	388	173	167	91	1247
Num. genes containing ID loci with read coverage							
Sense							
Targeting-competent	697	143	505	526	264	155	1617
Non-competent	137	27	99	100	40	28	345
Antisense	385	158	376	163	166	90	1109

[a]Strand of the ID element with respect to the gene.

of 63,784 mate pairs in six sequencing experiments, contained in 2590 genes; 2411 of these loci occur in the sense direction, and of these 2046 are predicted to be targeting competent (Table 4.7). The 60 intronic loci supported by three or more sequencing runs are summarized in Tables 4.8 and 4.9.

These data show a significant enrichment in sense-direction ID elements compared to antisense-direction ID elements, 1.93 fold, compared to the genomic occurrence of ID elements, which favors antisense elements by 1.43 fold (see Section 4.2.8). We note that it is possible that antisense elements are also functionally significant despite their lack of the correct BC1 targeting features; however, in the absence of such functionality, we would expect antisense elements to be present as part of the retained intron in which they occur, whose retention is mediated by other factors.

Table 4.8: Targeting-competent sense-strand ID element loci retained in a majority of soma samples

Coord		RefSeq: intron	Symbol	Num. samples[a]	
				Dendrite	Soma
chr11:66544722-66544790	+	NM_001029903:i5	Fam162a	3	3
chr4:66045882-66045955	+	NM_001134553:i13	Ubn2	2	3
chrX:89735678-89735751	+	NM_017107:i5	Ogt	1	3
chr4:149005586-149005659	+	NM_001106614:i18	Setd5	1	3
chr3:11440674-11440746	-	NM_080689:i18	Dnm1	2	2
chr10:75491508-75491562	+	NM_001108288:i22	Trim37	2	2
chr3:159115619-159115692	+	NM_012637:i8	Ptpn1	2	2
chr10:71576187-71576260	+	NM_001105824:i6	Taf15	2	2
chr1:85079016-85079089	-	NM_001100991:i2	LOC499124	2	2
chr6:25465217-25465290	-	NM_001126372:i11	LOC362710	2	2
chr8:77257195-77257267	+	NM_012986:i1	Nedd4	2	2
chr12:34866906-34866979	+	NM_031338:i2	Camkk2	1	2
		NM_001080147:i7	Anapc5		
chr7:127871488-127871560	+	NM_021676:i21	Shank3	1	2
chr3:34693830-34693901	-	NM_001106480:i5	Prpf40a	1	2
chr5:142363851-142363924	+	NM_001108676:i4	Trit1	1	2
chr7:2154725-2154798	+	NM_001033070:i9	Sarnp	1	2
chr13:40909312-40909385	+	NM_001134867:i19	R3hdm1	1	2
chr8:95142917-95142990	-	NM_001108175:i4	Tbc1d2b	1	2
chr9:59052166-59052231	+	NM_173143:i6	Abi2	1	2
chr1:80288452-80288525	-	NM_012506:i16	Atp1a3	1	2
chr7:117598975-117599048	-	NM_001130581:i1	LOC685444	1	2
chr6_random:442356-442429	-	NM_017359:i3	Rab10	1	2
chr12:33774296-33774369	-	NM_001134766:i4	Ccdc62	1	2
chr17:48502769-48502841	-	NM_001107354:i1	Hist1h2an	1	2
chr15:27752648-27752720	-	NM_001024794:i1	Mettl3	1	2
chr19:53569106-53569179	-	NM_001107440:i4		1	2
chr7:120656200-120656273	+	NM_019375:i2	Sept3	1	2
chr5:65069567-65069640	+	NM_001107932:i6	Invs	1	2

[a]Number of dendrite and soma sequencing experiments in which unique read coverage overlapped the ID element locus.

145

Table 4.9: Targeting-competent sense-strand ID element loci retained in a majority of dendrite samples

| | | | | Num. samples[a] | |
Coord		RefSeq: intron	Symbol	Dendrite	Soma
chr1:168012312-168012385	-	NM_001134970:i27	LOC691036	3	2
chr20:3614357-3614429	-	NM_001002807:i5	Clic1	3	2
		NM_133300:i6	Bat1		
chr6:74989905-74989978	-	NM_001134987:i6	Eapp	3	2
chr14:104496291-104496363	-	NM_001100971:i16	RGD1305110	3	2
chr18:74627136-74627209	-	NM_001107374:i1	RGD1308601	3	1
chr8:118512730-118512803	+	NM_053722:i17	Clasp2	3	1
chr5:91974244-91974317	+	NM_001106662:i12	Frmd3	3	1
chr17:11151503-11151576	+	NM_001106100:i26	Agtpbp1	3	0
chr7:115353524-115353596	-	NM_001079895:i1	Rbm9	3	0
chr5:130378024-130378096	+	NM_012993:i7	Nrd1	2	1
chr10:64382492-64382565	-	NM_001105803:i1	Sdf2	2	1
chr7:1658380-1658453	-	NM_001108727:i2	Coq10a	2	1
chr3:11440038-11440111	-	NM_080689:i18	Dnm1	2	1
chr13:107842967-107843040	+	NM_001109376:i5	LOC679692	2	1
		NM_001047894:i3	LOC317456		
chr10:71575745-71575818	+	NM_001105824:i6	Taf15	2	1
chr19:10031165-10031238	+	NM_001107409:i4	Csnk2a2	2	1
chr3:41009974-41010047	+	NM_001106482:i4	Pkp4	2	1
chr5:152641779-152641852	+	NM_001109358:i1	RGD1566319	2	1
chr1:226509592-226509665	-	NM_001106357:i17	Smc5	2	1
chr12:46591220-46591293	-	NM_001047901:i3	Ankle2	2	1
chr1:82534406-82534479	+	NM_001106236:i1	Blvrb	2	1
chr1:107337799-107337872	-	NM_001107518:i6	Nipa2	2	1
chr5:61407500-61407573	+	NM_001106658:i4	Zcchc7	2	1
chr1:56509832-56509904	-	NM_172323:i2	Has1	2	1
chr1:233830886-233830959	+	NM_001107585:i10	Uhrf2	2	1
chr9:84086618-84086691	+	NM_001108804:i1	Fbxo36	2	1
chr8:34997093-34997166	+	NM_001034150:i13	Srpr	2	1
chr11:60755156-60755228	-	NM_017242:i1	Lsamp	2	1
chr5:148336165-148336237	-	NM_001134628:i5	RGD1564943	2	1
chr8:113101807-113101871	-	NM_001108186:i5	Rbm6	2	1
chr20:19735536-19735609	-	NM_031805:i1	Ank3	2	1
chr3:89453560-89453633	-	NM_001109606:i10	Fbxo3	2	1

[a]Number of dendrite and soma sequencing experiments in which unique read coverage overlapped the ID element locus.

4.2.5 ID ELEMENT SEQUENCE IS ENRICHED IN SEQUENCING READS

Since we used a conservative policy of only considering unique read matches to the genome, a large number of reads are unassigned to genomic loci; on average, only about 30 percent of the reads per sequencing experiment matched uniquely to the genome (Table 4.4). Correspondingly, it is not possible to assign reads to all genomic loci containing ID elements if the surrounding sequence context is not genomically unique. Thus, to quantify the amount of excess read coverage purported to derive from some ID-element containing locus, we created BLAST databases using the full set of reads for each of the sequencing runs and performed nucleotide BLAST querying a prototypical rat ID element sequence from RepBase [75]. In each of the samples, a large number of reads ($\sim$ 30,000 - 200,000) had significant similarity to the 74-nt ID element hairpin with e-value < 0.001 (Table 4.10). To ensure that these reads were matching ID elements and not to BC1 RNA, we also queried the BC1 3$'$ domain (78 nt) and found only a single match across all sequence runs; thus, the vast majority of the ID-matching reads must derive from ID-element loci in the genome. Compared to the number of mate-pair results from the previous section, these data suggest that the degree of ID locus retention is somewhere between 7 and 34 fold under-reported using a policy of only allowing uniquely-aligning reads (Table 4.10).

We also compared the degree of read coverage of ID elements to that of B2 elements, another ubiquitous SINE element in the rat genome that occurs approximately 2.18 times more frequently than the ID element as annotated by RepeatMasker. BLAST matches to the B2 element were found throughout the sequencing runs, but at a much lower frequency than for the ID elements (Table 4.10). When normalized by genomic frequency, ID elements have a 10 to 28 fold greater proportion of aligning sequence reads than B2 elements. Assuming that the Illumina sequencing data is an

Table 4.10: Sequence reads aligning to ID elements

	Soma samples				
	S1	S2	S3	S4	S5
Total number reads	24,054,234	28,360,642	27,723,528	11,177,563	11,256,670
Num. reads aligning					
ID 5$'$ hairpin (74 nt)	205,658	105,821	78,824	30,259	60,855
BC1 3$'$ domain (78 nt)	0	0	0	0	1
B2 5$'$ portion (74 nt)	16,170	17,768	9,494	5,561	11,374
Genomic coverage[a]					
ID (161,321 total)	1.275	0.656	0.489	0.188	0.377
B2 (352,447 total)	0.046	0.050	0.027	0.016	0.032
Fold enrichment ID > B2	27.8	13.0	18.1	11.9	11.7

	Dendrite samples			
	D1	D2	D3	D4
Total number reads	25,923,420	25,428,726	21,647,526	11,463,613
Num. reads aligning				
ID 5$'$ hairpin (74 nt)	101,981	109,919	75,227	66,127
BC1 3$'$ domain (78 nt)	0	0	0	0
B2 5$'$ portion (74 nt)	11,793	8,603	11,143	14,062
Genomic coverage				
ID (161,321 total)	0.632	0.681	0.466	0.410
B2 (352,447 total)	0.033	0.024	0.032	0.040
Fold enrichment ID > B2	18.9	27.9	14.7	10.3

[a]Total number of elements in the genome divided by number of aligning sequencing reads. Genome element totals are as reported by RepeatMasker.

accurate representation of the rat transcriptome, these results imply a large number of transcripts contain ID element sequences, more than would be expected from any sort of background or random transcription phenomenon.

4.2.6 IN SITU ANALYSIS REVEALS TARGET COMPETENCY OF INDIVIDUAL ID ELEMENTS

We chose several ID elements from retained introns on the basis of host gene and structural characteristics for further analysis. PCR primers were designed to amplify these ID elements using genomic DNA as the template. Amplified fragments include the ID sequences (approximately 74bp) and approximately 500-1000bp of flanking sequence. The lengths of these PCR products were determined based on the availability of well-defined genomic sequence in the regions of interest. ID PCR products were subcloned into pCRII TOPO vectors and then further subcloned into pEGFP-N1 expression vectors (CMV promoter driven) to generate ID-EGFP transcripts upon transfection into primary rat hippocampal neurons (Figure 4.4a).

ID-EGFP vectors were transfected into primary cultures of neurons, and 48 hours later the cells were analyzed by *in situ* hybridization using probes directed to the EGFP portion of the exogenously expressed mRNA (Figure 4.4b). The *in situ* results in the figures, and quantified below, suggest that ID elements from the retained introns can confer dendritic targeting to the transgene mRNA.

To quantify the targeting capacity of the fusion construct we developed a custom program using Igor (WaveMetrics, Inc.) to measure probe intensity along a curve drawn in the *in situ* images. The quantification paths are manually drawn, tracing dendrites of selected cells based on MAP2 immunostaining. Paths originate at the somal end of the dendritic process. For each of the assays described below three dendrites were quantified per cell and 8 or 10 cells were quantified for each probe.
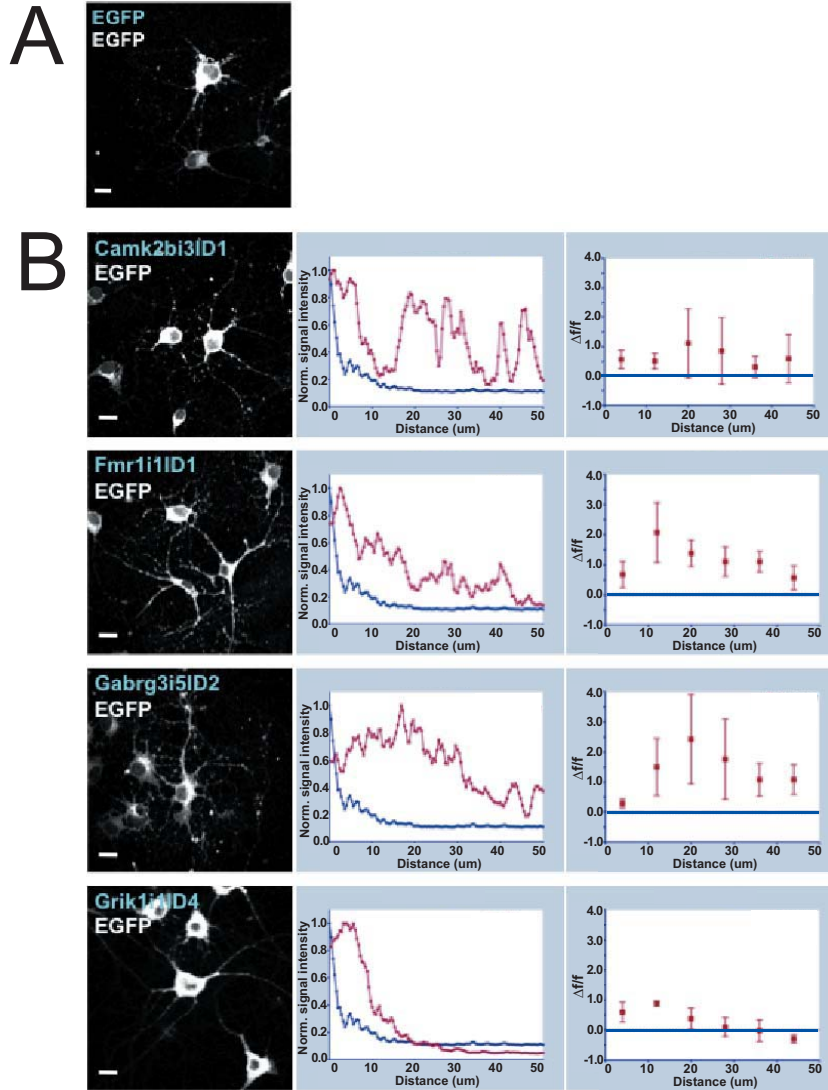
Figure 4.4: (a) pEGFP-N1 transfected control cells *in situ* hybridized with antisense biotinylated EGFP probe. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. (b) ID-EGFP targeting *in situ* hybridization results. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. A representative set of signals is shown for imaging results along with a plot of normalized signal intensity against distance from the cell soma in microns. Blue lines represent EGFP control signal, red lines represent ID-EGFP signal. Each transfected ID-EGFP experiment (red boxes) is then binned to find average intensity values across 8μm distances and subtracted from binned EGFP signal to generate mean and SEM values to distances of 48μm. These values are plotted as Δf/f against distance in microns. Blue lines represent level of EGFP signal. Scale bars = 20μm.

Comparing the relative distances in microns from the cell body at which *in situ* signal can be detected for an ID-EGFP transfected cell to an EGFP transfected control cell shows that ID-containing transcripts are localized to more distant regions along the dendrite (Figure 4.4c). A greater level of signal can be seen in ID-EGFP transfected cells at distances of 20μm for all four ID elements presented. At distances beyond 20μm, only signals for FMR1i1ID1 and GABRG3i5ID2 have mean and SEM values that are greater than EGFP transfected cells. These signals continue to be greater than control out to distances beyond 44μm along the length of dendrite. However, as 44μm is the length of the shortest dendrites analyzed in this study and we restricted our statistical analysis to this length of dendrite.

The main signature of RNA transport for our probes is in both the differential intensity levels and differential gradient of intensity along the dendrites. Actively transported RNAs are expected to have greater intensity and a shallower gradient along the length of the dendrite while non-transported RNA are expected to have less intensity and steeper gradients. We first tested the differential intensity levels along the dendrites (after normalizing the highest intensity pixel to 1). The RNA intensity differences were pooled in 8 μm intervals and paired t-tests were carried out to assess the significance of the difference between the test probe and control EGFP probe within each interval. The resulting set of t-tests may not be independent due to shared residuals from a gradient-like generating process along the dendrites. Therefore, we carried out a conservative Bonferroni correction for non-independent multiple tests. The significance of the overall differences along the entire dendrite(s) was assessed using Fisher's combined p value test for the Bonferroni corrected t-test p values. The Fisher's combined p statistic and probability of Fisher's combined p values were as follows: CAMK2Bi3ID1 (2*LogLikelihood = 80.36, p < 10^{-11}), FMR1i1ID1 (2*LogLikelihood = 83.61, p < 10^{-11}), GABRG3i5ID2 (2*LogLikelihood

151

= 83.24, p < 10^{-11}), GRIK1i1ID4 (2*LogLikelihood = 58.06, p < 10^{-7}).

We next tested the differential gradient by fitting the entire probe intensity curve to a negative hyperbolic function of the form $I = c - sd/(g + d)$, where I represents probe intensity and d represents distance from soma. The parameters c and s represent translation and scale of the curve with $(c - s)$ forming the asymptote of the curve. The parameter g represents the steepness of the curve – i.e., the steepness of the gradient – and is therefore the parameter of interest. The ISH signals for control EGFP probes (n = 8) and test probes (n = 10, each) were fitted using a nonlinear least-squares fitting procedure (R statistical package). The 95 percent confidence interval for the parameter g are: EGFP = 1.63±0.241, CAMK2 = 2.32±0.420, FMR1 = 4.94±1.133, GABR = 3.56±0.499, GRIK1 = 4.96±0.866. Thus, EGFP forms a significantly steeper gradient along the dendrites than any of the four quantified test probe ISHs suggesting more active transport of the mRNA corresponding to the test probes. It should be noted that the parameter estimate g provides an overall expression level independent assessment of the RNA gradient (because of the other fitted parameters). In effect, fitting a hyperbolic curve and then testing the steepness parameter establishes the spatial pattern as self-control that is invariant of expression levels or probe specific effects.

A variety of distribution patterns can also be observed across these distances and can best be described as diffuse (GABRG3i5ID2), punctate (CAMK2Bi3ID1) and intense (FMR1i1ID1). These findings suggest different targeting mechanisms for ID-containing sequences that may be governed by flanking sequence or subtle sequence/structural variations across the elements presented. Regardless of the mechanisms involved, an *in situ* signal can clearly be seen at greater distances from the cell soma in the dendrites of ID-EGFP expressing cells when compared to EGFP expressing cells. These results show that intronic ID elements can act as dendritic-targeting

elements for exogenously expressed fusion constructs.

The sequence required for independent folding of the ID element is approximately 74nt; however, hundreds of bases of flanking sequence are present in the "full-length" ID-EGFP fusion constructs. In an effort to narrow the amount of sequence required for dendritic targeting of reporter gene mRNA, constructs were generated with significantly less intronic flanking sequence than the full-length ID element PCR products. Using PCR primers designed to anneal approximately 40 bases upstream and downstream of the 74 base intronic ID element, PCR products were amplified which were only 137-152 bases in length (Figure 4.5a). These products were cloned into pEGFP-N1 expression vectors and assessed for targeting capacity by *in situ* hybridization with probes targeted at the EGFP sequence (Figure 4.5b).

