9 Glimpses of a Modern RNA World

The discovery of abundant small RNAs with functions beyond translation and the recognition of their target specificity by base-pairing prompted exploration of the regulatory potential of 'antisense' molecules.[1] In the late 1970s, Paul Zamecnik (whose group discovered tRNA, Chapter 3) and colleagues demonstrated that binding of short synthetic oligonucleotides to complementary sequences in Rous sarcoma virus and human T-cell lymphotropic virus blocked the replication and translation of viral RNA and the oncogenic transformation of cells.[2,3]

Such findings suggested that short antisense RNAs might exist naturally in cells, not as common species involved in core processes but as (individually rarer) regulators of specific genes or transcripts, but difficult to identify and characterize by the analytical techniques of the time.

RIBOREGULATORS

The clues were already there. Studies in the 1960s had revealed a number of small RNAs (sRNAs)[a] of unknown function in bacteria.[4,8,9] The existence of 'antisense' RNAs and 'bidirectional transcription' was first reported in 1972 in phage lambda, where it was proposed to control expression of the lambda repressor.[10] The subsequent sequencing of the genomes of bacteriophages and eukaryotic viruses showed that the occurrence of overlapping genes and transcripts was a general phenomenon.[11–14]

Conserved bacterial RNAs such as the 10Sa[b] and 10Sb[c] RNAs,[20,21] as well as regulatory motifs and structures, had also been reported around that time, for example, in feedback mechanisms controlling rRNA and ribosomal protein levels, akin to RNA structure-dependent regulatory mechanisms identified in bacteriophages.[22–26]

In 1975, Stuart Heywood and colleagues demonstrated that short RNA sequences from chicken muscle ribonucleoprotein fractions could control translation of the mRNA encoding myosin. They called the RNAs "translation control RNAs" (tcRNAs) and proposed that tcRNAs act by binding their mRNA targets in a sequence-specific manner.[27,28] In follow-up work a decade later, Heywood demonstrated that one of these tcRNAs, tcRNA102, recognizes a sequence in the 5′UTR of the myosin mRNA.[29,30]

In the 1980s, a number of studies identified bacterial plasmid-encoded small 'untranslatable' antisense RNAs that formed stable secondary structures and regulated plasmid replication, plasmid incompatibility, transposition and translation, among others.[31–35] For example, mutation analysis revealed that the ~108nt antisense transcript 'RNA I' blocked replication of the ColE1 plasmid (Chapter 6) by base pairing with the RNA that forms the replication primer[31] – one of the first regulatory roles demonstrated for any RNA (see Chapter 8). Soon after, a ~70nt RNA transcribed from a promoter of the Tn10 transposon was shown to repress transposition by preventing translation of the transposase mRNA, representing the first example of transposon regulation by antisense RNAs.[34] A ~170nt antisense RNA (micRNA) expressed in *E. coli* was found to inhibit translation of OmpF mRNA, which encodes a major outer membrane protein.[35] Packaging of phage DNA during infection was found to be directed by the phage-encoded ~120nt phi29 RNA, as part of the DNA-packaging machine.[36]

Before the end of the decade, enough examples had accumulated to allow generalizations around the theme of antisense RNA control of gene expression and the potential of fine-tuning interactions in a way not readily achieved by proteins.[37] Masayori Inouye speculated at the time that this "regulatory system may be a general regulatory phenomenon in *E. coli* and in other organisms, including eukaryotes"[38] and that "RNA species may have additional roles in the regulation of various cellular activities".[35]

[a] One of these sRNAs was discovered in 1967,[4] 9 months after the identification of the *lac* repressor. It was initially dubbed 6S (also known as SsrS) and later shown to be a structurally conserved molecule that regulates RNA polymerase promoter use.[5–7] One can only speculate what the impact on the conceptual framework of RNA and protein function in molecular biology might have been if this had come to light earlier.