All of the four discrete ID element constructs tested conferred significant dendritic targeting to exogenously expressed reporter transcripts compared to control EGFP transfected cells using the Fisher and Bonferroni statistical approach previously described (Figure 4.5c). The Fisher's combined p statistic and probability of Fisher's combined p values were as follows: CAMK2Bi3ID1dis (2*LogLikelihood = 71.99, p $< 10^{-9}$), FMR1i1ID1dis (2*LogLikelihood = 89.02, p $< 10^{-11}$), GABRG3i5ID2dis (2*LogLikelihood = 74.19, p $< 10^{-10}$), GRIK1i1ID4dis (2*LogLikelihood = 74.19, p $< 10^{-10}$). In all cases except GABRGi5ID2dis, discrete ID sequences have mean plus SEM signals greater than control at greater distances than their full-length counterparts. This may result from removing flanking intronic sequences that are important to dendritic targeting as separate targeting elements or as enhancers of this particular ID elements ability to target. In all cases, signals above control can be seen to distances of 28μm, for CAMK2Bi3ID1dis and FMR1i1ID1dis mean plus SEM signals can be seen out to 44μm.
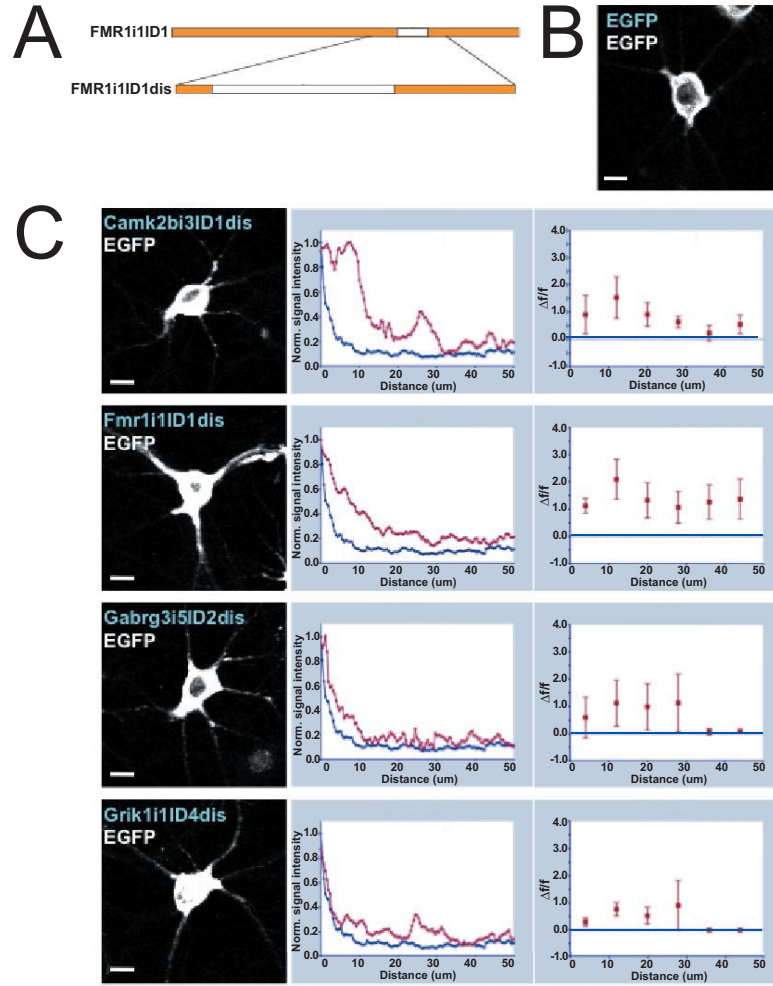
Figure 4.5: Intronic ID element sequences confer dendritic localization to reporter gene mRNA with minimal flanking sequences. (a) Schematic of discrete ID element generation. (b) pEGFP-N1 transfected control cells *in situ* hybridized with antisense biotinylated EGFP probe. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. (c) Discrete ID-EGFP targeting *in situ* hybridization results. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. A representative set of signals is shown for imaging results along with a plot of normalized signal intensity against distance from the cell soma in microns. Blue lines represent EGFP control signal, red lines represent ID-EGFP signal. Each transfected ID-EGFP experiment (red boxes) is then binned to find average intensity values across 8μm distances and subtracted from binned EGFP signal to generate mean and SEM values to distances of 48μm. These values are plotted as Δf/f against distance in microns. Blue lines represent level of EGFP signal. Scale bars $= 20\mu$m.

4.2.7 Transgenic intronic ID elements compete with endogenous transcripts for targeting machinery

Exogenous expression of ID elements was also used to assess their capacity to modify the localization of endogenous transcripts in an *in vivo* competition assay. Neurons were transfected with full-length ID-EGFP fusion constructs. Forty-eight hours post-transfection, *in situ* hybridization was performed using probes directed at the intronic region used for microarray analysis and absent in the ID-EGFP transcripts. Only endogenous transcripts containing this region of the introns will be detected, thus allowing the study of the ID-EGFP transcript's effect on endogenous intron-containing mRNA (Figure 4.6). In all cases tested, transfection of the full-length ID-EGFP fusion constructs disrupted the localization of their analogous endogenous intron-retaining transcripts compared to transfection with EGFP. Fisher and Bonferroni statistical analysis was performed as described previously for our targeting experiments, and again showed significant differences. The degree of competition and pattern of remaining dendritic signal for endogenous intron-retaining transcript ranges from a 0.2 to 0.4 fold decrease in dendritic signal at all distances along the dendritic length where endogenous intron-retaining transcripts are detectable. These data show that exogenous expression of an ID element from a particular intron limits the dendritic localization of endogenous transcripts retaining that intron.

Given the effect of exogenous expression of an ID-EGFP construct on related intron-retaining transcripts, cross-competition experiments were also performed to assess the potential for a transfected ID element to disrupt the localization of all intron-retaining transcripts or select intron-retaining transcripts. Exogenously expressed ID-EGFP sequences will compete with endogenous transcripts for targeting machinery as described previously, and probing for an unrelated intron-retaining tran-
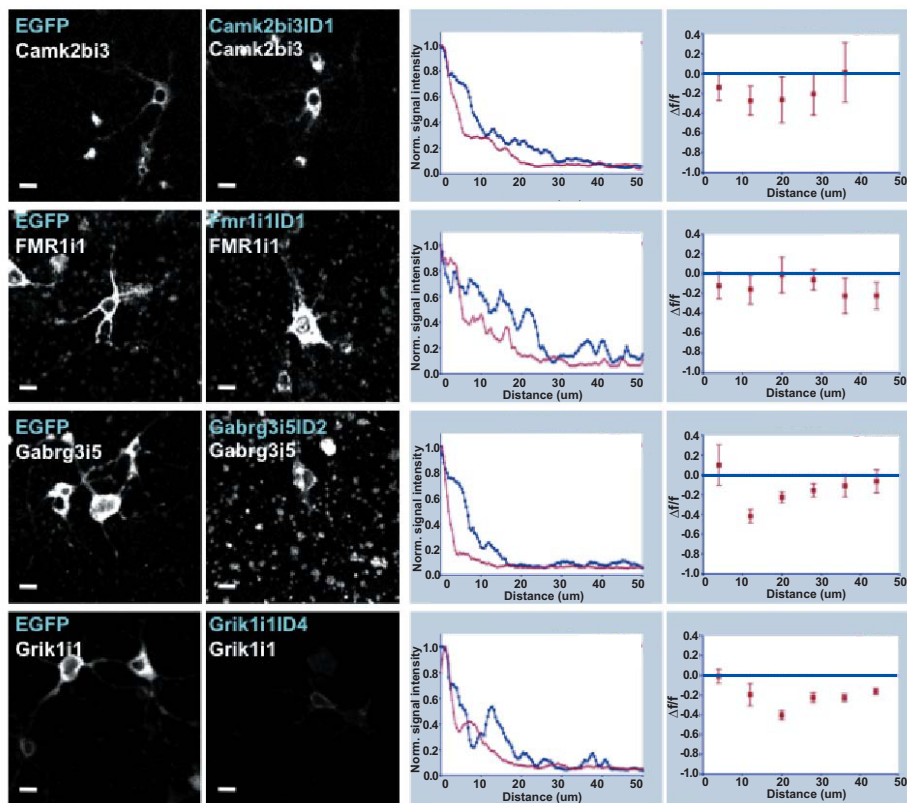
Figure 4.6: Intronic ID element sequences disrupt dendritic localization patterns of endogenous mRNA. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. A representative set of signals is shown for imaging results along with a plot of normalized signal intensity against distance from the cell soma in microns. Blue lines represent control signal for endogenous transcripts following EGFP transfection, red lines represent level of endogenous signal following ID-EGFP transfection. Each transfected ID-EGFP experiment (red boxes) is then binned to find average intensity values across 8μm distances and subtracted from binned EGFP signal to generate mean and SEM values to distances of 48μm. These values are plotted as Δf/f against distance in microns. Blue lines represent level of endogenous signal following EGFP transfection. Scale bars = 20μm.

script will show if the mechanisms governing dendritic targeting are specific to particular ID elements or common to all targeted transcripts that use this motif. This was tested using probes to introns from genes that do not contain the particular ID element being exogenously expressed (Figure 4.7).
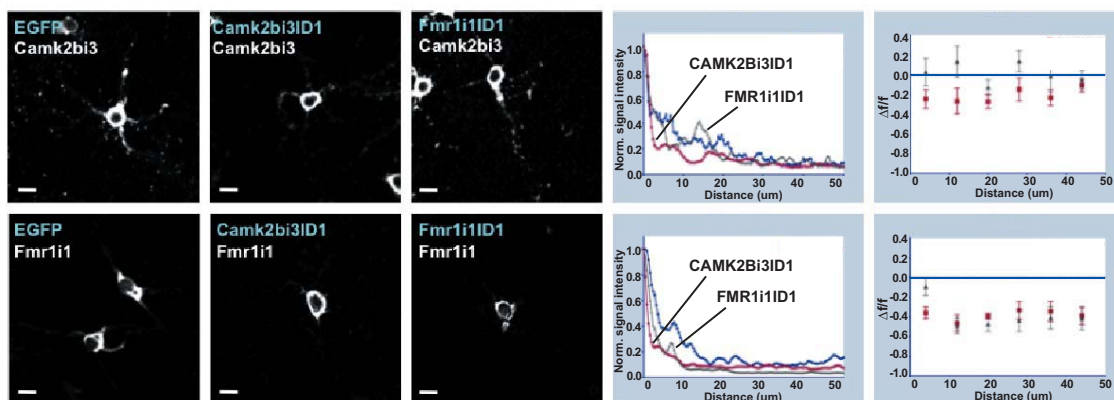
Figure 4.7: Intronic ID element sequences differentially cross-compete with endogenous mRNA of different genes. CAMK2Bi3ID1 disrupts localization of FMR1i1 intron-retaining transcript, but FMR1i1ID1 does not disrupt localization of CAMK2Bi3. Blue text indicates transfected DNA construct, white text indicates *in situ* probe sequence. A representative set of signals is shown for imaging results along with a plot of normalized signal intensity against distance from the cell soma in microns. Blue lines represent control signal for endogenous transcripts following EGFP transfection, red lines represent level of endogenous signal following ID-EGFP transfection. Each transfected ID-EGFP experiment (red boxes) is then binned to find average intensity values across 8μm distances and subtracted from binned EGFP signal to generate mean and SEM values to distances of 48μm. These values are plotted as Δf/f against distance in microns. Blue lines represent level of endogenous signal following EGFP transfection. Scale bars = 20μm.

Transfection of CAMK2Bi3ID1 or FMR1i1ID1 disrupts the intronic *in situ* pattern of their endogenous intron-retaining transcripts, namely CAMK2Bi3 and FMR1i1. While transfection of FMR1i1ID1 disrupts the intronic *in situ* pattern of FMR1i1 transcripts, dendritic targeting of CAMK2Bi3 transcripts is unaffected. This shows that the FMR1i1ID1 sequence specifically targets the endogenous *Fmr1* i1 containing sequence but not the endogenous *Camk2b* transcript containing the i3 sequence.

Conversely, transfection with CAMK2Bi3ID1 disrupts the dendritic targeting of not only CAMK2Bi3 transcripts, but FMR1i1 transcripts as well (Figure 4.7). This CAMK2Bi3ID1 disruption of endogenous FMR1i1 transcript targeting is of equal

or greater magnitude than transfection with FRM1i1ID1. Cross-competition of the CAMK2Bi3ID1 element with endogenous *Camk2b* and *Fmr1* intron-retaining transcripts indicates that a shared targeting mechanism is associated with those transcripts containing the CAMK2Bi3 ID element.

4.2.8 Genome-wide characterization of ID elements shows broad distribution in the rat genome

In an attempt to characterize the ID element landscape outside of the 33 genes of interest, we constructed a catalog of ID elements over the entire rat genome using a BLAST-based approach. We used the 74-nt BC1 RNA 5′ hairpin targeting motif as a query sequence for nucleotide BLAST and chose an e-value cutoff of $1x10^{-13}$ – this cutoff roughly corresponds to a p-value < 0.001 using a Bonferonni correction for multiple tests set to the approximate size of the rat genome, 2.5 gigabases. In all, we found 146,785 distinct ID element loci evenly distributed on the Watson and Crick strands (Table 4.11) – fewer overall than the approximate number of loci reported by RepeatMasker, 161,321, which is a result of our more stringent criteria in requiring high sequence similarity to BC1 in the 74-nt hairpin region, though in several cases, we found that RepeatMasker fails to annotate ID element loci overlapping other repetitive sequence. Thus, our results represent both a more specific and more sensitive catalog of ID element hairpins.

These 146,785 ID elements are comprised of 62,101 unique ID element sequences, indicating a high degree of degeneracy from the canonical BC1 RNA sequence, which may be due to selective pressure (positive or negative) or drift. Using the criteria defined in Section 4.2.4, we assessed the ID elements for targeting competency and found that 78 percent of the elements satisfied the criteria for being capable of functioning as dendrite-targeting elements. By comparison, using the same parameters

Table 4.11: ID elements in the rat genome

	Num.	Targeting competent Num.	Percent
Total rat genomic ID elements	146,785	115,004	78.4
Watson (+) strand elements	73,527	57,586	78.3
Crick (−) strand elements	73,258	57,418	78.4
Intergenic elements	106,412	82,839	77.8
Genic elements	40,373	32,165	79.7
Antisense elements	23,183	18,475	79.7
Sense elements	17,190	13,690	79.6
Unique rat ID sequences	62,101	37,960	61.1

and the mouse-specific BC1 RNA, we identified only 682 ID elements, consistent with previous estimates [76]. These results confirm that the majority of the ID elements in the rat genome arose after the mouse-rat lineage split.

We annotated each locus according to RefSeq gene annotations and found that 27.5 percent of genomic ID elements are found in 8,784 genes, indicating slight enrichment when compared to the proportion of the rat genome annotated by RefSeq as genic (< 25.2 percent). A slightly but significantly greater proportion of the genic elements are targeting competent compared to the genome-wide frequency (79.6 percent, p $=$ 3.6 x 10^{-5} by Binomial Test). Among gene-overlapping ID elements, there is a bias toward antisense placement with respect to the gene orientation, 57.4 percent, that is similar to values we calculated for B2 elements. There is no difference in rate of target competency between sense and antisense elements (Table 4.11).

Focusing our attention on putatively target-competent elements, ID element content per gene is correlated with the length of the gene for both sense ($r^2 = 0.42$, p $<$2 x 10^{-16}) and antisense ($r^2 = 0.32$, p < 2 x 10^{-16}) elements, though the proportion

of sense to antisense elements varies greatly per gene (Figure 4.8). In fact, there is almost no correlation between gene length and sense-strand surplus ($r^2 = 0.002$), which we define as the number of target-competent sense elements in a gene minus the number of target-competent antisense elements. The average sense-strand surplus over all genes containing ID elements is -0.57, which reflects the overall antisense bias among genic ID elements. (Figure 4.9)
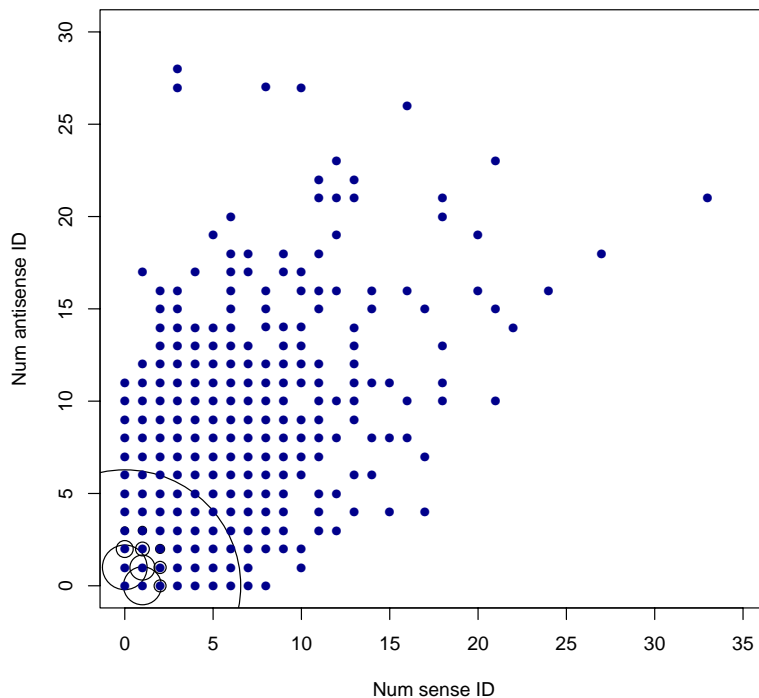
Figure 4.8: Number of sense-versus antisense-direction ID elements per gene. A concentric circle around a point defines an area that is proportional to the number of genes containing the given number of sense and antisense ID elements.

The vast majority of the gene-overlapping ID elements are intronic; however, 118 target-competent elements occur in 3′ UTR exons while eight occur in 5′ UTR exons. The role of these exonic elements is unclear, and given that they have not previously been characterized as functional dendritic targeting sequences suggests
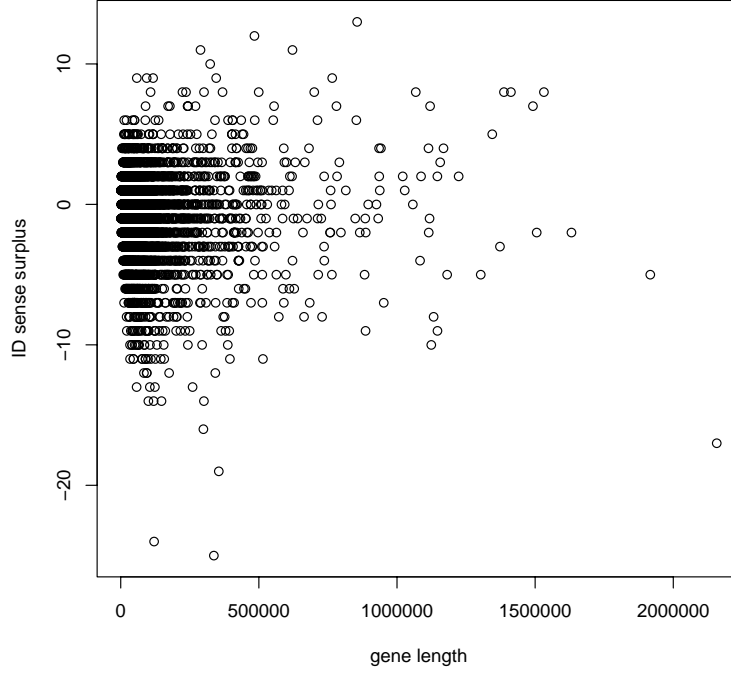
Figure 4.9: Surplus of sense-direction ID elements compared to gene length in nucleotides. Each point represents one gene.

that the placement of ID elements in introns is important in their efficacy as localization elements. Interestingly, there are three examples of ID elements spanning annotated splice boundaries – chrUn:25,471,736-25,471,806 (NM_001008882 exon 1 and intron 1), chr4:134,544,564-134,544,637 (NM_001047956 exon 1 and intron 1), and chr1:158,440,020-158,440,093 (NM_017255 exon 3, intron 3, and exon 4). Each of these elements is novel in the rat lineage, implying that in these instances, evolutionarily recent ID-element insertion has altered splicing patterns, though whether there are functional consequences of these changes is unknown. The fact that minor variants of the canonical ID element sequence can harbor both splice donor and acceptor sites may be significant in understanding their regulation as part of retained intronic sequence.

4.2.9 Neuronal function is associated with ID element-enriched genes

Given the wide distribution of gene-overlapping ID elements in the rat genome, we performed a Gene Ontology (GO) enrichment analysis to determine whether particular functions were associated with these genes. Since there is a length effect on the number of ID elements that appear in a gene, which may bias the results of a GO enrichment test, we focused on the subset of genes containing a surplus of target-competent sense-direction ID elements, which has no correlation with length (Figure 4.9). Since there is an overall antisense bias for ID element orientation, genes with excess sense-direction ID elements may reflect selection for target-competent ID elements in particular genes (or perhaps lack of selection against target-competent ID elements). A total of 2762 RefSeq-annotated genes contain a surplus of sense targeting-competent ID elements; performing a GO analysis on these genes yields highly significant enrichment in GO terms related to neuronal activity (Table 4.12), including several terms consistent with genes expected to be present in dendrites, such as "synapse," "plasma membrane," and "gated channel activity." These results indicate that neuronal function is associated with potentially functional ID elements on a genome-wide level.

We also looked for GO terms enriched in genes containing ID elements with sequencing support. Target-competent ID elements are spanned by uniquely-aligning mate pairs in 1617 genes. Significantly enriched GO terms include neuronal functions, as expected since the sequencing experiments were performed on neuron transcriptomes; interestingly, the most significant of these terms are specifically related to binding and localization (Table 4.13), which are consistent with genes involved in active transcript targetting.

Table 4.12: Most significantly enriched Gene Ontology terms in genes with a sense-strand-ID element surplus

Cat.[a]	Term	Description	p value	Enrich.[b]	Bonferroni[c]
MF	GO:0005488	binding	1.16E-22	1.17	3.22E-19
CC	GO:0045202	synapse	1.26E-19	3.17	1.03E-16
MF	GO:0005515	protein binding	3.57E-18	1.30	9.91E-15
CC	GO:0005737	cytoplasm	1.72E-17	1.34	1.41E-14
CC	GO:0044424	intracellular part	1.60E-16	1.21	9.07E-14
BP	GO:0019226	transmission of nerve impulse	4.19E-16	2.30	2.24E-12
CC	GO:0005886	plasma membrane	8.46E-16	1.55	7.26E-13
BP	GO:0007268	synaptic transmission	1.56E-14	2.33	7.88E-11
CC	GO:0000267	cell fraction	1.90E-14	1.76	1.55E-11
CC	GO:0044444	cytoplasmic part	3.34E-14	1.37	2.73E-11
CC	GO:0044459	plasma membrane part	4.90E-14	1.59	4.00E-11
BP	GO:0051234	establishment of localization	8.49E-14	1.43	4.28E-10
MF	GO:0022836	gated channel activity	1.30E-13	2.64	3.62E-10
MF	GO:0005216	ion channel activity	1.30E-13	2.43	3.62E-10
BP	GO:0006810	transport	1.47E-13	1.43	7.38E-10
MF	GO:0022838	substrate specific channel activity	2.62E-13	2.38	7.29E-10
CC	GO:0044456	synapse part	2.87E-13	3.42	2.34E-10
MF	GO:0015267	channel activity	5.95E-13	2.33	1.65E-09
MF	GO:0022803	passive transmembr. transporter act.	5.95E-13	2.33	1.65E-09
CC	GO:0005624	membrane fraction	8.55E-13	1.83	6.99E-10
BP	GO:0051179	localization	8.93E-13	1.37	4.50E-09
MF	GO:0005215	transporter activity	1.02E-12	1.63	2.84E-09
BP	GO:0007267	cell-cell signaling	5.04E-12	1.85	2.54E-08
MF	GO:0022892	substrate-specific transporter activity	5.11E-12	1.71	1.42E-08
MF	GO:0022857	transmembrane transporter activity	2.86E-11	1.72	7.96E-08

[a]Gene Ontology category: BP = biological process, CC = cellular compartment, MF = molecular function. [b]Fold enrichment over expectation. [c]p value for enrichment after Bonferroni multiple test correction.

Table 4.13: Most significantly enriched Gene Ontology terms in genes containing ID elements supported by sequencing reads

Cat.[a]	Term	Description	p value	Enrich.[b]	Bonferroni[c]
BP	GO:0051179	localization	1.44E-19	1.88	7.23E-16
MF	GO:0005488	binding	4.82E-18	1.26	1.34E-14
MF	GO:0005515	protein binding	6.66E-18	1.55	1.85E-14
BP	GO:0051234	establishment of localization	6.10E-16	1.86	2.79E-12
BP	GO:0006810	transport	8.41E-16	1.88	4.47E-12
BP	GO:0016043	cell. component organiz., biogenesis	6.29E-15	1.85	3.19E-11
BP	GO:0051649	establishment of cellular localization	1.24E-14	2.64	6.20E-11
BP	GO:0051641	cellular localization	2.31E-14	2.61	1.16E-10
CC	GO:0045202	synapse	3.48E-14	4.69	2.85E-11
CC	GO:0005737	cytoplasm	3.72E-14	1.56	3.04E-11
CC	GO:0044444	cytoplasmic part	3.21E-11	1.60	2.62E-08
CC	GO:0044424	intracellular part	3.41E-11	1.30	2.79E-08
BP	GO:0045045	secretory pathway	4.16E-11	3.89	2.09E-07
MF	GO:0000166	nucleotide binding	1.67E-10	1.85	4.64E-07
MF	GO:0017076	purine nucleotide binding	1.70E-10	1.93	4.72E-07
BP	GO:0032940	secretion by cell	2.02E-10	3.49	1.02E-06
CC	GO:0005622	intracellular	3.68E-10	1.26	3.01E-07
BP	GO:0046903	secretion	6.08E-10	3.04	3.06E-06
BP	GO:0019226	transmission of nerve impulse	1.50E-09	2.85	7.54E-06
BP	GO:0016192	vesicle-mediated transport	1.50E-09	2.85	7.54E-06
MF	GO:0032555	purine ribonucleotide binding	2.09E-09	1.89	5.80E-06
MF	GO:0032553	ribonucleotide binding	2.09E-09	1.89	5.80E-06
BP	GO:0033036	macromolecule localization	7.35E-09	2.36	3.70E-05
CC	GO:0043005	neuron projection	1.07E-08	3.66	8.76E-06

[a]Gene Ontology category: BP = biological process, CC = cellular compartment, MF = molecular function. [b]Fold enrichment over expectation. [c]p value for enrichment after Bonferroni multiple test correction.

4.3 Conversion of Alu sequence into Camk2a-style localization elements

In our search for localization factors in the retained introns of our 33 dendritically localized genes, we looked for sequence motifs similar to the two previously characterized dendrite-localization elements – the *Camk2a* element and the *Map2* element, both found in their respective gene's 3′ UTR. The minimal *Map2* element is large (∼640 nts) and is predicted to form an extensive secondary structure [16]. We were not able to find this element in any of the intronic sequences, though it is still possible that specific submotifs may be present.

In contrast, the minimal *Camk2a* targeting element is small (∼50 nts). Mori et al [15] characterized this element using deletion constructs as the necessary and sufficient element for directing *Camk2a* mRNAs to the dendritic compartment. Based on manual assessment of nucleotide sequence similarity, they were also able to identify a corresponding element in neurogranin that also conferred dendritic targeting that had ∼30 nt similarity with the *Camk2a* element.

4.3.1 Camk2a localization motif forms a local hairpin structure

Although Mori et al did not report any form of structural similarity among the *Camk2a* and neurogranin localization elements, we hypothesized that secondary structure may contribute to their functional specificity. We performed a consensus secondary structure prediction using RNAalifold [77] on the rat and mouse *Camk2a* localization elements as well as the rat neurogranin localization element, and found that the sequences folded into a stable hairpin structure spanning 19 nucleotides of the motif (Figure 4.10). A core region of three base pairs is perfectly conserved between the three elements.

165

```
                                    .............(((((((((......)))))))))..
mouse_minimal_camkii/1-37    CCT--------CGTTCGCAGAGA-TCCGCTCTTTGTC- 28
rat_minimal_camkii/1-37      CTT-------GC-TTCGCAGAGA-TCCATTCTTTGTC- 28
rat_neurogranin/1-38         ct-ccccgccgcgt-cccagagattcc-ctctcccagg 35
                                    .........10........20........30.......
```
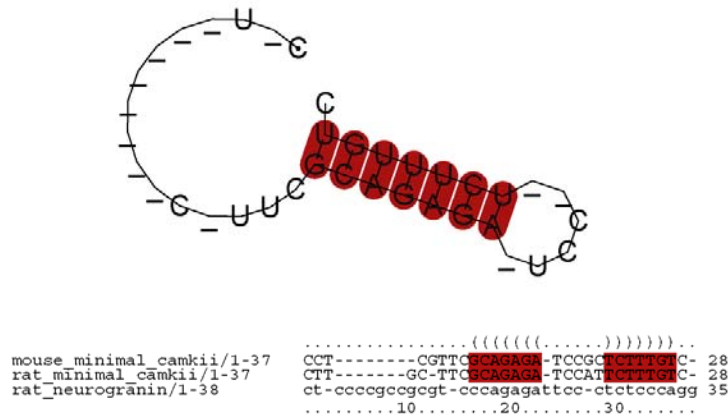
Figure 4.10: Predicted structure of the *Camk2a* localization element. Consensus secondary structure was determined using RNAalifold on rat and mouse *Camk2a* element and rat neurogranin element. Red shaded base pairs indicate high-probability interactions.

We investigated the stability of the hairpin in the context of the surrounding transcript. In the context of the full UTR, none of the base pairs has greater than 0.01 probability of forming, as determined using RNAplfold [78], which measures the local stability of structures in context. However, the hairpin does have high self containment (SC = 0.96), indicating that in other contexts, the hairpin would have a high probability of robustly folding. Given that transcript localization appears to be modulated by additional motifs in the *Camk2a* 3′ UTR [15], these results are not inconsistent, and may indicate that formation of the hairpin is regulated – i.e., by sequestering the sequence to allow it to fold properly – depending on whether localization is desired.

4.3.2 Genome-wide scan reveals a large number of similar motifs strongly associated with Alu elements

Using this structural specification along with the nucleotide similarity between the localization elements, we constructed a search protocol to scan nucleotide sequences

for candidate localization elements. We extracted all 19-nt segments with an N_4-AGA-N_5-TCT-N_4 sequence pattern (where N_k is any nucleotide sequence of length k) with a predicted hairpin secondary structure. We additionally filtered out any sequences with less than 11/19 nucleotide similarity to the rat *Camk2a* element, corresponding to the sequence distance separating the neurogranin and *Camk2a* elements.

When we ran the search over the 33 dendritic genes, we found 282 instances of candidate motifs spanning 29 of the genes, or an incidence rate of five per 100,000 nucleotides. Genome wide we found 138,967 candidates, corresponding to a roughly equivalent incidence rate.

Upon closer examination of the genome-wide results, we found an enrichment in candidate elements associated with repetitive sequence – 54 percent of the elements overlapped a RepeatMasker-annotated repetitive element, compared to an overall rat repeat genomic frequency of about 41 percent. Given that repetitive elements, consisting of different classes of transposons plus low complexity sequence, may have a higher tendency to form secondary structures, this result is not necessarily surprising. However, when we broke down the repeats into their respective families, we found a marked enrichment in *Camk2a* elements overlapping SINE/Alu-derived sequence – nearly a five-fold enrichment over genomic frequency (Table 4.14), which is an extremely significant deviation ($p < 10^{-323}$ by a binomial test). These elements include various B1 and proto-B1 (PB1 and PB1D) classes of retrotransposons, all of which show enrichment over genomic frequencies (data not shown).

To verify that we were not biasing the results due to the secondary structure of the *Camk2a* element, we constructed a "decoy" element with the same secondary structure but shuffled sequence, and performed an identical genome scan to the original *Camk2a* scan, and found that the *Camk2a* candidates were enriched in SINE/Alu sequence over the decoy frequencies as well (5.8 fold) (Table 4.14). We additionally

Table 4.14: Rat repeat element families most frequently overlapping candidate *Camk2a*-style elements

Repeat Family	Genomic Prop.[a]	DECOY MOTIF			CAMK2A MOTIF			
		Count[b]	Prop[c]	Enrich[d]	Count	Prop	Enrich	Enrich (d)[e]
LINE/L1	0.202	13018	0.232	1.150	32944	0.237	1.174	1.021
LTR/MaLR	0.036	4365	0.078	2.145	5774	0.042	1.145	0.534
LTR/ERVK	0.030	1664	0.030	1.006	4682	0.034	1.142	1.135
Simple repeat	0.022	943	0.017	0.775	3465	0.025	1.149	1.483
SINE/B4	0.020	1377	0.025	1.233	5973	0.043	2.159	1.750
SINE/B2	0.020	2433	0.043	2.222	1952	0.014	0.719	0.324
SINE/Alu	**0.015**	**698**	**0.012**	**0.814**	**9981**	**0.072**	**4.699**	**5.770**
Unknown	0.010	447	0.008	0.825	976	0.007	0.727	0.881
LTR/ERV1	0.009	406	0.007	0.797	1603	0.012	1.270	1.593
LTR/ERVL	0.008	518	0.009	1.131	2000	0.014	1.762	1.558

[a]Proportion of rat genome annotated with the repeat family by RepeatMasker. [b]Number of motifs overlapping the repeat family. [c]Proportion of the motifs that overlap the repeat family. [d]Frequency of repeat overlaps for the motif divided by the genomic frequency. [e]Frequency of the repeat overlaps for the *Camk2a* motif divided by the frequency for the decoy.

generated 1000 such decoy elements using an inverse folding algorithm (RNAinverse [68]) and compared the frequency with which the *Camk2a* motif overlaps Alu elements to the frequencies for the decoys. Decoys overlap an average of 127.45 Alu elements on chromosome 1, corresponding to an average Alu-overlap incidence rate of 0.025 for decoy motifs. In contrast, 1070 of the 12,701 total *Camk2a* matches on chromosome 1 overlap an Alu element, yielding a significantly higher incidence rate of 0.084 (p $\ll$ 1 x 10^{-16} by a Chi-squared goodness of fit test). Thus, the specific *Camk2a* motif and not just hairpin motifs in general has significant similarity to Alu elements.

We performed a similar analysis on the mouse genome and found 141,907 candidate elements that again showed a marked preference for overlapping Alu elements (Table 4.15), at a rate 3.7 times the expected frequency. Additionally, we found that B4 SINE elements as well as ERVL LTR elements occurred at more than twice the

Table 4.15: Mouse repeat element families most frequently overlapping candidate *Camk2a*-style elements

| Repeat Family | Genomic Prop.[a] | CAMK2A MOTIF | | |
		Count[b]	Prop[c]	Enrich[d]
LINE/L1	0.193	24945	0.176	0.910
LTR/MaLR	0.043	6368	0.045	1.037
LTR/ERVK	0.042	7146	0.050	1.198
SINE/Alu	**0.025**	**13044**	**0.092**	**3.738**
Simple repeat	0.024	4705	0.033	1.382
SINE/B2	0.022	1945	0.014	0.626
SINE/B4	**0.022**	**6524**	**0.046**	**2.103**
LTR/ERVL	**0.011**	**3750**	**0.026**	**2.343**
LTR/ERV1	0.010	1597	0.011	1.171
Low complexity	0.007	1167	0.008	1.108

[a]Proportion of mouse genome annotated with the repeat family by RepeatMasker. [b]Number of motifs overlapping the repeat family. [c]Proportion of the motifs that overlap the repeat family. [d]Frequency of repeat overlaps for the motif divided by the genomic frequency.

expected frequency.

4.3.3 Genes with candidate Camk2a elements have neuronal function

We performed a Gene Ontology enrichment analysis to determine if there was any functional coherence in the genes containing candidate *Camk2a* motifs. Due to the possible confounding effect of gene length – longer genes will be more likely to contain matching motifs by chance, which would bias an enrichment analysis toward terms associated with longer genes – it was necessary to construct a background set of genes with similar length characteristics to the test set. 7779 RefSeq-annotated rat genes contain *Camk2a* candidate motifs. From these, we partitioned a subset (~ 10 percent) with the highest degree of sequence similarity to the actual *Camk2a* localization element, and tested enrichment in GO functional categories in this subset compared to a background consisting of the entire gene list. Any overall length differences between

Table 4.16: Most significantly enriched Gene Ontology terms in genes containing motifs similar to the *Camk2a* localization element

Cat.[a]	Term	Description	p value	Enrich.[b]	Bonferroni[c]
CC	GO:0016020	membrane	3.67E-09	1.26	2.41E-06
CC	GO:0045202	synapse	7.04E-09	2.44	4.63E-06
MF	GO:0060089	molecular transducer activity	1.77E-07	1.55	3.56E-04
MF	GO:0004871	signal transducer activity	1.77E-07	1.55	3.56E-04
CC	GO:0044456	synapse part	2.56E-07	2.72	1.68E-04
BP	GO:0003001	gener. of a signal / cell-cell signaling	6.78E-07	2.93	2.64E-03
BP	GO:0022610	biological adhesion	6.91E-07	2.00	2.69E-03
BP	GO:0007155	cell adhesion	6.91E-07	2.00	2.69E-03
BP	GO:0042476	odontogenesis	1.09E-06	6.94	4.23E-03
MF	GO:0004872	receptor activity	1.56E-06	1.57	3.12E-03
BP	GO:0007269	neurotransmitter secretion	2.98E-06	3.16	1.15E-02
BP	GO:0007154	cell communication	3.94E-06	1.30	1.52E-02
BP	GO:0007156	homophilic cell adhesion	7.08E-06	3.11	2.72E-02
CC	GO:0045211	postsynaptic membrane	1.42E-05	2.71	9.28E-03
BP	GO:0007399	nervous system development	1.50E-05	1.71	5.68E-02

[a]Gene Ontology category: BP = biological process, CC = cellular compartment, MF = molecular function. [b]Fold enrichment over expectation. [c]p value for enrichment after Bonferroni multiple test correction.

the genes in the foreground and background sets should not be due to random effects, since there is no *a priori* reason to expect that motifs with more sequence identity to the *Camk2a* element would be found more often in longer genes than equal-length motifs with less sequence identity. Significantly enriched GO terms are show in Table 4.16 and include several terms relevant to neuronal biology such as "membrane," "synapse," and "molecular transducer activity."

A similar GO analysis performed using genes containing the decoy motif produced no significantly enriched terms after multiple test correction.

4.4 Discussion and conclusions

In this chapter we presented examples of modular RNA functionality mediated by transposable elements. In the first section, we characterized the phenomenon of intron sequence retention among a large number of dendritically localized mRNA of primary rat hippocampal neurons; these introns provide a previously unreported context in which localization elements may reside. Many of these introns harbor ID elements, a class of SINE retrotransposon, that we showed confer dendritic targeting capacity for the host transcript. Genome-wide characterization of ID elements revealed a wide distribution across many transcripts, particularly neuron-function associated genes. In the second section, we showed that the RNA structural motif responsible for dendritic targeting of the *Camk2a* mRNA occurs throughout the rat genome and is significantly associated with sites of SINE Alu element insertion, and again appears to show a preference for arising in neuron-function genes. Both the ID element and the *Camk2a*/Alu element show the capacity for the functionalization of transposable elements in a way that affects fundamental neurobiology.

ID elements were previously implicated in brain-specific regulation [79]. Members of the Brosius lab created transgenic mice with various ID elements, as well as the 5′ ID domain of BC1, fused to the 3′ UTR of EGFP and found that these sequences were not sufficient for dendritic targeting *in vivo*. Additionally, ID elements occurring endogenously in the 3′ UTR of neuronally expressed genes were tested for dendritic localization and also found to be restricted to cell bodies [80]. In our experiments, ID elements along with some degree of flanking intronic sequences were fused to the 5′ end of EGFP and transfected into cultured neurons. The sequence context in which ID elements are presented in this system is, therefore, different from that of ID sequences arising from coding or UTR sequences. There is evidence that targeting

171

mechanisms can depend on intronic sequence. In *Drosophila*, correct localization of *oskar* mRNA to the posterior pole of a developing oocyte requires the presence of an intron [81], as the localization mechanism appears to be coupled to splicing.