[b] 10Sa RNA is also known as tmRNA (a 'transfer-messenger RNA' with properties of a tRNA and an mRNA) or SsrA.[15–17] It was later shown to play a key role in the symbiosis between the bacterium *Vibrio fischeri* and squid[18] (Chapter 12).

[c] 10Sb RNA was later found to be the bacterial homolog of the RNase RMP ribozyme[19] (Chapter 8).

DOI: 10.1201/9781003109242-9

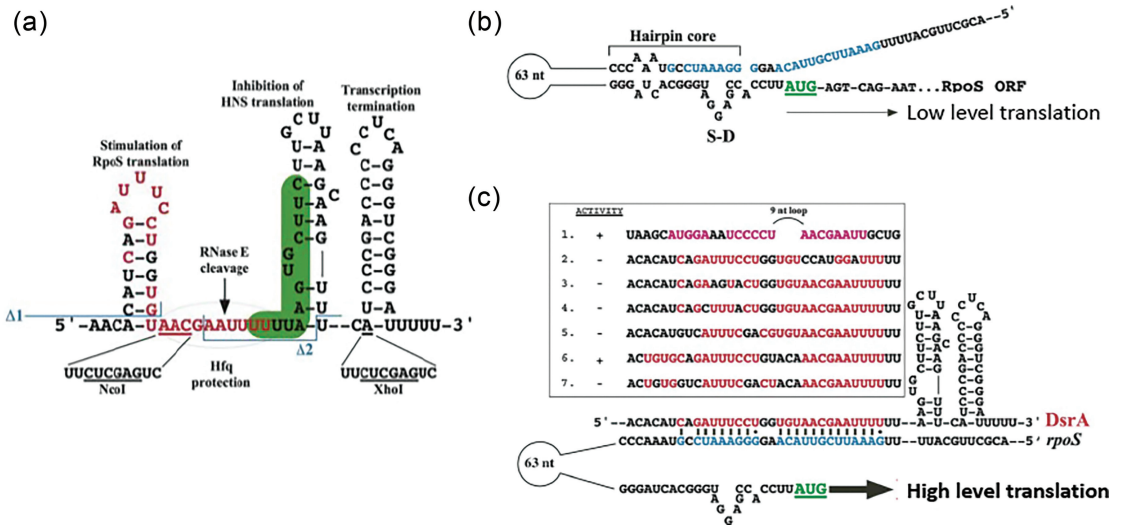

FIGURE 9.1 Structure (a) and mode of action of DsrA in controlling RpoS translation. The Shine-Dalgarno (S-D) ribosome binding sequence and the AUG start codon are sequestered by the 5' sequence of the RpoS mRNA (b) but released by DsrA binding (c). (Reproduced from Majdalani et al.[59])

In subsequent decades, it was shown that sRNAs regulate many bacterial processes,[d] including virulence, quorum (community) sensing, symbiosis, stress responses, the physiological transition from growth to stationary phase, other aspects of metabolism and environmental responses, bacteriophage packaging, DNA exchange, transcription and translation, among others.[18,38,42,43] Thousands of bacterial sRNAs that regulate gene expression at both transcriptional and post-transcriptional levels[42] have now been described,[44–46] aided by new high-throughput RNA-protein interaction technologies.[47–50]

The following examples are illustrative. The *E. coli* DsrA RNA (Figure 9.1), which is induced at low temperatures, inhibits transcriptional silencing by the nucleoid-associated H-NS protein and stimulates translation of the stress sigma factor RpoS, both depending on association with the RNA-binding protein Hfq.[51–53] Another ~109nt sRNA, oxyS, was found to repress translation of RpoS by interacting with Hfq and altering its activity, acting as a global regulator to activate or repress the expression of approximately 40 genes involved in stress responses.[54,55] In fact, many sRNAs, such as the Spot 42 sRNA that regulates the galactose operon (Chapter 3), also require Hfq[e] for their stability and function.[57,58]

The common involvement of Hfq, which acts as a general cofactor for stabilizing small antisense RNAs, facilitating RNA-RNA interactions and gene expression control in many bacteria,[57,60–64] including the regulation of utilization of the intestinal metabolite ethanolamine by a Hfq-dependent sRNA,[65] indicated the existence of broader RNA-regulated networks.[42,60] This was also an early example of the use of a generic protein infrastructure to execute RNA-directed regulatory events, a theme that would later be writ large in eukaryotes (Chapters 12 and 16).