An alternate explanation may be that the ID elements in rats have acquired novel functional roles restricted to this species in comparison to mice. ID elements have undergone great expansion in rats, with more than 145,000 instances of the 5′ targeting domain according to our analysis, while the mouse genome contains two orders of magnitude less (approximately 680 instances). These numbers are consistent with a previous survey of ID elements in rodents, which suggested a wide variety of genomic distributions from estimates of 200 (guinea pig) to hundreds of thousands (rat) [76]. This suggests a surprising finding that evolutionarily-novel element expansion may play a critical functional role in neuronal physiology. The acquisition of this functional role may be mediated by the novel processing of retained introns, which creates a different sequence context from the 3′ UTR sequences as well as a substrate for other specificity-determining RBP factors. Functionalization of retroelements has been suggested to provide a dynamic reservoir of rapid genome evolution [82, 83]. Here, we provide strong evidence for evolutionarily rapid functionalization of a mobile element.

The variety of distribution patterns in our intronic *in situ* hybridization results also highlights that there are multiple mechanisms for targeting of intron-retaining transcripts in dendrites. The fact that exogenous expression of any particular intronic ID element does not disrupt targeting of all intron-retaining transcripts shows that targeting of intron-retaining transcripts involves multiple targeting mechanisms. If a single mechanism were in place, transfection of any intronic ID element would block the targeting all endogenous intron-retaining transcripts containing an ID element. Our data indicate that at least three targeting mechanisms exist for intron-retaining transcripts in dendrites: one which is common to CAMK2Bi3iD1 and FMR1i1ID1,

one which is distinct for FMR1i1ID1, and one which is ID element independent. These mechanisms coupled with those already proposed for the targeting of non-intron containing mRNAs – i.e. those specific to *Map2* and *Camk2a* – indicate that minimally five distinct dendritic targeting mechanisms exist. The fact that neither these characterized localization elements nor ID elements appear in all transcripts hypothesized to be localized to the dendrites suggests that there are even more mechanisms in play.

The complexity of the dendritic mRNA targeting mechanisms further highlights the fact that localization of mRNAs within the dendrite is important for neuronal function. Indeed, when particular RNAs are present in dendrites their translation can cause cell death [84], whereas other RNAs are important for aspects of learning and memory [85, 86, 87, 88]. The evolutionary novelty of the ID elements within the rat genome also suggests that the variety of localization mechanisms may rapidly evolve and modulate species-specific characteristics of individual neuronal function. As these targeting mechanisms come to be understood, insight into how they may be regulated promises to provide important information with regard to maintaining dendrite viability and function.

ID elements constitute an example of wholesale cooption of the sequence of a transposable element, which, coupled with regulation in the form of intron retention and possibly other mechanisms, can rapidly attain functionality. In the case of the *Camk2a* localization motifs, our results suggest that transposable elements may also provide the raw material for "cooption with modification," such that a minimal number of nucleotide changes allow transposon-derived sequence to become functional RNA elements. This appears to have been the case in *Drosophila*: versions of the *gurken* mRNA localization signal are also found in G2, Jockey, and I factor transposable elements, whose targeting competency and specificity was verified by injecting labeled transposons into oocytes [89]. It is possible that the *Camk2a* element is itself

derived from a transposon sequence or shares common ancestry with the Alu master gene, which interestingly is believed to be the 7SL RNA [90, 91], the RNA component of the signal recognition particle, which is responsible for protein localization. The fact that many retrotransposons derive from ancestral RNAs [45], and thus may still contain a high degree of biologically-relevant structure, reinforces the idea that transposons constitute a supply of RNA building blocks from which novel function can arise.

Many questions remain. Given the frequency with which ID elements appear throughout the rat genome, regulation of intron retention must play a role in determining which of these are allowed to drive localization; however, the nature of this regulation is still unclear. In particular, many of the ID-containing introns are orders of magnitude longer than the coding exon, which would result in the export of an extremely long quasi-mature mRNA out of the nucleus if the entire intron sequence is retained; thus, perhaps only portions of the full intron are retained, possibly through the use of cryptic splice sites.

As we described above, ID elements are not found in large number in the mouse genome, implying that an alternate yet analogous mechanism could exist for mouse dendritic localization. Preliminary evidence from our lab suggests that intron retention is also present in mouse dendritic transcripts, so it remains to be seen if a corresponding localization element can be found among these introns. If such an element turns out to be a different characteristic mouse transposon, it would be an example of conserved (or analogous) function in the absence of conserved mechanism. Evidence of this pattern of functional analogy has already been found in the introns of humans and mice, where there exist characteristic overrepresented sequence motifs whose sequence is not conserved between the two species but whose pattern of occurrence in genes with specific functions appears to be correlated [92].

Finally, we note that neuronal genes tend to be longer overall than genes not associated with neuron function. Our set of 33 dendritic associated genes are on average twice the length of the average RefSeq-annotated gene, a phenomenon that is due in large part to the length and number of introns appearing in these genes. We have taken measures to prevent length confounds in our analysis, specifically with respect to gene enrichment, since longer sequences will tend to accumulate a greater number of transposon insertions. However, is it correct to assume that these long introns have no biological significance? Neuronal function is vital to the fitness of higher organisms, so we might expect a high degree of conservation among essential genes encoding the various neuron components – e.g. ion channels or vesicular proteins. And yet, the diversity of neuronal function is apparent among species as closely related as rats and mice, implying rapid rates of sequence change. But perhaps we can reconcile these two seemingly conflicting evolutionary forces with a more holistic view of functionally-significant sequence change that includes intronic sequence. The fact that neuronal gene introns are long may reflect some sort of selection for accumulating a reservoir of RNA elements that initially have a neutral effect, by virtue of being in a non-translated region of the gene, but then can rapidly become functional in a way that affects the expression pattern of the transcript rather than the protein code. Such a differential in the conservation rates of coding versus intronic sequences is in fact observed among presynaptic genes, whose exons are highly similar but whose introns are highly divergent between eight vertebrate species [93]. In this way, rather than rapid evolution of the "what" – i.e., amino acid sequence – there is rapid evolution of the "how," "where," and "when," as a way to effect phenotypic change.

4.5 MATERIALS AND METHODS

4.5.1 WET PROCEDURES

CULTURING CONDITIONS   Primary cultures from E18.5 rat embryos are plated at 100,000 cells per ml of Neurobasal medium and B27 (Invitrogen). Neurons are grown on 12mm round German Spiegelglas coverslips (Bellco) coated with poly-L lysine (Peptide Institute). mRNA amplification and cDNA labeling: Probes were generated from mRNA isolated from dendrites. Dendrites were harvested by mechanical isolation from primary rat hippocampal neurons (harvested day E18, cultured for 13 days). Approximately 150 dendrites were used as template material for aRNA amplification. Following three rounds of aRNA, labeled single stranded cDNA was generated by incorporation of amino-allyl labeled dUTP and conjugation with Cy3. Labeled material from dendrites was hybridized to our custom microarrays and screened for positives.

MICROARRAY SAMPLE PREPARATION   Fragments were amplified using forty rounds of PCR with an annealing temperature of 50°C. The template used was rat genomic DNA isolated from rat liver. 1μg of each of 96 PCR products were submitted to the University of Pennsylvania Microarray Facility for printing on Corning UltraGap slides. These samples were dried and resuspended in 10μl of Corning Spotting Buffer. 1nl of each sample was then denatured and printed in each spot on individual slides and cross-linked using ultra-violet light for immobilization.

MICROARRAY DETECTION   Slides were blocked (pre-hybridized) at 42°C for 3 hours in 1% bovine serum albumin (BSA), 1% sodium dodecyl sulphate (SDS), and 3X saline-sodium citrate (SSC). Hybridization was carried out in Corning slide chambers for sixteen hours at 42°C in a 25% formamide, 0.1% SDS, 4X SSC buffer with human

Cot-1 DNA, single stranded (SS) poly dA and poly dT DNA, yeast transfer RNA (tRNA) and T7-oligo dT primer as blocking agents. Slides were washed two times for five minutes at room temperature (RT) in 2X SSC, 0.1% SDS, two times for five minutes at 42°C in 0.2X SSC, 0.1% SDS, and two times for five minutes at RT in 0.2X SSC. Slides were scanned using an Axon Instruments GenePix 4200 series scanner provided by the University of Pennsylvania Microarray Facility, and analyzed with GenePix 6.0 software.

INTRONIC SEQUENCE SUBCLONING   It should be noted that the ID PCR products do not contain the sequence found in the microarray PCR product. ID sequences were found up to 600kb from the upstream splice site making it problematic to represent the entire intron or the regions of interest in a single PCR product (Figure 4.5b).

IN SITU HYBRIDIZATION AND IMAGING   Antisense digoxigenin or biotin-labeled probes were produced as runoff transcripts from plasmid DNAs that were digested at a site downstream of the region to be transcribed. Primary rat hippocampal neurons were fixed for 15 minutes in 4% paraformaldehyde, washed in PBS and permeabolized with 0.3% TritonX-100. Cells were prehybridized at 42°C with 50% formamide, 1X Denhardt's solution, 4X SSC, 10mM DTT, 0.1% CHAPS, 0.1% Tween-20, 500μg/ml yeast tRNA, 500μg/ml salmon sperm DNA. *In situ* hybridization was performed at 42°C with 10ng/μl (for EGFP probes) or 20ng/μl (for intron probes) probe in prehybridization buffer with additional 8% Dextran sulfate. Rabbit anti-MAP2 antibody was added to cells after probe hybridization followed by goat anti-rabbit antibody and streptavidin conjugated to Qdot molecules for imaging. The samples were visualized by confocal microscopy. The emission wavelengths for each fluorescent dye were selectively collected by specific spectral ranges of dyes with either slit width (Olympus fluoview 1000, 60x N.A.1.2 or 20x N.A.0.7) or meta detector (Zeiss 510

meta, 40x N.A 1.0). The collected images were minimally processed in Metamorph image analysis software and extracted information in regions of interest was transferred to Excel. The images were background subtracted and scaled 0 to 2000 in 12bit bit depth unless indicated in text.

4.5.2 Computational procedures

Software and implementation   All computation was performed using custom-written Python and R code run on quad-core Linux machines with 16GB of memory. RNA structure prediction was performed using Vienna RNAFold 1.7 [68]). Rat genome sequence (v 3.4 [30]), mouse genome sequence (Build 37), RefSeq gene annotations, and RepeatMasker annotations were obtained from the UCSC Genome Browser [94]. GO enrichment analysis was performed using the NIH DAVID server (http://david.abcc.ncifcrf.gov/).

BLAST sequence analysis   For the initial intron comparisons, pairwise NCBI BLAST [95] was run on each pair of intronic sequences represented on the microarray using an e-value cutoff of 1e-10, and results were clustered (single-linkage) based on overlapping gene coordinates. Individual clusters were annotated for presence of repetitive elements, including the ID element, using RepeatMasker [67]. To construct the ID element genome-wide catalog, BLAST was run querying the canonical RepBase ID element sequence against the entire rat genome sequence. Gene annotations are based on RefSeq gene annotations for gene feature boundaries and strand. Overlapping genes were considered ambiguous and not included for the purposes of annotating ID element strand and feature preference.

Alignment of Illumina sequencing reads to rat genes   Specific read coverage for our 33 genes of interest was performed using Bowtie [72] version 0.9.8 using

the default parameters on the rat genome. Only reads uniquely aligning to the gene loci were defined as retained, except in reads aligning *Stx1a*, where we corrected for the fact that the gene has two genomic copies. Paired-end reads were used to define high-confidence regions present in the transcriptome samples, while additional read coverage from unpaired single reads was used to augment the transcriptome maps to mitigate reduced sensitivity from the paired-end analysis on shorter features and lower-complexity sequences [72].

ALIGNMENT OF ILLUMINA SEQUENCING READS TO INTERGENIC REGIONS To determine whether read alignment was specific to the intronic regions, and not endemic to general non-coding regions, we analyzed read coverage in intergenic regions upstream and downstream of the 33 genes of interest such that the amount of repeat-masked sequence roughly corresponded to the amount of intronic sequence per gene. For each sequencing run, we compared the ratio of the number of aligning reads to the cumulative nucleotide length for the intergenic regions and the intronic regions using an Exact Binomial Test. Significantly higher ratios in the intronic regions indicate enrichment in read coverage compared to the presumed background level represented by the intergenic regions.

IGOR CALULATIONS Quantification paths are manually drawn tracing 3 dendrites of selected cells based on MAP2 immunostaining. Path origins were chosen at the somal end of the dendritic process. Generated paths are 11 pixels wide (4.4μm). The average signal intensity along the paths were computed for the *in situ* hybridization channel. These average intensities were normalized to the maximum signal along the path. The average of the normalized values was computed for each cell and then plotted against the distance from the path origin, using Graphpad Prism.

4.6 ACKNOWLEDGEMENTS

REFERENCES

[1] Gerbi SA (1986) The evolution of eukaryotic ribosomal DNA. Biosystems 19:247–58.

[2] Yokoyama T, Suzuki T (2008) Ribosomal RNAs are tolerant toward genetic insertions: evolutionary origin of the expansion segments. Nucleic Acids Res 36:3539–51.

[3] Zeng Y, Yi R, Cullen BR (2005) Recognition and cleavage of primary microrna precursors by the nuclear processing enzyme drosha. The EMBO Journal 24:138–148.

[4] Cooper DN, Krawczak M (1991) Mechanisms of insertional mutagenesis in human genes causing genetic disease. Hum Genet 87:409–15.

[5] Kazazian HHJ (1998) Mobile elements and disease. Curr Opin Genet Dev 8:343–50.

[6] Callinan PA, Batzer MA (2006) Retrotransposable elements and human disease. Genome Dyn 1:104–15.

[7] Myerowitz R, Costigan FC (1988) The major defect in Ashkenazi Jews with Tay-Sachs disease is an insertion in the gene for the alpha-chain of beta-hexosaminidase. J Biol Chem 263:18587–9.

[8] Karpuj MV, Garren H, Slunt H, Price DL, Gusella J, et al. (1999) Transglutaminase aggregates huntingtin into nonamyloidogenic polymers, and its enzymatic activity increases in Huntington's disease brain nuclei. Proc Natl Acad Sci U S A 96:7388–93.

[9] Lesort M, Chun W, Johnson GV, Ferrante RJ (1999) Tissue transglutaminase is increased in Huntington's disease brain. J Neurochem 73:2018–27.

[10] Lewinski MK, Bushman FD (2005) Retroviral DNA integration–mechanism and consequences. Adv Genet 55:147–81.

[11] Vagner S, Galy B, Pyronnet S (2001) Irresistible IRES. Attracting the translation machinery to internal ribosome entry sites. EMBO Rep 2:893–8.

[12] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–54.

[13] Macdonald PM, Struhl G (1988) cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. Nature 336:595–8.

[14] Chabanon H, Mickleburgh I, Hesketh J (2004) Zipcodes and postage stamps: mRNA localisation signals and their trans-acting binding proteins. Brief Funct Genomic Proteomic 3:240–56.

[15] Mori Y, Imaizumi K, Katayama T, Yoneda T, Tohyama M (2000) Two cis-acting elements in the 3' untranslated region of alpha-CaMKII regulate its dendritic targeting. Nat Neurosci 3:1079–84.

[16] Blichenberg A, Schwanke B, Rehbein M, Garner CC, Richter D, et al. (1999) Identification of a cis-acting dendritic targeting element in MAP2 mRNAs. J Neurosci 19:8818–29.

181

[17] Miyashiro KY, Bell TJ, Sul JY, Eberwine J (2009) Subcellular neuropharmacology: the importance of intracellular targeting. Trends Pharmacol Sci 30:203–11.

[18] Eberwine J, Belt B, Kacharmina JE, Miyashiro K (2002) Analysis of subcellularly localized mRNAs using in situ hybridization, mRNA amplification, and expression profiling. Neurochem Res 27:1065–77.

[19] Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297:2053–6.

[20] Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16:1616–26.

[21] Washida H, Kaneko S, Crofts N, Sugino A, Wang C, et al. (2009) Identification of cis-localization elements that target glutelin RNAs to a specific subdomain of the cortical endoplasmic reticulum in rice endosperm cells. Plant Cell Physiol NIL:NIL.

[22] Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–99.

[23] Shabalina SA, Ogurtsov AY, Spiridonov NA (2006) A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 34:2428–37.

[24] Moore MJ, Silver PA (2008) Global analysis of mRNA splicing. RNA 14:197–203.

[25] Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–4.

[26] Bell TJ, Miyashiro KY, Sul JY, McCullough R, Buckley PT, et al. (2008) Cytoplasmic BK(Ca) channel intron-containing mRNAs contribute to the intrinsic excitability of hippocampal neurons. Proc Natl Acad Sci U S A 105:1901–6.

[27] Konig H, Matter N, Bader R, Thiele W, Muller F (2007) Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. Cell 131:718–29.

[28] Morgan WF, Corcoran J, Hartmann A, Kaplan MI, Limoli CL, et al. (1998) DNA double-strand breaks, chromosomal rearrangements, and genomic instability. Mutat Res 404:125–8.

[29] Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314.

[30] Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521.

[31] The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

[32] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297:1003–7.

[33] Lynch M (2002) Genomics. Gene duplication and evolution. Science 297:945–7.

[34] Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363–7.

[35] Svensson O, Arvestad L, Lagergren J (2006) Genome-wide survey for biologically functional pseudogenes. PLoS Comput Biol 2:e46.

[36] Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260:91–5.

[37] Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 111:433–44.

[38] Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35:41–8.

[39] Bomar JM, Benke PJ, Slattery EL, Puttagunta R, Taylor LP, et al. (2003) Mutations in a novel gene encoding a CRAL-TRIO domain cause human Cayman ataxia and ataxia/dystonia in the jittery mouse. Nat Genet 35:264–9.

[40] Santangelo AM, de Souza FSJ, Franchini LF, Bumaschny VF, Low MJ, et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3:1813–26.

[41] Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90.

[42] Smalheiser NR, Torvik VI (2005) Mammalian microRNAs derived from genomic repeats. Trends Genet 21:322–6.

[43] Kuryshev VY, Skryabin BV, Kremerskothen J, Jurka J, Brosius J (2001) Birth of a gene: locus of neuronal BC200 snmRNA in three prosimians and human BC200 pseudogenes as archives of change in the Anthropoidea lineage. J Mol Biol 309:1049–66.

[44] McDonald JF (1995) Transposable elements: possible catalysts of organismic evolution. Trends in Ecology and Evolution 10:123 – 126.

[45] Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–34.

[46] Aakalu G, Smith WB, Nguyen N, Jiang C, Schuman EM (2001) Dynamic visualization of local protein synthesis in hippocampal neurons. Neuron 30:489–502.

[47] Bassell GJ, Kelic S (2004) Binding proteins for mRNA localization and local translation, and their dysfunction in genetic neurological disease. Curr Opin Neurobiol 14:574–81.

[48] Crino PB, Eberwine J (1996) Molecular characterization of the dendritic growth cone: regulated mRNA transport and local protein synthesis. Neuron 17:1173–87.

[49] Eberwine J, Miyashiro K, Kacharmina JE, Job C (2001) Local translation of classes of mRNAs that are targeted to neuronal dendrites. Proc Natl Acad Sci U S A 98:7080–5.

[50] Huang YS, Carson JH, Barbarese E, Richter JD (2003) Facilitation of dendritic mRNA transport by CPEB. Genes Dev 17:638–53.

[51] Job C, Eberwine J (2001) Identification of sites for exponential translation in living dendrites. Proc Natl Acad Sci U S A 98:13037–42.

[52] Kindler S, Wang H, Richter D, Tiedge H (2005) RNA transport and local control of translation. Annu Rev Cell Dev Biol 21:223–45.

[53] Miyashiro K, Dichter M, Eberwine J (1994) On the nature and differential distribution of mRNAs in hippocampal neurites: implications for neuronal functioning. Proc Natl Acad Sci U S A 91:10800–4.

[54] Steward O, Worley P (2001) Localization of mRNAs at synaptic sites on dendrites. Results Probl Cell Differ 34:1–26.

[55] Bramham CR, Wells DG (2007) Dendritic mRNA: transport, translation and function. Nat Rev Neurosci 8:776–89.

[56] Glisovic T, Bachorik JL, Yong J, Dreyfuss G (2008) RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 582:1977–86.

[57] Hastings ML, Krainer AR (2001) Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol 13:302–9.

[58] Huang Y, Steitz JA (2005) SRprises along a messenger's journey. Mol Cell 17:613–5.

[59] Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418:236–43.

[60] Yue C, Ponzio TA, Fields RL, Gainer H (2008) Oxytocin and vasopressin gene expression and RNA splicing patterns in the rat supraoptic nucleus. Physiol Genomics 35:231–42.

[61] Denis MM, Tolley ND, Bunting M, Schwertz H, Jiang H, et al. (2005) Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets. Cell 122:379–91.

[62] Wang Z, Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14:802–13.

[63] Glanzer J, Miyashiro KY, Sul JY, Barrett L, Belt B, et al. (2005) RNA splicing capability of live neuronal dendrites. Proc Natl Acad Sci U S A 102:16859–64.

[64] Dirks RW, Molenaar C, Tanke HJ (2001) Methods for visualizing RNA processing and transport pathways in living cells. Histochem Cell Biol 115:3–11.

[65] Garcia-Vitoria M, Garcia-Corchon C, Rodriguez JA, Garcia-Amigot F, Burrell MA (2000) Expression of neuronal nitric oxide synthase in several cell types of the rat gastric epithelium. J Histochem Cytochem 48:1111–20.

[66] Rhee WJ, Santangelo PJ, Jo H, Bao G (2008) Target accessibility and signal specificity in live-cell detection of BMP-4 mRNA using molecular beacons. Nucleic Acids Res 36:e30.

[67] Smit AFA, Hubley R, Green P. Repeatmasker Web Server. Available: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.

[68] Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–31.

[69] Kim J, Martignetti JA, Shen MR, Brosius J, Deininger P (1994) Rodent BC1 RNA gene as a master gene for ID element amplification. Proc Natl Acad Sci U S A 91:3607–11.

[70] Muslimov IA, Santi E, Homel P, Perini S, Higgins D, et al. (1997) RNA transport in dendrites: a cis-acting targeting element is contained within neuronal BC1 RNA. J Neurosci 17:4722–33.

[71] Phillips J, Eberwine JH (1996) Antisense RNA Amplification: A Linear Amplification Method for Analyzing the mRNA Population from Single Living Cells. Methods 10:283–8.

[72] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

[73] Muslimov IA, Iacoangeli A, Brosius J, Tiedge H (2006) Spatial codes in dendritic BC1 RNA. J Cell Biol 175:427–39.

[74] Lee MT, Kim J (2008) Self containment, a property of modular RNA structures, distinguishes microRNAs. PLoS Comput Biol 4:e1000150.

[75] Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–7.

[76] Kass DH, Kim J, Deininger PL (1996) Sporadic amplification of ID elements in rodents. J Mol Evol 42:7–14.

[77] Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474.

[78] Bernhart SH, Hofacker IL, Stadler PF (2006) Local RNA base pairing probabilities in large sequences. Bioinformatics 22:614–5.

[79] Milner RJ, Bloom FE, Lai C, Lerner RA, Sutcliffe JG (1984) Brain-specific genes have identifier sequences in their introns. Proc Natl Acad Sci U S A 81:713–7.

[80] Khanam T, Raabe CA, Kiefmann M, Handel S, Skryabin BV, et al. (2007) Can ID repetitive elements serve as cis-acting dendritic targeting elements? An in vivo study. PLoS One 2:e961.

[81] Hachet O, Ephrussi A (2004) Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. Nature 428:959–63.

[82] Kazazian HHJ (2004) Mobile elements: drivers of genome evolution. Science 303:1626–32.

[83] Volff JN, Brosius J (2007) Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. Genome Dyn 3:175–90.

[84] Barrett LE, Sul JY, Takano H, Bockstaele EJV, Haydon PG, et al. (2006) Region-directed phototransfection reveals the functional significance of a dendritically synthesized transcription factor. Nat Methods 3:455–60.

[85] Grooms SY, Noh KM, Regis R, Bassell GJ, Bryan MK, et al. (2006) Activity bidirectionally regulates AMPA receptor mRNA abundance in dendrites of hippocampal neurons. J Neurosci 26:8339–51.

[86] Ju W, Morishita W, Tsui J, Gaietta G, Deerinck TJ, et al. (2004) Activity-dependent regulation of dendritic synthesis and trafficking of AMPA receptors. Nat Neurosci 7:244–53.

[87] Kacharmina JE, Job C, Crino P, Eberwine J (2000) Stimulation of glutamate receptor protein synthesis and membrane insertion within isolated neuronal dendrites. Proc Natl Acad Sci U S A 97:11545–50.

[88] Kang H, Jia LZ, Suh KY, Tang L, Schuman EM (1996) Determinants of BDNF-induced hippocampal synaptic plasticity: role of the Trk B receptor and the kinetics of neurotrophin delivery. Learn Mem 3:188–96.

[89] Hamilton RS, Hartswood E, Vendra G, Jones C, Bor VVD, et al. (2009) A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in Drosophila retrotransposons. RNA 15:200–7.

[90] Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312:171–2.

[91] Labuda D, Sinnett D, Richer C, Deragon JM, Striker G (1991) Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol 32:405–14.

[92] Tsirigos A, Rigoutsos I (2008) Human and mouse introns are linked to the same processes and functions through each genome's most frequent non-conserved motifs. Nucleic Acids Res 36:3484–93.

[93] Hadley D, Murphy T, Valladares O, Hannenhalli S, Ungar L, et al. (2006) Patterns of sequence conservation in presynaptic neural genes. Genome Biol 7:R105.

[94] Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37:D755–61.

[95] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–10.

Chapter 5

Identifying functional building blocks of RNA

5.1 Introduction

Visual inspection of RNA structures reveals that there exist repeating structural motifs that occur repeatedly in different RNAs regardless of their evolutionary relationship. As we saw in Chapter 3, several different classes of RNA adopt similar hairpin shapes, perhaps by virtue of the enhanced stability that that degree of base pairing confers [1]. Loops, bulges, and stems all occur commonly throughout various RNAs, and it would seem that there is some sort of structural vocabulary – a common set of RNA building blocks – from which all RNAs are constructed. To what extent are we able to identify and characterize these fundamental units?

Structural components of RNA   There are several possible levels of resolution to use when assembling a collection of RNA building blocks, many of which we have alluded to in previous chapters. Our focus up to this point has been on RNA modules on a somewhat macro scale – all of our RNA species of interest have been on the

length scale of functionally independent RNAs. As modular building blocks, they are components of multimeric RNA species – e.g., one pre-miRNA among several in a primary transcript, or one of many localization signals in an mRNA. Structure catalogs exist for large numbers of these RNAs, ranging from the general – e.g., the RFAM RNA family database [2] – to the specific – e.g., *cis*-regulatory motifs in mRNA untranslated regions are cataloged in UTRsite [3], and microRNAs in miRBase [4].

To generalize away from the level of well-defined RNA families, we can zoom in or zoom out. Zooming out yields an abstract shapes approach, in which RNAs are categorized based on their general shape properties rather than their specific base pairing pattern or sequence. In [5], RNA structures are represented as generalized stems and loops, de-emphasizing specific aspects such as exact stem length, which are hypothesized to be of less importance. Thus, RNA structures will map onto a smaller space of generalized RNA shapes, which can facilitate the discovery of large-scale patterns in natural RNA structures.

At the opposite end of the spectrum is the level of nucleotide sequence. In a strict sense, individual nucleotides – A, C, G, and U – are the fundamental building blocks of RNA structures, and there do exist statistical properties at the level of nucleotide content that both unite and distinguish various classes of RNAs [6]. However, given the compact size of the nucleotide alphabet, a single-nucleotide representation of RNA building blocks will not have enough power to capture much functional significance, in addition to failing to capture any structural information, which by definition involves multiple nucleotides.

Identifying and cataloging groups of nucleotides as RNA structural building blocks has mostly proceeded through a biophysical approach, relying on three-dimensional models of small RNA structures derived from crystallography experiments (e.g., [7, 8, 9]). In this sense, the prototypical RNA building block is defined as a fully-

specified structural element – i.e., a full three-dimensional description rather than just the base-pairing configuration – that appears in natural RNAs at a high rate. Examples of such elements include various tetraloop hairpin motifs such as the GNRA or UNCG, distinguished by the tertiary interactions between bases in the loop; the "kissing hairpin loop" formed by base pairing of two loop sequences [10]; and the U turn, characterized by a sharp bend in the RNA backbone followed by a base stacking interaction between two nucleotides [11].

Characterizing the full repertoire of RNA structural motifs at this level of resolution will be vital to understanding the ways in which larger RNAs are composed; such a task will require additional physical data, perhaps aided by more accurate molecular simulations from increased computing power. In the meantime, heavy reliance on primary and secondary structure as surrogates for full three-dimensional specification is necessary. A large number of strategies exist for RNA motif identification that consider only sequence and secondary structure (reviewed in [12]). The general procedure consists of assembling a set of RNA sequences, then searching for a previously identified structural element or mining the sequences for *de novo* common structure patterns.

FUNCTIONAL COMPONENTS OF RNA   An implicit assumption we make when we identify a common structural motif in two RNAs is that there is a functional analogy present that is due to the structural similarity. RNA function does follow directly from structure [13], a fact that is often exploited for inferring the function of novel RNAs that are sufficiently similar in structure to RNAs that have already been well characterized. Structural motif finding proceeds under this assumption. Common function is assumed either from the groups of sequences compared (e.g. [14]) or from the motif specification used in the search (e.g., [15]), which can be an effective ap-

proach to find novel instances of specific structural motifs when they are hypothesized to exist.

In assembling a set of RNA structural building blocks, our goal is actually to assemble a set of RNA *functional* building blocks, which possess structures specific to their function. Thus, fully attaining this goal requires characterization of the functional repertoire of RNAs and elucidating the mapping between these functions and the structured components that carry them out. As we described in Chapter 2, various combinations and flavors of four basic RNA functions – nucleotide recognition, catalysis, scaffolding, and biomolecule binding – define the functional specificity of an RNA.

To our knowledge, an adequate catalog of the diversity of RNA function does not exist. Specific functional annotations do exist for individual RNA families such as tRNAs or snoRNAs [2], but comparisons between distinct classes of RNAs is difficult due to the lack of a unified vocabulary of RNA function. In response to this deficit, the creation of an RNA ontology has been proposed [9], but as of September 2009, its status is unclear.

Some attempts have been made to broadly classify RNA functionality. In [16], the authors created an integrated RNA database called NONCODE that collated RNA sequences from disparate sources to unify their nomenclature. They used a semi-automatic process to extract research articles of interest containing examples of known RNAs, then manually curated the results into 111 different RNA classes. Additionally, they created a process-function classification (PfClass) scheme consisting of 26 keywords such as "RNA_editing" and "Protein_transport" with which to annotate the classes. However, since over half of the PfClass keywords annotate only one or two classes, the PfClass terms are of limited use for finding common functionality among multiple RNA families.

Creating an RNA ontology manually would be a laborious and subjective process, given the growing number and diversity of RNA families that have been characterized thus far. An appealing strategy would be to construct the ontology automatically using the information and annotations that already exist for many RNA families in databases such as RFAM. Information extraction from free-text sources has been exploited for biological ontology construction and augmentation, notably for the Gene Ontology (GO) [17], a manually curated, hierarchical vocabulary specific to the function and components of protein-coding genes. In [18], the authors used the statistical properties of words found in MEDLINE abstracts associated with specific genes to assemble a set of relevant ontology terms with which to annotate those genes. Two challenges exist for this sort of approach: an appropriate vocabulary must be generated that is general but at the same time induces different, meaningful partitions on the set of annotated items (genes); and the ontology terms must be used to accurately annotate a list of genes, such that each term is a function or characteristic belonging to a gene in question. Ideally, an ontology is specific to a particular domain or subdomain, such that using an ontology for a different purpose than it was intended may not be an optimal solution.

CHAPTER OVERVIEW   In this chapter, we take a functionally-driven approach to identifying fundamental RNA building blocks. In Section 5.2 we undertake the completely automated construction of an RNA ontology using information extraction techniques on free-text RNA family descriptions from Wikipedia. We test this ontology in various ways to assess its applicability to RNA biology and show that it can be used as a framework for identifying the functional components of RNAs. In Section 5.3, we use the ontology for two pattern-discovery tasks to map the functions encoded by the ontology to structured components. In a "forward" approach, we as-

194

semble groups of RNAs from unrelated families that share ontology annotations and show that the RNAs contain significantly unique structural motifs that distinguish them from RNAs that are not similarly annotated. These motifs may represent the structured components that confer the functionality in question. In a complementary "reverse" approach, we use a low-level structural encoding to decompose individual RNAs into sets of motifs, then show that specific motifs shared among unrelated RNAs are significantly associated with RNA functions defined by the ontology. In this way, we show that there does exist a repertoire of functional building blocks common among different RNA families, which can be characterized in further detail using wet-experimental approaches and further refinements on the methods presented here.

5.2 FUNCTIONAL CLASSIFICATION OF RNA FAMILIES USING AN AUTOMATICALLY GENERATED ONTOLOGY

Our strategy was to generate a set of semantically meaningful words, each of which describes some aspect of RNA biology, and together span the diversity of RNA function and constitute an RNA ontology. For each RNA family, a subset of these words will be relevant and will serve as a functional annotation the family. Closely related families are expected to be annotated with many words in common, while more distant families with some common functional aspect may share one or a few words. The goal was to create this ontology in a completely automated fashion – i.e., neither the ontology construction nor annotation of RNA families should require human intervention.

Fortunately, a large number of RNA families are included RFAM, the database of RNA families [2], which as of January 2009 contained 1372 families defined by

structural covariance models. Each RNA family is associated with a description that consists of unstructured free text written by experts specifically for that RNA family or superfamily. Recently these descriptions were ported to the Web encyclopedia Wikipedia [19] and thus became freely editable by the scientific community at large.

The Wikipedia entries provide a framework in which each RNA family in RFAM is associated with a set of words relevant to that RNA, though since the words are organized into human-readable descriptions, they do not constitute a workable ontology in their raw state. Our task then was to extract relevant ontology terms from the free-text descriptions, and then annotate RNA families based on whether those ontology terms appear in their respective Wikipedia descriptions.

5.2.1 Constructing the RNA ontology

There are a total of 633 Wikipedia entries spanning the 1372 RFAM families; in several instances, more than one family points to the same Wikipedia entry, which is written broadly enough to encompass all of the RNAs it annotates.

We assembled the Wikipedia documents and scrubbed them of their html and Wikipedia-control content, then normalized word usage using a pipeline that included stemming (i.e., removal of orthographic suffixes such as "-s" and "-ing") (see Materials and Methods). A total of 3306 words appear across all of the Wikipedia documents, with 1557 words appearing in at least two different documents. We used a bag-of-words document model [20], treating each document as an arbitrarily ordered 1557-dimensional vector of word counts.

Next, our goal was to identify a subset of words that are uniquely informative about RNA documents. Of the starting set of 1557 words, we anticipated three broad categories of words: 1) words such as "the," "and," and "is" that will appear very frequently in these documents, but contain little to no semantic information; 2) words

such as "polymerase", "spliceosome," and "riboswitch" that are exclusively found in RNA documents; and 3) words such as "involve," "recent," and "recognize" that are semantically rich but may not be informative about RNA documents specifically. We sought to include only words belonging to the second category and a subset of words in the third category in the final ontology.

To distinguish these word categories, we constructed a background corpus of 67,299 Wikipedia documents from [21] on topics unrelated to RNA biology, and compared the document frequencies of each of the 1557 words in the RNA corpus to their frequencies in the background corpus; document frequency is defined as the number of different documents in which a word appears at least once – i.e. the number of instances of a word in a single document is not considered. Words appearing more frequently in the RNA corpus compared to the background corpus will tend to be specific to RNA biology, while words appearing equally frequently in the RNA and background corpora will tend to be uninformative.

We used several methods to identify the informative subset of the 1557 words and took the union of each of these methods, with the goal of creating the most comprehensive, if not compact, ontology. First, we used a stop list consisting of 424 domain-independent commonly found utility words [22] to serve as a gold standard of exclusion – these will consist of words belonging to the first category described above. Next, we included all words that appear in Gene Ontology [17] category titles but not in the stop list (632 words total).

Finally, we constructed two decision boundaries – a chi-square boundary and an odds-ratio boundary [23] – and retained words with greater RNA document occurrence than background occurrence by an amount defined by the less restrictive of the two boundaries (which was the odds-ratio boundary in most cases). We defined each boundary by calculating the respective value for the set of stop-list words and setting

the boundary such that 95 percent of the stop words are excluded. Accordingly, each boundary would allow 11 stop words among the RNA-specific word set; however, we retained only the six stop words common to both boundaries, comprising 1 percent of the total number of stop words. These words were "small," "known," "non," "c,", "o," and "d," each of which having biological significance: "small" as in "small RNA," known in the sense of experimentally verified, "non" as in "non coding," "c" and "d" referring to the C box and D box sequence motifs respectively, and "o" referring to oxygen. Figure 5.1 illustrates these decision boundaries and the words that were ultimately retained.

The final ontology consisted of 991 words (see Appendix), which is 30 percent of the total number of words appearing in any RNA document and 63 percent of those words appearing more than once. Figure 5.2 shows the distribution of document frequencies for each of the terms. The top 5 percent of the terms according to document frequency each annotate from 25 percent ("element") to 90 percent ("rna") of the 633 Wikipedia RNA classes. The majority (843) of the terms annotate fewer than 5 percent of the RNA classes.

5.2.2 Verifying specificity of the ontology

We verified the specificity of the ontology through a series of classification tasks using a 991-dimensional document vector representation for the RNA documents.

First, we performed hierarchical clustering on the RNA document vectors. Visual inspection of the dendrograms revealed that easily identifiable large groups of RNA families such as the miRNAs and snoRNAs tend to cluster together in word space (Figure 5.3). Given the diversity of document content especially for the individual miRNA families, the ontology captures sufficient essential information from the RNA documents that is commonly shared among closely related RNAs.

Figure 5.1: Document frequencies of candidate RNA ontology words showing decision boundaries used to determine relevance. Words are plotted according to their document frequency in the set of background documents and the frequency in the set of RNA documents (points are jittered for better visualization). Retained words (including words retained because they are used in GO annotations) are colored blue, discarded words are colored in orange, and stop words are colored in dark orange. The two dashed lines indicate the decision boundaries defined by the chi-square statistic (left-most dashed red line) and the odds ratio (dashed green line). The solid blue line indicates equal background and RNA document frequency of occurrence.