Other RNAs that control global processes in bacteria were also discovered, such as the inducible CsrB and CsrC RNAs of *E. coli*, which bind (via conserved sequences and hairpin structures) and inhibit the RNA-binding protein CsrA, a translational regulator, by outcompeting mRNA targets.[59,66] Homologs of this system have been implicated in the regulation

[d] As explained by Kai Papenfort and Jörg Vogel: "The late appreciation of regulatory RNA might be attributed to the fact that loci encoding such regulators were rarely selected in genetic screens for virulence factors, likely owing to a usually smaller gene size, missing annotations in genome sequences, and typically subtle phenotypes, as compared to virulence-associated proteins."[39] For example, the ~514nt RNAIII was originally described as the δ-hemolysin mRNA of *Staphylococcus aureus* but subsequent molecular analysis revealed that, in addition to expressing hemolysin from its 5′ region, RNAIII acts as an antisense regulator of virulence and surface protein synthesis through its 3′ region, a dual coding and regulatory RNA.[40,41]

[e] Hfq was originally described in 1968 as an *E. coli* host factor required for the synthesis of bacteriophage Qβ RNA and the replication of the bacteriophage Qβ RNA genome.[56]

of gluconeogenesis, biofilm formation and virulence factor expression in a variety of bacterial pathogens,[59] and represent some of the first examples of mimicry, protein-sequestration or sponging by regulatory RNAs at a post-transcriptional level.

In 2020, the late promoter of the Shiga toxin-encoding bacteriophage in enterohemorrhagic *E. coli* was found to produce an abundant regulatory RNA to silence the expression of the toxin during lysogeny (the Shiga toxins cause renal failure and neurological damage), which had been hiding in plain sight despite decades of research on Shiga toxin.[67]

Synthetic riboregulators have been constructed for eukaryotic translational control.[68] A common principle underlying the functions of these small regulatory RNAs is the ability to combine secondary structures that can bind proteins or small ligands (as exemplified by some ribozymes[69,70] and SELEX) with exposed nucleotide stretches that can recognize other RNAs or DNA in a sequence-specific manner.

RIBOSWITCHES

In 2002 and following years, Ron Breaker, Wade Winkler, Alexander Mironov, Evgeny Nudler and others showed that the ability of RNA to sense ligands, previously thought to be the sole province of proteins, is widely used by bacteria to connect the regulation of transcription and translation to metabolic and environmental signals, including thiamin (vitamin B1), riboflavin-5′-phosphate (vitamin B2), biotin (vitamin B7), cobalamin (vitamin B12), fluoride, various amino acids, S-adenosyl methionine (SAM) and glucosamine-6-phosphate, among many others, and even temperature (RNA 'thermometers').[71–77] These RNA ligand-sensing modules have become known as 'riboswitches',[78–81] with more evident in genomic analyses,[82] and high-resolution studies revealing the molecular dynamics involved.[83]

For example, the SAM riboswitch (the 'S-box leader') is a highly conserved RNA domain that responds to the coenzyme SAM with high affinity and specificity. In *Bacillus subtilis*, it occurs in the 5' region of dozens of genes encoding proteins involved in methionine or cysteine biosynthesis, where it allosterically regulates their expression at the level of transcription termination. When SAM is unbound to the RNA aptamer, the anti-terminator sequence sequesters the terminator, which is then unable to form, whereas when SAM is bound, the anti-terminator is sequestered and transcription is terminated (Figure 9.2).[73,84,85]