Next, we constructed a series of support vector machine (SVM) classifiers to test the efficacy of the ontology in distinguishing between RNA documents and non RNA documents. First, we trained an SVM to distinguish the 633 RNA documents from a randomly selected 1266-subset of background Wikipedia documents. Training accuracy was 100 percent for both the positive and negative training sets. To test accuracy on independent documents, we generated four additional corpora: a disjoint subset

Figure 5.2: Document frequencies of ontology terms. The 991 ontology terms are ordered on the x axis by document frequency. Individual examples of terms with document frequency in parentheses are shown.
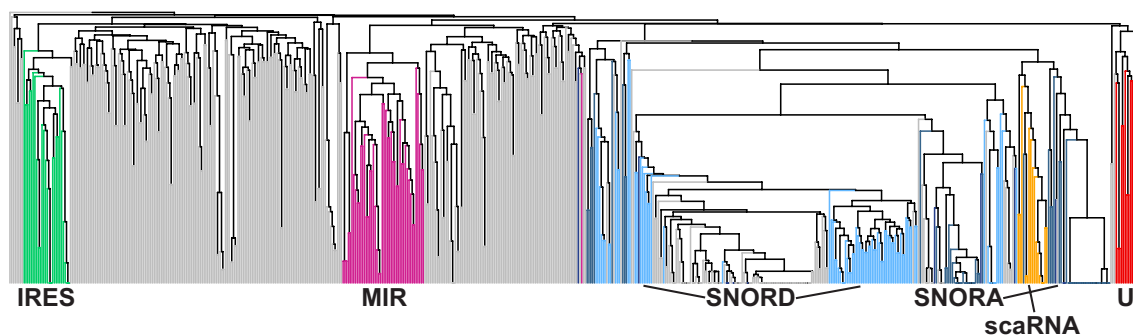
Figure 5.3: Cluster dendrogram of RNA documents showing co-clustering of related RNA families. Documents are clustered using average linkage on the Euclidean distances of the feature vector, defined as presence/absence of ontology terms. Highlighted RNA families are IRES RNAs (green), microRNAs (MIR, purple), snoRNAs of the H/ACA (SNORA) and C/D (SNORD) box varieties (dark blue and light blue respectively), scaRNAs (orange), and spliceosomal RNAs (U, red).

of 633 background Wikipedia documents, a set of general-topic biological abstracts extracted from PubMed, a set of RNA-specific abstracts from PubMed, and a set of Reuters news articles from the Reuters-21578, Distribution 1.0, corpus [24], which

200

Table 5.1: Document classification accuracy using the RNA ontology

Test set	Doc. count	Percent documents correctly classified		
		Full ontology	Low-frequency[a]	Random average[b]
RNA training set	633	1.00	0.90	1.00
Background training set	1266	1.00	1.00	0.96
Background test set	633	1.00	0.99	0.92
Pubmed general	633	0.83	0.66	0.58
Pubmed RNA	633	0.85	0.83	0.63
Reuters news articles	633	1.00	0.99	0.89

[a]Subset of full ontology containing the 85 percent least frequently used terms (n = 849). [b]Average of 10 randomly-generated 849-word subsets of the full vocabulary from which the RNA ontology was extracted.

constitute a truer negative test set than the Wikipedia articles since they were not consulted during ontology construction. Test accuracy was high for all four test sets, with the Wikipedia and Reuters sets achieving 100 percent classification as non-RNA related, the general-biology abstracts achieving 83 percent classification as non-RNA related, and the RNA-specific abstracts achieving 85 percent classification as RNA related (Table 5.1).

To assess the performance of the ontology terms less frequently appearing in the RNA documents, we partitioned a subset of the 849 terms with the lowest RNA document frequency ($\sim$ 85 percent of the total ontology) that spanned 90 percent of the RNA documents; the most common term in this subset, "spliceosome," appears in 33 different RNA documents, or $\sim$ 5 percent. We repeated the SVM classifier tests using the subset (low-frequency subset), and again found favorable performance (Table 5.1). Correct classification of general-topic biology abstracts did decrease to 66 percent, but considering that some of the ontology terms likely have applicability to non-RNA-specific functionality, this result does not necessarily detract from the efficacy of the ontology.

Finally, we randomly extracted 10 849-term subsets from the original candidate

word list of 1557 words and compared performance of these random subsets to the performance of the ontology low-frequency subset when used to train an SVM for RNA document classification. These random subsets exclude the 207 terms occurring in more than 33 different RNA documents, in order not to bias the subsets in favor of highly identifiable RNA words such as "rna" or semantically generic words such as "the." The low-frequency subset outperformed all of the random subsets in all test cases; given that training accuracy on the RNA documents actually decreased in the low-frequency subset, we can conclude that the ontology terms generalize better to unseen RNA-related documents than randomly selected words appearing in RNA Wikipedia articles (Table 5.1).

5.3 Functionally related RNAs display characteristic structural signatures

With characteristic descriptors in place for a large number of RNA families, we looked for enrichment of sequence and structural motifs associated with functionally related subsets of RNA families. We took two approaches. In one approach, we identified groups of RNA families sharing similar characteristics and performed structural motif finding to identify enriched structures shared between different families. In a second approach, we encoded the RNAs using a low-level structural representation and looked for enrichment in ontological terms associated with these low-level motifs.

5.3.1 Ontologically similar RNA families contain common motifs

Assuming accurate annotation of the RNA families in the Wikipedia documents, we would expect that the groups of RNA families annotated with the same ontological terms would share some sort of biological function or characteristic, which may be

correlated with a set of structural features. To test this, we performed structural motif finding on the sets of RNA families annotated by each of 689 of the ontology terms. We excluded both frequently used terms (ones that annotate greater than one-third of all of the RNA families) and infrequently used ones (annotating fewer than three RNA families), with the goal of isolating biological functions with some specificity.

We used the motif-finding pipeline implemented in RNApromo [25], which is tuned to find common sequence/structure motifs in the size range of 15 to 70 nts; these motifs each constitute well-formed structures – i.e., all base pairing occurs within the motif and not to sequences outside of it. For each ontology term, we assembled a balanced subset of RFAM RNA sequences belonging to the families associated with the term, ensuring that no one family dominated the sequence set, which would bias the motif finder toward structural elements specific to that RNA family (see Materials and Methods). Where common motifs exist, up to 10 possibly overlapping structure definitions are returned in the form of a set of covariance models.

Each structural motif induces a distribution of likelihood scores for each set of input RNA sequences; if the motif is in fact characteristic of some shared structure/function unique to the specific RNA families, then these likelihood scores should be significantly higher than those produced for a background set of unrelated RNA sequences. Assembling an appropriate background set is a non-trivial task, as length, nucleotide content, and other structural features has an impact on the structural richness of an RNA sequence [26]. Using nucleotide-shuffled same-length versions of the input sequences does control for length, but much of the structural potential is lost [27]. Thus, we constructed custom background sequence sets for each group of input RNAs, drawing randomly from real RFAM RNA sequences belonging to families not included in any of the input RNAs. These sequences are truncated or concatenated

in order to match the length distribution of the input set.

Given the input and background RNA sets, we calculated likelihood scores for each of the RNAs and performed a t-test, testing for the null hypothesis that the two sets of RNAs are drawn from the same distribution of likelihood scores. We applied a conservative Bonferonni correction for multiple hypothesis testing (number of tests $\sim 10,000$) and found that 339 of the ontology terms annotated RNA families sharing significantly similar structural components, using a $p < 0.05$ cutoff.

The 10 most significant motifs and their associated ontology terms are presented in Figure 5.4. The terms "mi#" and "mipf#" are both gene symbol prefixes for miRNA genes, and as expected their motif is a long stem-loop structure characteristic of a pre-miRNA. Interestingly, there seem to be weak nucleotide preferences along the stem and in the loop that may indicate Drosha- or Dicer-imposed constraints on substrate specificity.

Two additional highly significant terms are "movement" and "attenu[ate]," both of which annotate a group of RNA motifs that occur upstream of bacterial operons. These RNAs, exemplified by the Tryptophan operon leader, are regulators of transcription by a mechanism of negative feedback. During transcription of the trp operon, the leader sequence is transcribed first and is immediately bound by a ribosome to commence protein translation (a phenomenon that occurs only in prokaryotes, which do not sequester transcription inside a nucleus). The leader encodes a short peptide consisting of several consecutive tryptophans. Under high concentrations of tryptophan, this peptide is rapidly translated, which causes a conformational change in the mRNA at the site of the leader RNA. The hairpin structure that forms blocks the *movement* of RNA polymerase, leading to the *attenuation* of transcription [28, 29].

Other significant ontology terms shown include "nucleophil[e]," which annotates 23S rRNA, U2 spliceosomal RNA, Hammerhead ribozyme, and Group II intron, and

Figure 5.4: Most significant common motifs occurring in RNAs annotated by ontology terms. P-values are calculated as described in the text and are Bonferonni corrected. Nucleotide color indicates the degree of similarity at that position in the RNAs where the motif appears, ranging from red, indicating 100 percent shared, to green, 50 percent shared. Structure elements (free bases and base pairs) are similarly shaded according to the degree of similarity among the RNAs, with black indicating 100 percent shared.

thus refers to the nucleophilic attack reactions involving or catalyzed by these RNAs; "coronavirus," which annotates several separate RNAs specific to coronaviruses; and "minus," which annotates several viral RNAs involved the regulation of minus-strand RNA synthesis.

In the case of the term "muscl[e]," the annotated RNAs appear all to be RNAs that are expressed in muscle cells, consisting of three miRNAs and a snoRNA. Based on the structure of the motif, a long hairpin with extensive nucleotide similarity, it is likely that the motif in fact captures similarity in only the muscle-specific miRNAs.

To refine the search, we repeated the procedure using pairs of ontology terms – i.e., Wikipedia documents annotated by both terms in a pair are purported to share functionality defined by the two terms. Starting with each single ontology term, we selected a second term that co-occurs with the first term in a subset of the Wikipedia documents. This second term was selected to maximize informativeness as measured by entropy – i.e., if a set S of Wikipedia documents is annotated by the first term, and a subset $T \subseteq S$, where $|T|$ (size of T) is a fraction p of $|S|$, is additionally annotated by some second term x, we calculated $I_x = -p \lg p - (1 - p) \lg(1 - p)$ for each x and chose x to maximize I_x. Intuitively, if the two terms annotate nearly all of the same documents annotated by the first term alone, the second term adds little information; conversely, if the two terms annotate a very small fraction of documents, the second term is too specific. After grouping together term pairs annotating identical document subsets, we obtained 316 groups of ontology terms that contain significantly similar motifs.

The most significant motifs are annotated by ontology terms which themselves already define highly significant structural motifs in isolation. In some cases the additional terms serve to disambiguate the biological characteristic. For instance, the term "movement" now unambiguously is associated with attenuation, since it appears

Table 5.2: Most significant ontology term combinations

Ontology terms	p-value of most significant motif[a]
small, mi#	1.08E-64
threeprim[e], extent	2.15E-62
imped[e], upstream, defici[ent], attenu[ate], movement	2.60E-51
small, muscl[e]	3.46E-47
nucleotid[e], purin[e]	1.03E-46
ribosom[e], excess	2.16E-46
nucleotid[e], coronavirus	6.81E-46
bacteri[a], antitermin[ate]	1.29E-44
known, bacillus	1.46E-44
ribosom[e], 23s	7.79E-44

[a]P-values are calculated as described in the text and are Bonferonni corrected.

along with "attenu[ate]" and other associated terms (Table 5.2). However, in other cases, using multiple ontology terms to generate RNA subsets does narrow down the functional aspect of interest. Figure 5.5 shows examples of these, including "proteobacteria" + "upstream," which annotate a set of riboswitches in Gram-positive bacteria; and "mrna" + "transport," which annotate transcripts that undergo sub-cellular compartment localization.

It is worth noting that arbitrary combinations of ontology terms can also be used to return a set of related RNA families that may contain common structural motifs, as an alternative to manual curation of relevant RNA sequences to investigate a function of interest.

5.3.2 SMALL RNA MOTIFS ARE ENRICHED IN SPECIFIC ONTOLOGY TERMS

Characterizing structure-function relationships can also proceed in a reverse direction – we can ask whether a specific structural motif appears in different RNAs from families that share a biological function, as defined by the RNA ontology. In an analogous process to performing Gene Ontology enrichment analysis [17], we sought
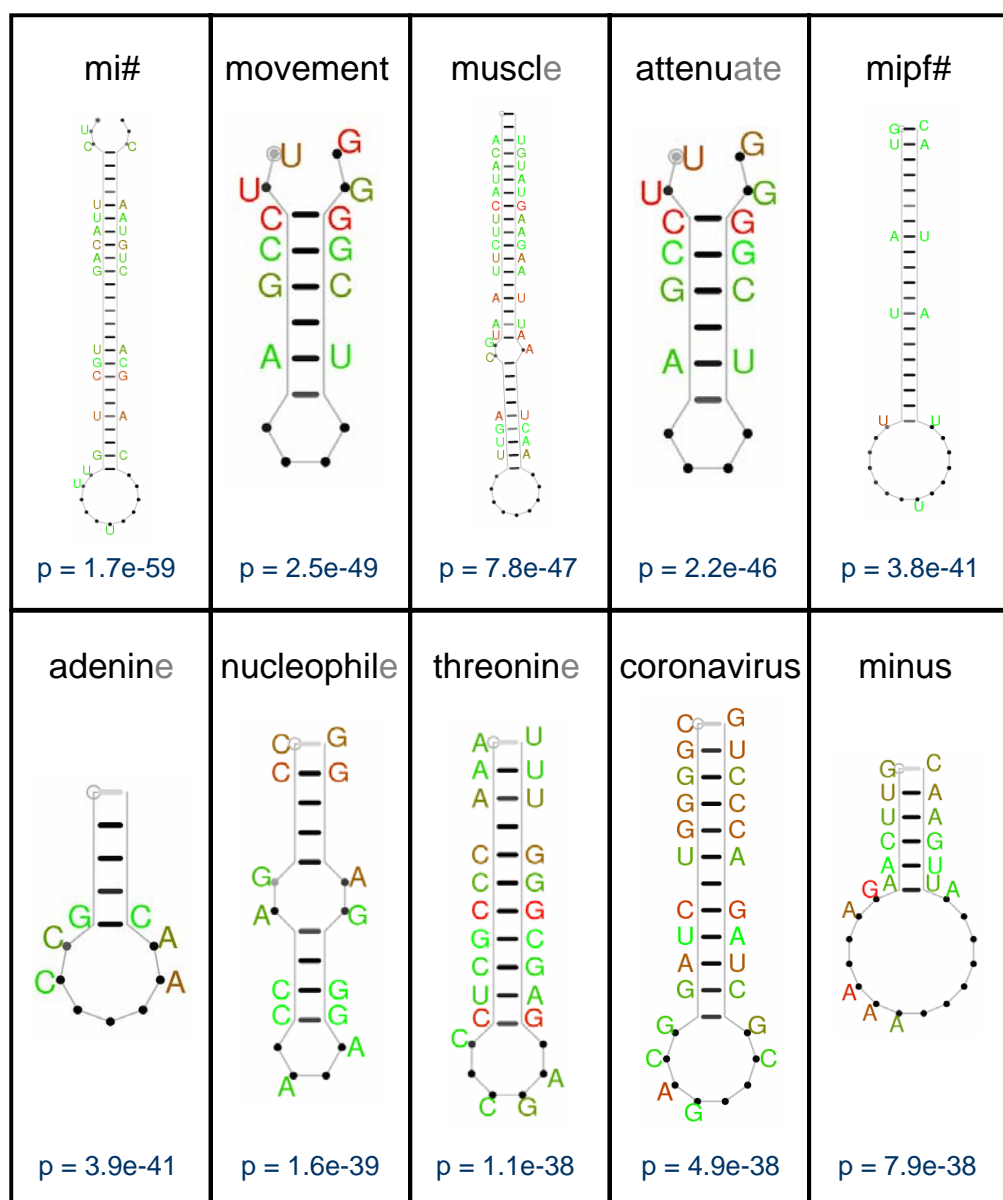
Figure 5.5: Significant common motifs occurring in RNAs annotated by pairs of ontology terms. P-values are calculated as described in the text and are Bonferonni corrected. Nucleotide color indicates the degree of similarity at that position in the RNAs where the motif appears, ranging from red, indicating 100 percent shared, to green, 50 percent shared. Structure elements (free bases and base pairs) are similarly shaded according to the degree of similarity among the RNAs, with black indicating 100 percent shared.

to look for enrichment of specific RNA ontology terms in small structural motifs spanning multiple RNA families.

In the previous section, we focused on small, well-formed motifs in the length range of 15-70 nucleotides, which was a consequence of the motif finding algorithm we used (RNApromo). The motifs returned were predominantly hairpin shapes (Figures 5.4 and 5.5) that corresponded to larger-scale characteristics – i.e., the miRNA precursor hairpin. However, we were also interested in identifying smaller, less structured components that may have specific functionality – e.g., a catalytic domain or a recognition motif. For example, C/D snoRNAs are defined by two small motifs, a six-nucleotide C box (UGAUGA) and a four-nucleotide D box (CUGA) [30]. The

BC1 RNA contains at least three localization motifs, one of which is a single U nucleotide bulge in an otherwise uninterrupted stem region [31]. Although these motifs generally require appropriate context in which to function – for example, the C and D boxes must be positioned opposite each other in the secondary structure in order to facilitate tertiary interactions – we hypothesized that it is possible to recognize these signals in isolation, on the basis of the functional annotations of the RNAs that contain them.

In designing an RNA structure database search protocol, Xue et al. defined a representation scheme in which the linear sequence/structure of the RNA is partitioned into a set of k-nucleotide segments [32]. Each segment encodes a motif that consists of the structure of the k consecutive nucleotides (unpaired or paired) as well as some amount of sequence information, depending on the degree of generality desired. In this way, an RNA structure is scanned from $5'$ to $3'$ using a sliding k-nt window, and can be represented as an unordered count of the number of times each motif appears, or as a binary vector indicating presence/absence of a particular motif in the RNA (Figure 5.6); thus, the presence of common motifs among different RNAs is manifested in the similar vector components in the RNA representations.

For the RScan database scanning task, the authors used k=7-11 which led to a highly specific representation for individual RNAs, thus an efficient database lookup. Xue et al. also used their RNA representation in their work predicting miRNA genes, and used k=3 to capture general properties of miRNA hairpins compared to hairpins occurring in protein-coding regions [33]. Based on initial tests, we chose to use a degree of specificity conferred by a motif size of k=4. Each nucleotide in an RNA sequence can be in one of three structural states – base paired with a downstream nucleotide, "("; base paired with an upstream nucleotide, ")"; and unpaired, "." – and have one of four base identities – A, C, G, U – for a total of 3x4 = 12 different

GCCACUUUG**GGAU**GGCGGG**GCAC**UGGGGAUGCUC
((((. . . . . **. .)**))) . ((**((.** . . . . . .))))))

..))GAUG	2	))))UGCU	1	((((GGCA	1
...)GGAU	2	))))UGGC	1	((((GGGC	1
....CUGG	1	))).GGCG	1	(((.CCAC	1
....CUUU	1	)).(GCGG	1	(((.GCAC	1
....GGGA	2	).((CGGG	1	((..CACU	2
....GGGG	1	.(((GGGG	1	(...ACUG	1
....UGGG	2	.)))AUGC	1	(...ACUU	1
....UUGG	1	.)))AUGG	1	))))GCUC	1
....UUUG	1	((((GCCA	1		
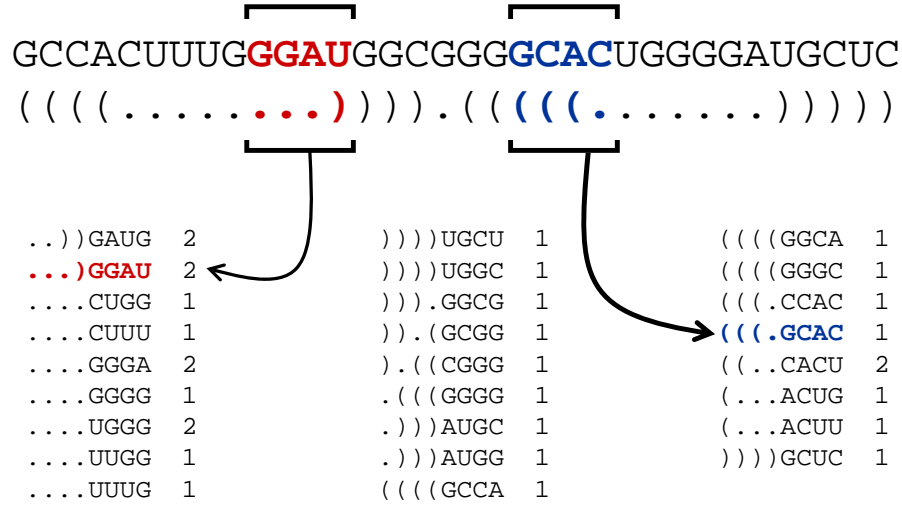
Figure 5.6: Example of structural motif encoding. Two different motifs are illustrated on the RNA sequence. Numbers next to each motif indicate the number of times the motif appears in the RNA.

characters per nucleotide. For $k=4$, there are $12^4 = 20,763$ possible motifs, though not all of these are biophysically feasible.

Using this representation, we scanned the RFAM sequence set, which consists of 154,875 sequences containing a total of 18,991 different 4-nt motifs. To determine whether any of these motifs possessed putative functional significance, we associated each motif with a set of ontology terms based on the RNAs that contain that motif: given a motif m, we generated a list of the n individual RNAs that contain m; a single RNA may contain multiple copies of m, in which case that RNA is counted twice. Each of the RNAs in the list will belong to some RNA family that is annotated by a set of ontology terms. Then, for each motif, we can assemble counts of the number of times an ontology term is used to annotate an RNA in the list; these counts will sum to n.