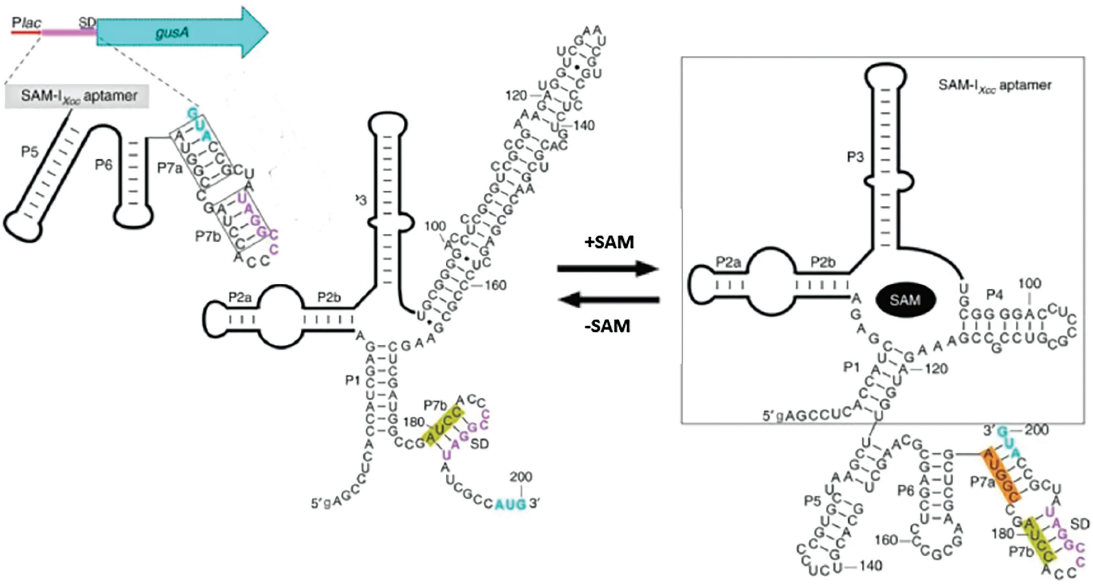

FIGURE 9.2 Structure and function of the S-adenosyl methionine (SAM) riboswitch in the 5' untranslated region of the mRNA in the polycistronic *met* operon of *Xanthomonas campestris*, which encodes three enzymes for the biosynthesis of methionine, replaced here by the reporter gene *gusA*. Binding of SAM to the RNA aptamer in the riboswitch (boxed) causes an allosteric structural rearrangement that sequesters the Shine-Dalgarno sequence (purple) and AUG start codon (cyan) to inhibit translation. (Adapted from Tang et al.[85] under license from Creative Commons.)

Thiamin riboswitches also occur in plants, fungi and protists,[86–88] and shown, *inter alia*, to regulate RNA splicing.[89] Riboswitches likely also exist in animals, although their repertoire is not so well explored, possibly because of the difficulty of their characterization in complex organisms. Artificial riboswitches have been constructed to respond to pH[90] and light[91] and to control RNA splicing.[92]

Ligand-induced allosteric changes in RNA structure are similar to those observed in proteins upon binding of small molecules, such as nucleotides (ATP, AMP, GTP, etc.) to sense energy status or transduce extracellular signals, and even the *lac* repressor's recognition of lactose, which causes a conformational change in the repressor so that it can no longer bind DNA to block transcription of lactose metabolizing enzymes.

Riboswitches may have predated proteins and have been suggested to be the oldest mechanism for the regulation of gene expression.[93] As Breaker speculated: "The characteristics of some riboswitches suggest they could be modern descendants of an ancient sensory and regulatory system that likely functioned before the emergence of enzymes and genetic factors made of protein."[94]

ANTISENSE RNAs AND COMPLEX TRANSCRIPTION IN EUKARYOTES

The first evidence of regulatory antisense RNAs in eukaryotes was obtained in 1987, also by Zamecnik's group, who reported endogenous small (<30nt) RNA oligonucleotides in mammalian cells using radioactive labeling, proposing that they "may play a regulatory role in intracellular metabolism and may conceivably travel from one cell to another in a similar role".[95] It was another 13 years before the ubiquity and power of such regulatory 'microRNAs' in eukaryotes started to be revealed and appreciated (Chapter 12).

Nonetheless, based on the principles of the activity of antisense RNAs in bacteria, as well as experiments by Zamenick's group with exogenous antisense oligonucleotides, "anti-message" RNAs began to be used from 1984 as a tool for suppressing the expression of specific genes[f] in eukaryotes, including globin, at the level of transcription, translation and/or RNA stability[1,99–102] before the discovery of natural antisense transcripts in eukaryotes.[103,104]

The use of synthetic antisense oligonucleotides was quickly adopted and is still widely employed to study gene function in a wide range of eukaryotes, including frogs, insects, plants and mammalian cells, as specificity of inhibition is easy to achieve independently of any knowledge of the function of the gene under investigation.[103]

Antisense oligonucleotides also form a vital component of the toolkits for genetic engineering and gene therapy,[g] aided by artificial chemistries, such as peptide or phosphorothioate linkages and methylene bridges ('locked nucleic acids'), to increase the target affinity and half-life of the molecules *in vivo*.[107–111] The use of synthetic nucleic acids has since been given new impetus by the discovery of another natural antisense RNA regulatory pathway, the small RNA-guided 'CRISPR' systems that have revolutionized genetic engineering (Chapter 12).

At that time, however, despite the emerging examples of regulatory sRNAs in bacteria and the ability of antisense molecules to artificially modulate gene expression in eukaryotes, "the extent to which this novel form of regulation of gene expression is utilized in prokaryotes and eukaryotes … [remains] … to be established".[112]

The first discoveries of natural regulatory RNAs in eukaryotic cells were serendipitous[h] – a pattern repeated over the next ~15 years, until the genome projects revealed the full extent of RNA expression (Chapter 13). Nevertheless, an unexpected by-product of genetic screens and conventional gene cloning and mapping approaches were many early observations that hinted

[f] Antisense interactions between cDNAs and mRNAs ('hybrid-arrested translation') was developed in the late 1970s for gene mapping and identification.[96–98]

[g] The first RNA therapeutics company, Isis, now Ionis, was established in 1989 by Stanley Crooke.[105] As of 2021, eight antisense oligonucleotide drugs had been approved for commercial use.[106]

[h] One of the relevant discoveries during this period was the HIV TAR (trans-activating response element), an RNA stem-loop structure located at the 5′ ends of nascent HIV-1 transcripts, which was proposed to be a "novel type of regulatory element" for transcriptional activation.[113–115] In addition to the viral regulatory RNAs mentioned in the previous chapter, several other non-coding RNAs were subsequently characterized from DNA and RNA viruses that infect eukaryotic cells. These include highly abundant small RNAs and lncRNAs discovered in the late 1970s and 1980s, such as the EBV-encoded RNAs found to have different roles including recruitment of transcription factors to control expression[116] (Chapter 16) and the 2.7 kb repeat-derived RNA that comprises ~20% of the early transcription from the human cytomegalovirus Beta2.7 gene[117] and whose function only started to be identified decades later.[118]

at the existence of longer (>200nt) non-protein-coding RNAs in eukaryotes.

These early studies also revealed the existence of 'nested' genes and non-coding transcripts in intensely studied genomic regions, such as developmental and cancer-related loci, as well as in studies using differential cDNA cloning and hybridization strategies[i] to identify transcripts from genes that are active or repressed in specific tissues and/or developmental stages.