Enrichment of any one ontology term for a motif is determined by comparing the

observed number of times the term annotates one of the n RNAs in the motif list to the expected number of times that term should appear in a list of size n, given population-wide frequencies. These expected frequencies are calculated by summing the total number of RNA segments associated with an ontology term and dividing by the total number of motif instances (22,901,664). To determine significance, we calculate a Chi square goodness-of-fit statistic and apply a conservative Bonferonni correction for multiple tests (2×10^4 motifs $\times$ 10^3 ontology words $= \sim 10^8$ individual tests, leading to a Chi square critical value of approximately 36 to achieve a corrected p value < 0.01). We additionally filtered the ontology terms to exclude any term that annotates fewer than five different RNA families containing the given motif, to ensure that we were not selecting ontology terms that were overly specific.

In all, 6659 of the motifs showed significant ontology term enrichment. The most significant of these are shown in Table 5.3. Motifs are represented as an eight-character string, with the first four characters specifying the structure of the motif and the second four characters the sequence. Based on the ontology terms returned, these motifs appear to be characteristic structures in specific superfamilies of RNA that are highly represented in RFAM. For example, two of the top motifs are annotated with terms pertaining to H/ACA snoRNA function ("aca#," "snora#," "uridine"), and are presumably highly conserved motifs among these RNAs. The motif enriched in "revers[e]," "transcriptas[e]," and "step" are drawn from RNAs with *cis* 3′ regulatory function that happen to include retrotransposon-associated structures such as the R2 RNA element and the eel UnaL2 LINE 3′ element.

Among the motifs with slightly less significant term enrichment are ones that have previously been functionally characterized. Of note is the D box motif, which consists of an unstructured CUGA region ("....CUGA"), and is highly enriched in terms specific to C/D type snoRNAs including the easily identifiable "cuga" and "cslashd"

Table 5.3: Top 5 motifs with most highly enriched ontology terms

Motif	Term	Observed[a]	Expected[b]	Enrichment	Chi-square[c]	N. families[d]
(...GUUC	modifi	14032	2057.2	6.82	79214.2	28
	remov	13898	2053.3	6.77	77631.3	5
	arm	14531	2265.9	6.41	76505.6	37
	enzym	13932	2117.6	6.58	75209.6	19
	recognit	14024	2236.2	6.27	71464.1	28
((..GGUU	modifi	14595	2490.4	5.86	66863.1	33
	remov	14464	2485.6	5.82	65584.3	6
	polypeptid	14937	2669.9	5.59	64689.8	6
	arm	15090	2743.0	5.50	64045.6	41
	enzym	14455	2563.4	5.64	62942.1	13
))..CUAC	mbi	7049	796.2	8.85	53435.3	13
	aca#	7183	985.8	7.29	43302.9	40
	molecular	7156	1013.0	7.06	41530.9	32
	uridin	7273	1174.9	6.19	35947.6	66
	snora#	7261	1180.5	6.15	35593.3	59
....CAGU	mbi	9794	1434.7	6.83	52995.7	19
	aca#	10255	1776.4	5.77	44977.3	61
	molecular	10131	1825.5	5.55	42127.4	42
	uridin	10472	2117.2	4.95	37442.8	109
	snora#	10467	2127.4	4.92	37153.9	99
.)))AAAA	transcriptas	1264	32.8	38.59	46658.2	5
	revers	1266	39.3	32.19	38639.6	6
	determin	1816	350.3	5.18	6712.5	61
	step	1897	424.6	4.47	5702.8	86
	element	2064	627.9	3.29	3885.9	64

[a]Number of motifs observed among all RNAs annotated by a term (154,875 total RNAs with 22,901,664 total 4-mers). [b]Number of 4-mers expected to be annotated by a term, given the RFAM-wide frequency. [c]Chi-square statistic per term given observed and expected frequencies. All p-values are $<< 10^{-300}$. [d]Total number of families with RNAs containing the motif and annotated by the term.

Table 5.4: Top 10 ontology terms enriched in RNAs containing the D box motif

Term	Observed[a]	Expected[b]	Enrichment	Chi-square[c]	N. families[d]
cuga	1974	411.9	4.79	6119.8	270
cslashd	2245	521.8	4.30	5931.9	288
ugauga	1942	410.0	4.74	5912.7	268
snord#	1544	337.1	4.58	4437.5	189
snorna	2541	782.2	3.25	4210.9	332
biogenesi	1970	521.2	3.78	4197.9	277
box	2875	968.1	2.97	4062.5	323
ribos	1416	315.5	4.49	3935.4	184
modifiy	1095	276.9	3.95	2470.5	151
ncrna	1240	339.5	3.65	2453.0	176

[a]Number of D box motifs ("....CUGA") observed among all RNAs annotated by a term (154,875 total RNAs with 22,901,664 total 4-mers). [b]Number of 4-mers expected to be annotated by a term, given the RFAM-wide frequency. [c]Chi-square statistic per term given observed and expected frequencies. All p-values are $\ll 10^{-300}$. [d]Total number of families with RNAs containing the motif and annotated by the term.

annotations (Table 5.4). Similarly, the ACA box motif has significant enrichment of terms relevant to H/ACA type snoRNAs. In snoRNAs, ACA boxes are normally 3′ of an adjacent stem region, leading to a motif representation of ")...NACA" where the first nucleotide is base paired, with arbitrary base identity. The ")...UACA" motif in particular is strongly enriched in terms relating to mouse brain-specific snoRNAs ("mbi," "mouse," "nucleolar") and may represent a nucleotide preference for this subset of snoRNAs (Table 5.5)

Finally, we sorted all of the significantly enriched ontology terms across all motifs and extracted terms that appear rarely in the enrichment lists. Twenty-two terms are enriched for only one motif, including "gc," "stress," and "pathogen" while an additional 20 are enriched in only two motifs, including "shock," "hydrolysi[s]," and "fibrillarin." Two examples are shown in Table 5.6: "gc" appears in a term enrichment list along with "untransl[ate]" and "riboswitch"; and "antitermin[ate]" appears in two lists along with "gram," "posit[ive]," "bacteria," and "leader." Based on these annotations, these motifs appear to be characteristic to RNAs involved in bacterial

Table 5.5: Top 10 ontology terms enriched in RNAs containing the ACA box motif

Term	Observed[a]	Expected[b]	Enrichment	Chi-square[c]	N. families[d]
mbi	5380	659.3	8.16	36778.5	14
aca#	5492	816.3	6.73	29766.1	39
molecular	5469	838.9	6.52	28491.0	31
snora#	5691	977.6	5.82	25827.0	71
uridin	5665	972.9	5.82	25699.3	77
hslashaca	5782	1089.4	5.31	23337.3	95
nucleolar	5789	1180.9	4.90	21031.8	111
class	6270	1640.7	3.82	16358.9	122
mous	5545	1344.1	4.13	15726.8	67
modif	5914	1685.7	3.51	13375.6	109

[a] Number of ACA box motifs ("...UACA") observed among all RNAs annotated by a term (154,875 total RNAs with 22,901,664 total 4-mers). [b] Number of 4-mers expected to be annotated by a term, given the RFAM-wide frequency. [c] Chi-square statistic per term given observed and expected frequencies. All p-values are $<< 10^{-300}$. [d] Total number of families with RNAs containing the motif and annotated by the term.

attenuation.

5.4 Discussion and conclusions

In this chapter, we showed it is possible to identify functional building blocks common to ostensibly unrelated RNAs. Using an automated information extraction approach on a set of Wikipedia free-text RNA family descriptions, we constructed an ontology of RNA-specific terms that encode functional aspects of RNA biology. We showed that this ontology has a range of specificity in terms of the numbers of distinct families each term annotates, and that together even the rarest of these terms can distinguish RNA-related content. Under the hypothesis that the terms in the ontology constitute a set of fundamental functions, we sought to associate these functions with structural patterns across all RNA families. RNAs annotated by specific ontology terms were found to contain significant characteristic structural motifs, and individual structural motifs spanning multiple different families were found to be significantly associated

Table 5.6: Motifs containing rare ontology terms

Motif	Term[a]	Observed[b]	Expected[c]	Enrichment	Chi-square[d]	N. families[e]
..((CAAC	**gc**	25	2.1	11.83	248.1	5
	untransl	76	25.2	3.02	104.4	5
	low	28	6.5	4.32	71.8	8
	riboswitch	107	50.0	2.14	67.0	11
	modif	454	331.2	1.37	57.5	49
))))UCGU	**antitermin**	325	46.7	6.96	1673.5	5
	gram	410	82.9	4.95	1313.4	8
	leader	565	177.5	3.18	878.1	21
	posit	1437	713.6	2.01	860.6	67
	bacteria	1737	1033.8	1.68	608.7	33
....CGUU	**antitermin**	185	33.8	5.47	682.4	5
	gram	216	60.0	3.60	413.0	11
	box	535	263.1	2.03	303.9	101
	bulg	373	171.9	2.17	247.6	36
	element	855	539.4	1.59	218.4	91

[a]Rare ontology terms are bolded. [b]Number of motifs observed among all RNAs annotated by a term (154,875 total RNAs with 22,901,664 total 4-mers). [c]Number of 4-mers expected to be annotated by a term, given the RFAM-wide frequency. [d]Chi-square statistic per term given observed and expected frequencies. [e]Total number of families with RNAs containing the motif and annotated by the term.

with specific ontology terms. Together these results reflect an underlying structure-function relationship that can be encapsulated in the form of elemental building-block units, which combine together to form a functional RNA species.

The fact that common structural motifs exist between related families is perhaps not surprising; many of the families are defined in a way that facilitates natural groupings, such as snoRNAs. However, the framework provided by an RNA ontology allows longer-distance relationships and commonalities to be discovered. For example, we found a motif annotated by the term "nucleophil[e]" that unites several different ribozymes and reflects the presence of common catalytic domains that perform the same biochemical function. Similarly, we found motifs associated with general "mRNA" "transport," a mechanism known to involve disparate RNA factors in different settings, which may have previously unreported similarities reflected in these motifs.

Our results are notable in that we used a relatively sparse corpus, consisting solely of short Wikipedia descriptions, to construct the ontology. Previous attempts at ontology construction (e.g., [18]) relied on extracting statistical signals from several documents per annotation, which would likely result in a more precise and focused ontology. That we are able to construct a functionally rich ontology that can facilitate in identifying functionally relevant RNA structures is evidence that there exists a strong association between structure and function beyond family-level functional annotations, that can be elucidated using computational techniques.

Further refinements of the ontology will lead to stronger associations of function with structure. Though the ontology in its current state seems to be semantically rich, it is not sufficiently compact, as evidenced by the inclusion of terms with high degrees of semantic similarity. Some of this is due to limited detection of morophological forms, a situation that can be helped by replacing the Porter stemmer with a more

216

complex morphology detection scheme. Whereas the Porter stemmer is a rule-based method, a statistical-based approach may be appropriate here, in which stemming behavior is based on the occurrences of words and roots as measured in a large corpus [20]. The fact that a large amount of scientific jargon with irregular derivations appears in biological text suggests that word usage should be taken into account in this task.

A large part of the ontology redundancy is also due to synonymy, where many words are different ways of expressing the same thing. Given a larger corpus, we might be able to take advantage of machine-learning methods (e.g., as used in [34]) that detect word associations and can collapse synonyms (e.g., different gene names) into single unified terms.

In applying the ontology to identify functionally related structural motifs, we used the motif detection algorithm implemented in RNApromo, as well as a version of the low-level motif representation used in RScan. However, the pipeline we present is general, such that any of a large number of motif-finding methods may be used in their place, depending on the nature of the task. For instance, a search for function-ally relevant stem structures could proceed using a search algorithm well suited to such structures (e.g., [35]). One particular area of interest is bistable RNAs such as riboswitches [36], in which two or more energetically-similar conformations exist for the same sequence [37]. In this case, a structural representation that simultaneously considers both conformations could be used to detect similarities with other RNAs that may not be found using a single static structure.

Our work represents the first large-scale attempt to encapsulate RNA functionality in way that reveals aspects of the organizing principles that define RNA structures. Thus, we present a model in which RNAs are composed of conceptual and physical building-block components that can be individually characterized. A broad under-

standing of these components will help in reconstructing the evolutionary histories of the wide diversity of extant and ancestral RNAs and will facilitate the annotation of novel RNAs.

5.5 MATERIALS AND METHODS

SOFTWARE AND IMPLEMENTATION   All computation was performed using custom-written Python and R code run on quad-core Linux machines with 16GB of memory. SVM construction and prediction was done using the R package e1071. RNA structure prediction was performed using Vienna RNAFold 1.7 [38]. RNA sequences were obtained from RFAM 9.1 full sequence lists [2] and filtered to exclude highly similar sequences using Cd-hit, which implements a greedy clustering algorithm [39]. Version 1.2 of the Gene Ontology was obtained from the Gene Ontology Website [40].

CORPUS PREPARATION   The 633 Wikipedia documents were converted to clean ASCII using the Linux `tr` command and stripped of all html tags and Wikipedia special syntax, defined as any set of characters nested in angled or square brackets (e.g., "<tag>" or "[1]"). All whitespace was converted to single spaces, and a set of known abbreviations were converted to a standard nomenclature (see Table 5.7). We performed sentence boundary detection to disambiguate abbreviations containing the period (".") character and treat such abbreviations as single words. Subsequently, all non-alphanumeric characters were replaced with spaces and all alphabetic characters were converted to lower case.

Next, we extracted a dictionary of all alphanumeric strings (5143 words total) appearing anywhere in the cleaned Wikipedia documents. Each word was stemmed using a two-pass scheme starting with the standard Porter stemmer [41] implemented in Snowball [42], which uses orthographic cues to detect common suffix structure (e.g.,

218

Table 5.7: Known abbreviations converted prior to ontology creation

Abbreviation	Converted form
3'	threeprime
5'	fiveprime
2'0	twoprime o
2'	twoprime
H/ACA	HslashACA
B/C	BslashC
C/D	CslashD
P10/11	P10slash11
J2/3	J2slash3
C'/D	CprimeslashD
G/C	GC
A'	Aprime
C'	Cprime
D'	Dprime
5.8S	fivepointeightS
5.8 S	fivepointeightS
i.e.	i.e
et al.	et.al
vs.	vs
C. elegans	C.elegans
B. subtilis	B.subtilis
E. coli	E.coli
C. difficile	C.difficile
S. typhimurium	S.typhimurium
D. melanogaster	D.melanogaster
H. influenzae	H.influenzae
S. coelicolor	S.coelicolor
A. thaliana	A.thaliana
S. aureus	S.aureus
Y. pestis	Y.pestis
V. cholerae	V.cholerae

plurals, verb tense), followed by a custom protocol to handle suffixes of biological words: 1) any word greater than two letters long that ends with "s" has the "s" removed if the resulting root also appears in the dictionary (e.g., "miRNAs" becomes "miRNA"); and 2) gene names of the form [a-z]+[0-9]+ – i.e., one or more letters followed by one or more numbers – are truncated to include only the alphabetic portion plus a generic number marker (e.g., "mrpl20" becomes "mrpl#"). All of the original words in the Wikipedia documents were translated to corresponding words in the reduced dictionary, which consisted of 3306 words.

A background corpus of 67,299 non-RNA-related Wikipedia documents was collected from the Wikimedia XML Corpus [21] main English collection; documents in the subject categories "Agriculture," "Chemistry," and "Physics" were excluded due to possible similarity to biology-related documents. The two PubMed journal abstract corpora were constructed from March 2009 downloads of abstracts using the NCBI Web interface [43]. RNA-specific abstracts were obtained by searching using the keyword "RNA." The Reuters corpus consisted of a random subset from the Reuters-21578 Distribution 1.0 Corpus [24]. All corpora were processed identically to the procedure used for the RNA Wikipedia documents and filtered to exclude documents containing fewer than 17 alphabetic words, corresponding to the minimum-length RNA Wikipedia document.

DE NOVO MOTIF FINDING USING RNAPROMO   RNAPromo was run using default settings on positive-example input sequence sets selected from the RFAM families annotated by each ontology term or pairs of ontology terms. If significant structural motifs are present, up to 10 are returned in the form of individual covariance models and consensus structure diagrams (e.g., as presented in Figure 5.4). To avoid overrepresentation of any single RNA family in each input sequence set, sets were required

to contain at least three different RNA families in balanced proportions such that no one family constituted more than one-third of the total number of sequences. For computational tractability, sequence set sizes were limited to 100 sequences.

Background sequence sets for significance testing were selected individually for each input set to control for possibly confounding length effects. For each sequence in an input set, a background sequence is selected from a pool of RNA sequences not containing family members represented in the input set; the pool of candidate background sequences was filtered to exclude sequences more than 70 percent similar using Cd-hit. If the background sequence is at least the length L of the input sequence, a random subsequence of length L is extracted from the background sequence and retained. Otherwise, a second background sequence is selected and concatenated to the first sequence, then an L-length subsequence is extracted and retained; this process continues as necessary. The resulting background set thus contains an identical length distribution to the input set.

5.6 Appendix: The RNA Ontology

13q14	align	attenu	bound
15a	alloster	au	box
18s	alpha	autoregulatori	bp
23s	alter	axial	brain
25s	altern	bacillus	branchpoint
28s	alu	bacteri	brucei
2c	am#	bacteria	bslashc
43s	amino	bacteriophag	bsubtili
45s	aminoacyl	bacterium	bulg
50s	amp	bakin	bushi
5s	amphibian	bam	bypass
7s	anneal	barley	c
a#	anterior	barr	c#
absenc	anti	bart#	cajal
abund	antibodi	base	canon
ac	anticodon	bcl	cap
aca	antisens	bear	capsid
aca#	antitermin	beet	carboxyl
acceptor	apc	belong	cardiovascular
access	apic	bend	cascad
accumul	apoptosi	beta	catalys
acid	apoptot	bhrf#	catalysi
acquisit	appar	bind	catalyst
act	aptam	biochem	catalyt
action	arabadopsi	biofilm	catalyz
activ	arabidopsi	biogenesi	celegan
addit	archaea	bioinformat	cell
adenin	archaeal	biolog	cellular
adenosin	arm	biosensor	central
adenosyl	array	biosynthesi	cercopithicin
adenosylmethionin	assay	biosynthet	cerevisia
adult	assembl	block	chain
aeruginosa	associ	blot	chang
affect	atp	blue	channel
agrobacterium	attach	bodi	chaperon
alfalfa	attack	bond	characteris

characterist
charg
chemic
chloramphenicol
chloroplast
chromosom
chronic
ciliat
circl
cis
class
classic
cleav
cleavag
cll
clone
cloverleaf
cluster
cm#
co
coaxial
code
codon
coenzym
cofactor
coli
coloni
common
complement
complementar
complementari
complet
complex
compon
compris
concentr

condit
conduct
confirm
conform
connect
connexin
conserv
construct
contact
contain
content
control
convers
coordin
copi
core
coronavirus
coupl
coval
covari
cprime
cre
cress
crick
crinkl
cross
crystal
crystallographi
cslashd
csra
csrb
cuga
cyanobacteria
cycl
cyclic
cyclin

cystein
cytoplasm
cytosin
d
dalgarno
darzacq
dbpa
death
decarboxylas
decay
decreas
defici
degrad
delet
delta
densiti
depend
deriv
destabilis
detect
determin
develop
development
di
dicer
differenti
dimeris
direct
diseas
disrupt
distal
distanc
distinct
distribut
dival
divis

dna
domain
donor
doubl
doublet
downstream
dprime
drosha
drosophila
dsra
dsrna
duplex
dyskerin
e#
ear
ecoli
edit
effector
effici
eif#
eif4f
electron
electrostat
element
elev
elimin
elong
embryo
embryogenesi
embryon
encapsid
encod
endonucleas
endoribonucleas
endotheli
energi

enhanc	famili	genbank	hgcg
enter	fast	gene	hinfluenza
enterobacteri	ferric	genera	hing
enterovirus	ferritin	generat	histidin
entri	fgf	genet	histon
envelop	fibrillarin	genom	hiv
environ	fibroblast	global	holoenzym
environment	fino	glutamin	homeostasi
enzym	finop	glycin	homolog
enzymat	finp	glycogen	homologu
epsilon	fivepointeight	gm#	host
epstein	fiveprim	gram	hox
equin	flavivirus	green	hslashaca
escherichia	flexneri	growth	hsp#
essenti	flj#	gtp	human
establish	fluorescen	guanin	hydrogen
ester	fold	guanosin	hydrolysi
eukarya	follow	guid	hydroxyl
eukaryot	form	hairpin	hyperthermophil
evolutionari	format	half	hypertroph
evolutionarili	fraction	hammerhead	hypothes
evolv	fragment	hbi	hypothesi
excess	frameshift	hbii	ictvdb
excis	free	hbv	ident
exclus	fruitfli	hcv	identifi
exogen	function	heal	ii
exon	g#	heart	iii
experiment	gac	heat	immunodefici
exponenti	gaca	hela	immunoprecipi
export	gag	helic	immunoprecipit
express	gamma	helix	imped
extens	gar#	hepat	imprint
extent	gar1p	herpesvirus	inactiv
f#	gas#	heterolog	incomplet
facilit	gate	hfq	increas
factor	gc	hgca	independ

224

indirect
individu
induc
infect
infecti
inflammatori
inhibit
inhibitor
initi
inosin
insert
interact
interfer
interferon
intergen
intermedi
intern
intra
intracellular
intron
involv
ion
ionic
ire
iron
is#
isomer
isomeris
iv
j#
join
junction
kilobas
kinas
kinet
klug

known
kv#
l#
l13a
l23a
laevi
late
latent
lead
leader
length
leukaemia
leukemia
level
life
ligand
ligat
limit
line
lineag
link
linkag
linker
listeria
live
local
localis
locat
loop
low
lower
ltr
lung
lymphocyt
lymphoma
lysin

lytic
m#
machineri
macrophag
magnesium
main
major
mammal
mammalian
map
mass
matern
matur
mbi
mbii
me#
mechan
mediat
melanogast
membran
messeng
metabol
metabolit
metal
metazoan
methanococcus
methionin
methyl
methyltransferas
mg
mgc#
mi#
mice
micf
microarray
microbi

microrna
minor
minus
mipf#
mir
mirna
mitochondri
mitochondria
mobil
model
modif
modifi
modifiy
modul
moieti
molecul
molecular
monocytogen
mosaic
motif
motil
mottl
mous
movement
mrna
mrp
ms3d
multifunct
multipl
muscl
mutagenesi
mutant
mutat
myc
nascent
natur

ncrna	organ	pocket	promot
near	organell	pol	protect
negat	origin	poli	protein
nematod	ornithin	polyadenyl	proteobacteria
nervous	orthologu	polyamin	proton
neural	oryza	polymeras	proxim
neuron	outer	polypeptid	pseudogen
nmr	overexpress	polyprotein	pseudoknot
nol5a	overlap	pomb	pseudomona
nomenclatur	oxi	porin	pseudouridin
non	oxygen	portion	pseudouridyl
noncod	p#	posit	psi
nop10p	packag	post	psi#
ns5b	pair	potassium	purifi
nt	paralogu	potenti	purin
nuclear	particl	prader	putat
nucleic	patern	pre	pws
nucleolar	pathogen	precursor	py
nucleolin	pathway	predict	pyrimidin
nucleolus	pattern	pregenom	q#
nucleophil	pcr	prematur	queuosin
nucleoplasm	peptid	preq	quorum
nucleosid	peptidyl	preq#	r#
nucleotid	pf#	presum	radiat
nucleus	pfam	pri	rarer
o	phase	primari	rat
occlud	phosphat	primer	rate
ofengand	phosphatas	prior	ratio
oh	phosphodiest	probe	ray
oligonucleotid	phosphoryl	process	reaction
ompf	phylogenet	produc	readthrough
oncogen	physiolog	product	rearrang
onto	picornavirus	program	receptor
oocyt	plant	project	recognit
operon	plasmid	prokaryot	reconstitut
opportunist	plus	prolifer	recruit

recycl
red
reduc
reduct
refer
region
regul
regulatori
relat
releas
remov
repeat
replic
replicas
replicon
repress
repressor
requir
residu
resist
resolut
respons
restrict
result
retain
retent
retrovir
retrovirus
revers
rf#
rfam
rhinovirus
rho
ribonucleas
ribonucleoprotein
ribos

ribosom
ribosos
riboswitch
ribozym
rice
rich
rightward
rna
rnai
rnaii
rnaiii
rnaprim
rnase
rnp
rodent
roll
rpl#
rpos
rpra
rps#
rrna
rsma
rsmb
rsmz
rybb
ryea
ryeb
ryhb
rz#
s#
saccharomyc
salmonella
sam
sarcoma
satellit
sativa

scaffold
scarna
scarna#
schizosaccharomyc
screen
secondari
secret
segment
select
self
sens
sensor
separ
sequenc
sequest
serotonin
shape
shigella
shock
short
shown
sigma
signal
silenc
silent
simian
similar
singl
singlet
site
size
sl
sl#
sm
small
snor#

snora#
snord#
snorna
snornp
snoz#
snr#
snrna
snrnp
sodb
spacer
speci
specif
spectrometr
splice
spliceosom
spread
srac
srna
srp
srp#
srpdb
ssu
stabil
stabilis
stack
stage
start
stationari
stem
step
stimul
stop
storag
strand
stress
stretch

stripe
structur
styphimurium
subgenom
substitut
substrat
subtili
subtyp
subunit
support
suppressor
surfac
surround
surviv
switch
symmetri
syndrom
synthas
synthes
synthesi
synthetas
system
tag
tail
tandem
target
tbr#
tcl#
tcv
telomer
telomeras
temperatur
templat
term
termin
termini

terminus
tertiari
thaliana
therapeut
thermodynam
thiamin
third
threeprim
threonin
tight
time
tissu
tomato
tombus
tombusvirida
tombusvirus
traj
tran
transcriptas
transduct
transesterif
transfer
transferas
transform
translat
transloc
transport
transposit
trigger
trna
trypanosoma
tumefacien
tumour
turnip

turnov
twoprim
tymovirus
type
typhimurium
u#
u45a
u45b
u4atac
u6atac
u83a
u83b
ubiquit
ugauga
uhg
um#
unpair
untransl
upregul
upsk
upstream
uptak
uridin
usual
utr
valin
variabl
variat
verifi
vertebr
via
viabil
vibrio
vii
viral
viroid

virolog
virul
virus
vitamin
vitro
vivo
voltag
wasserman
weak
wide
wound
xenopus
ydan
yeast
ykkc
yxkd
z#
zebrafish
zinc

References

[1] Svoboda P, Cara AD (2006) Hairpin RNA: a secondary structure of primary importance. Cell Mol Life Sci 63:901–8.

[2] Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, et al. (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–40.

[3] Mignone F, Grillo G, Licciulli F, Iacono M, Liuni S, et al. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 33:D141–6.

[4] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–4.

[5] Giegerich R, Voss B, Rehmsmeier M (2004) Abstract shapes of RNA. Nucleic Acids Res 32:4843–51.

[6] Schultes E, Hraber PT, LaBean TH (1997) Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. RNA 3:792–806.

[7] Leontis NB, Westhof E (2003) Analysis of RNA motifs. Curr Opin Struct Biol 13:300–8.

[8] Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38:221–43.

[9] Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, et al. (2006) The RNA Ontology Consortium: an open invitation to the RNA community. RNA 12:533–41.

[10] Chang KY, Tinoco IJ (1994) Characterization of a "kissing" hairpin complex derived from the human immunodeficiency virus genome. Proc Natl Acad Sci U S A 91:8705–9.

[11] Kim SH, Sussman JL (1976) pi turn is a conformational pattern in RNA loops and bends. Nature 260:645–6.

[12] George AD, Tenenbaum SA (2009) Informatic resources for identifying and annotating structural RNA motifs. Mol Biotechnol 41:180–93.

[13] Tinoco IJ, Bustamante C (1999) How RNA folds. J Mol Biol 293:271–81.

[14] Pavesi G, Mauri G, Stefani M, Pesole G (2004) RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Res 32:3258–69.

[15] Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, et al. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 29:4724–35.

[16] Liu C, Bai B, Skogerbo G, Cai L, Deng W, et al. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res 33:D112–5.

[17] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–9.

[18] Blaschke C, Valencia A (2002) Automatic ontology construction from the literature. Genome Inform 13:201–13.

[19] Wikipedia. Available: `http://en.wikipedia.org`.

[20] Manning CD, Schuetze H (1999) Foundations of Statistical Natural Language Processing. The MIT Press, 1 edition.

[21] Denoyer L, Gallinari P (2006) The wikipedia xml corpus. SIGIR Forum 40:64–69.

[22] Fox C (1990) A stop list for general text. SIGIR Forum 24:19–21.

[23] Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305.

[24] Reuters-21578, distribution 1.0. Available: `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`.

[25] Rabani M, Kertesz M, Segal E (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proc Natl Acad Sci U S A 105:14885–90.

[26] Freyhult E, Gardner PP, Moulton V (2005) A comparison of RNA folding measures. BMC Bioinformatics 6:241.

[27] Clote P, Ferre F, Kranakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11:578–591.

[28] Gusarov I, Nudler E (1999) The mechanism of intrinsic transcription termination. Mol Cell 3:495–504.

[29] Yarnell WS, Roberts JW (1999) Mechanism of intrinsic transcription termination and antitermination. Science 284:611–5.

[30] Samarsky DA, Fournier MJ, Singer RH, Bertrand E (1998) The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. EMBO J 17:3747–57.

[31] Muslimov IA, Iacoangeli A, Brosius J, Tiedge H (2006) Spatial codes in dendritic BC1 RNA. J Cell Biol 175:427–39.

[32] Xue C, Liu GP (2007) RScan: fast searching structural similarities for structured RNAs in large databases. BMC Genomics 8:257.

[33] Xue C, Li F, He T, Liu GP, Li Y, et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6:310.

[34] Yu H, Agichtein E (2003) Extracting synonymous gene and protein terms from biological literature. Bioinformatics 19 Suppl 1:i340–9.

[35] Hamada M, Tsuda K, Kudo T, Kin T, Asai K (2006) Mining frequent stem patterns from unaligned RNA sequences. Bioinformatics 22:2480–7.

[36] Roth A, Breaker RR (2009) The structural and functional diversity of metabolite-binding riboswitches. Annu Rev Biochem 78:305–34.

[37] Freyhult E, Moulton V, Clote P (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection. Bioinformatics 23:2054–62.

[38] Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–31.

[39] Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–9.

[40] Gene ontology website. Available: `http://www.geneontology.org/`.

[41] Porter MF (1997) An algorithm for suffix stripping. In: Readings in information retrieval, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 313–316.

[42] Snowball: A language for stemming algorithms. Available: `http://snowball.tartarus.org/index.php`.

[43] National center for biotechnology information. Available: `http://www.ncbi.nlm.nih.gov/`.

CHAPTER 6

CONCLUSIONS

Over the last several chapters, we presented examples of the organizing principles underlying RNA biology and revealed a common theme of modularity, in which the mechanisms and structures of RNAs can be understood by decomposing them into building-block units of function.

We first showed that certain classes of natural RNA structures, particularly precursor microRNAs, possess an intrinsic robustness that allows them to maintain a specific shape regardless of their sequence context. This form of structural modularity reflects the requirements of specific biogenerative processes as well as a mode of neo-functionalization in which copies of highly modular RNAs can retain shape specificity in new contexts.

Next, we illustrated modular functionalization in the context of rat dendritic transcript localization by characterizing the role of the ID element, a retrotransposon that can be co-opted to serve as a localization motif when associated with regulated intron sequence retention. We also showed that a similar role may be associated with Alu elements, which have high similarity to the previously-characterized *Camk2a* dendritic localization element. Together these observations show the potential for ubiquitously occurring transposable sequence to become functional RNA modules in a novel con-

text for regulatory modules, retained introns.

Finally, we undertook a low-level functional characterization of all RNA families using automated techniques and revealed the existence of common units of RNA function that link together diverse RNA families. We associated a large number of these basic functions with small structural motifs, and in this way highlighted the fundamental RNA structure-function relationship at a higher resolution than commonly considered. We hypothesize that these structure-function units represent a subset of the elemental building blocks that can combine in different ways to form the diversity of RNA species.

One of the hallmarks of modular evolution reflected in these findings is the accelerated rate at which innovation can appear. ID elements exist in high copy number only in the rat genome, and we found no instances of conserved ID elements between orthologous dendritic rat and mouse genes. If ID elements now play a role in dendritic localization, then this represents either novel functionality in the rat lineage or replacement of more ancestral functionality with a new mechanism. Similarly, the miRNA complement of closely related species varies [1], due to expansions in specific lineages, many times resulting from gene duplications that preserve the mature sequence [2]. Although a large number of these duplications seem to have arisen from ancestral large-scale genome duplications [3], there is evidence that localized and evolutionarily recent duplications can occur [4]. In both of these examples, modular properties allow rapid functionalization, bypassing the need to evolve RNA features *de novo*.

Our results also highlight the existence of organizing centers for modular RNAs. Polycistronic primary miRNA genes presumably exist to facilitate coordinated expression; however, it is not the case that all miRNAs in a cluster are always expressed at the same level [5], suggesting that the pri-miRNA gene provides avenues for individ-

ual regulation, perhaps via the miRNA loop sequence [6]. It is still poorly understood why miRNA clusters contain the miRNAs that they do, and why many other miRNAs are not clustered. If there is in fact selection for particular miRNA gene architectures, suggesting that the primary transcript constitutes a hierarchical functional module, then the downstream effects may also reflect a modular regulatory program. Detailed comparisons of expression data for both miRNA genes and target transcripts, as well as the pattern of miRNA binding sites in the targets, could uncover some of this underlying structure.

In the case of introns, an appealing hypothesis emerges in which regulatory modules can exist in non-exonic regions of protein-coding transcripts, affecting not the protein message but rather the manner in which it is produced. The key to this phenomenon is the retention of intronic sequence, which determines whether these regulatory modules are present when the transcript is exported from the nucleus. Functionalization, then, would seem to require at least two discrete, though not necessarily coincidental evolutionary steps. We propose that intron retention serves as an on/off switch for the proto-regulatory mechanisms contained in intronic sequence, and is thus the second of the two steps. Meanwhile, introns can accumulate sequence, through the action of transposable elements, some of which will have the potential to serve in functional roles. As long as an intron is spliced out, the elements contained within it will have little or no effect on phenotype; but when random change causes intron retention, the proto-elements can become "activated" and potentially affect the fitness of the individual. As such, introns may serve as sandboxes for evolutionary innovation, and it remains to be seen whether it is possible to identify additional evidence in favor of this hypothesis. One approach to address this question would be the creation of a comprehensive catalog of the transposable elements that appear in the introns of related species, such as rat and mouse, and an enumeration of the

instances where the repeat content differs. There might turn out to be one or several ID element-analogs in the mouse genome that may have a role in dendritic targeting, or some other regulatory mechanism.

In Chapter 5 we framed our analysis in terms of characterizing abstract functional modules, an approach that shifts the emphasis away from *a priori* assumptions about the physical form of a module. In formulating the self-containment property, we defined structural robustness to be robust maintenance of a static shape, but it is possible that other physical manifestations of functional modularity are relevant. For example, the salient feature of riboswitches is their ability to adopt alternate conformations after binding a metabolite [7], suggesting that some degree of structural plasticity is consistent with the definition of a riboswitch module. In fact, riboswitches do exhibit bistable conformations [8], so a reformulation of self containment to handle characteristic conformation changes and interactions with context might be able to detect modules of this form. Similarly, extraction of low-level structure representations, used in Chapter 5 and RScan [9], may benefit from enumerating the motifs of multiple stable structures that a sequence can adopt. In general, the concept of a single minimum-free energy secondary structure, while convenient, is not always an accurate characterization of an RNA *in vivo*. Richer models, in the form of probabilistic representations of the ensemble of possible structures, may lead to a broader understanding of the ways in which functional modularity can be attained in RNA structures.

To abstract even further from specific forms, we might look for higher-order structural tendencies associated with modular RNA structures. Graph-theoretic definitions of modularity (e.g., Bonner's gene net [10]) describe a high degree of connectivity between units within a module compared to a low number of extra-modular connections. Translated into RNA structure, this implies that the number of base-pairing

relationships within a highly-self contained RNA should exceed the potential number involving nucleotides outside the sequence. From an energetic standpoint, this must be true to some extent, yet base-pairing proportion is not strongly correlated with self containment (Table 3.6). Perhaps there is a subtler pattern of nucleotide composition and sequence associated with modularity, which we might be able to discover given a large number of examples of highly self-contained sequences, which could be artificially generated. The existence of geometric properties such as regularities or symmetries in RNA structures could manifest themselves more generally in other modular architecture.

These various lines of investigation all funnel into a general "RNomics" research plan [11]. As we come to recognize the prevalence of modularities in RNA biology, we can begin to construct RNA discovery pipelines in which putative novel RNA species are evaluated in an evolutionary context beyond nucleotide or shape conservation. We have already shown the efficacy of using self containment to distinguish RNAs with modular characteristics, and in fact, the self-containment index has already been used for *de novo* miRNA discovery [12]. However, what additional information can we leverage from genomic context, repeat structure, or geometry? To what extent is it possible to define the language of RNA structures in terms of intermediate-level elemental units that are above the level of the functionally-ambiguous structure motif, but more general than an RNA family-specific domain? A heightened understanding of the common evolutionary histories of RNAs – marked by insertions, duplications, shuffling, recombination, modification – may bring us closer to the goal of identification and characterization off the cellular RNA repertoire.

In achieving this goal, we will need to continue to refine and develop techniques for the accurate representation and quantification of RNAs – the analysis of short-read sequencing data, particularly with respect to read alignment to high-copy number

genomic loci; the full realization of a comprehensive functional annotation of RNAs; the formulation of functionally-motivated metrics for RNA structure distance.

RNA is organized. The nuances of how or why remain unclear, but in light of the last fifty years of fruitful RNA research, it seems likely that we can come to understand the big picture, one building block at a time.

References

[1] Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, et al. (2006) Diversity of microRNAs in human and chimpanzee brain. Nat Genet 38:1375–7.

[2] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–4.

[3] Gu X, Su Z, Huang Y (2009) Simultaneous expansions of microRNAs and protein-coding genes by gene/genome duplications in early vertebrates. J Exp Zoolog B Mol Dev Evol 312B:164–70.

[4] Heimberg AM, Sempere LF, Moy VN, Donoghue PCJ, Peterson KJ (2008) MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A 105:2946–50.

[5] Yu J, Wang F, Yang GH, Wang FL, Ma YN, et al. (2006) Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. Biochem Biophys Res Commun 349:59–68.

[6] Michlewski G, Guil S, Semple CA, Caceres JF (2008) Posttranscriptional regulation of miRNAs harboring conserved terminal loops. Mol Cell 32:383–93.

[7] Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, et al. (2002) Genetic control by a metabolite binding mRNA. Chem Biol 9:1043.

[8] Freyhult E, Moulton V, Clote P (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection. Bioinformatics 23:2054–62.

[9] Xue C, Liu GP (2007) RScan: fast searching structural similarities for structured RNAs in large databases. BMC Genomics 8:257.

[10] Bonner J (1988) The evolution of complexity by means of natural selection. Princeton Univ Pr.

[11] Huttenhofer A, Brosius J, Bachellerie JP (2002) RNomics: identification and function of small, non-messenger RNAs. Curr Opin Chem Biol 6:835–43.

[12] Lazzari B, Caprera A, Cestaro A, Merelli I, Corvo MD, et al. (2009) Ontology-oriented retrieval of putative microRNAs in Vitis vinifera via GrapeMiRNA: a web database of de novo predicted grape microRNAs. BMC Plant Biol 9:82.