In 1986, Steven Henikoff and colleagues reported the first case of a "gene within a gene" in *Drosophila*, showing that a pupal cuticle protein is encoded within the intron of an unrelated gene, on the opposite strand and independently expressed. They described their findings as "an unambiguous exception to the classical linear model of gene organization" and, considering the possible commonality of genes nested within large introns and extended loci, remarked that "it is interesting to consider the genetic complexity that could result".[120]

In the same year, Trevor Williams and Mike Fried showed that a region of the mouse genome encodes two RNAs that are transcribed in opposite direction and overlap at their 3' ends, contemplating the implications in the light of the findings of experimentally introduced antisense RNAs inhibiting gene activity.[121]

Similarly, in the same issue of *Nature*, Charlotte Spencer and colleagues reported that a transcript of unknown function overlaps that of the dopa decarboxylase (*Ddc*) gene on the opposite strand in *Drosophila*.[122] Given that the transcripts showed differences in temporal and spatial expression, they proposed that the antisense transcript could have regulatory function based on either RNA-RNA base pairing or via transcriptional interference and that "such arrangements in eukaryotes may be more common than previously supposed",[122,123] a prediction that was confirmed 20 years later when high-throughput transcriptome analyses were undertaken in the wake of the genome projects[124–126] (Chapter 13).

Also in 1986, Alain Nepveu and Kenneth Marcu showed that the protein-coding and opposite complementary strands of the *c-Myc* locus in mice are transcribed and regulated independently,[127] confirmed the following year by Gail Sonenshein and colleagues,[128] suggesting a role of the antisense RNAs in c-Myc processing or transcriptional interference.[127] The *TP53* tumor suppressor locus was shown in 1989 to also express a long antisense RNA, 'inRNA', speculated to be involved in the maturation of p53 mRNAs.[129]

Other examples in different organisms followed. An antisense RNA expressed in the silk moth *Bombyx mori* was found to display extensive complementarity to the chorion gene *Hcb.12* and to be co-expressed in follicular cells during development.[130] An RNA antisense (aHIF) to the 3'UTR of the human Hypoxia-inducible factor-1 alpha (HIF1α) mRNA was found to be co-expressed in cancer and hypoxia.[131] The expression of other antisense RNAs was found to negatively correlate with that of their complementary protein-coding mRNAs, such as the intronic antisense RNA from the human *eIF2A* locus,[132] transcripts antisense to the chicken alpha-I collagen gene[133] and transcripts antisense to the *EB4* locus in the slime mold *Dictyostelium*, whose functional analysis suggested a role in regulating the stability of EB4 mRNA.[134]

In some cases, such as the tcRNA identified by Heywood, antisense RNAs had only limited sequence complementarity to their targets, suggesting the potential existence of 'trans-acting' RNAs that originated from different loci.[30,135] Other cases involved large regions of overlap between antisense RNA and mRNA, sometimes spanning most of the length of the transcription units.[136]

A large number of similar but disparate observations followed at the end of the decade, including demonstrations of other nested genes and sense-antisense pairs in plants, insects, birds and mammals,[136–142] including in well-studied loci such as the globin clusters, which also showed evidence of transcription of non-coding regions encompassing enhancers and *cis*-regulatory elements,[143,144] as did loci exhibiting parental imprinting (see below).

Some of these early reports considered the conceptual and practical implications of the interleaved organization of genes and transcripts, challenging orthodox concepts, especially that the exons of protein-coding transcripts are the only biologically relevant portions of genes. As put by Adelman and colleagues in 1987:

> this situation may be a significant form of molecular evolution. By using both strands of the same DNA, the information content (regulatory and/or structural) of a particular genetic segment becomes amplified, adding a new complexity to the concept of a eukaryotic gene.[145]

[i] Subtractive cDNA hybridization and differential display.[119]

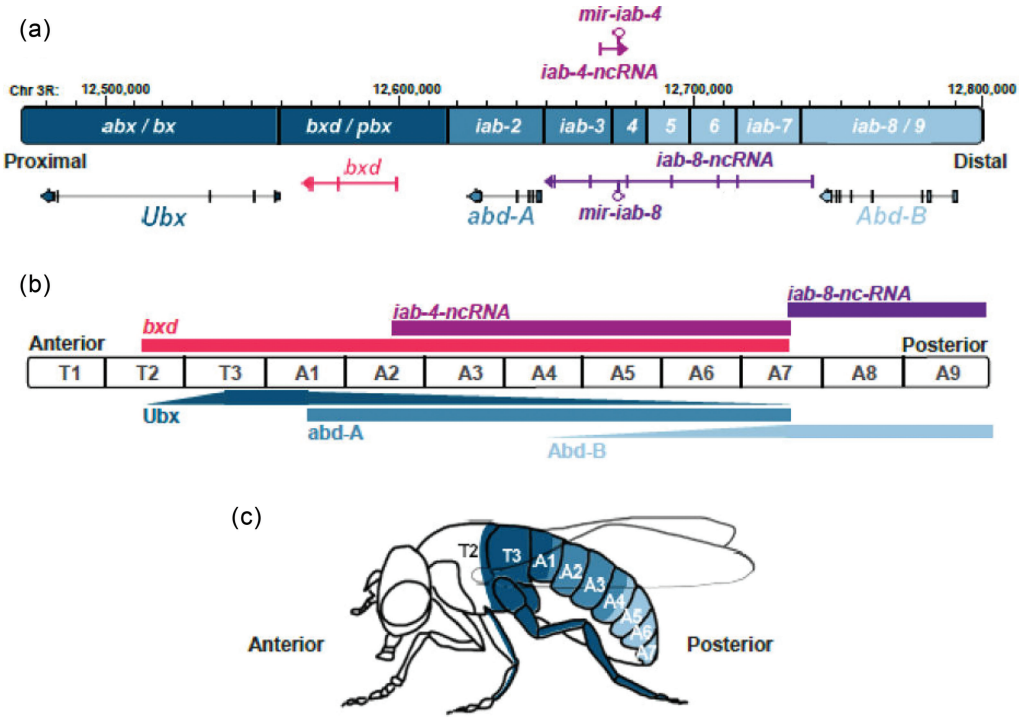

FIGURE 9.3 Map of the *bithorax* complex in *Drosophila*, showing the coding and non-coding transcripts produced from each locus – the protein-coding genes *Ubx*, *abd*-A and *Abd-b*, and the regulatory genes *bxd*, *iab-4* and *iab-8* – and their expression in different segments of the fly. The non-coding RNAs expressed from *iab-4* to *iab-8* are also regulated by microRNAs (Chapter 12). (Reproduced from Garaulet and Lai[150] with permission from Elsevier.)

LONG UNTRANSLATED RNAs

In addition to antisense RNAs, a number of other "unconventional" RNAs were found to be transcribed from 'intergenic'[j] regions of eukaryotic genomes and associated with genetic effects, notably in the well-studied regulatory regions of the *bithorax* complex studied by Lewis in *Drosophila*.[146]

As noted in Chapter 5, David Hogness, Michael Akam and colleagues discovered in 1985 that only one of the five mapped mutations ('pseudoalleles') in the *bithorax* complex corresponded to a protein-coding gene (*Ultrabithorax* or *Ubx*), while the others are derived from a much larger region containing regulatory elements. The pseudoallelic mutations were located in introns or in the upstream *bxd* region: the latter was found to be transcribed into a ~27 kb RNA that has a number of large introns and is subjected to differential splicing to produce various smaller (~1.2 kb) polyadenylated non-coding RNAs, none of which has protein-coding potential.[147,148] The expression of these transcripts was also shown to be highly regulated during embryogenesis, in a pattern that is partially reflective of *Ubx*.[147,149]

Moreover, while the extended *BX-C* cluster contained three protein-coding genes (*Ubx*, *abd-A* and *Abd-B*), it produced at least seven distinct RNAs that are co-linearly transcribed during development[149,151] (Figure 9.3). Other non-protein-coding transcripts were also reported in the nearby *iab-4* locus.[150,152] As their relevance was unclear, it was mooted that these transcripts are "functionless"[149] or "might function in cis by some unprecedented mechanism".[151] Some suggested that transcription of these loci is a passive by-product of the recruitment of transcription factors to enhancer sites that act distally by looping to contact promoters of protein-coding genes, or that it is simply the act of transcription (not the transcribed RNA) that is relevant by remodeling and/or exposing chromatin and underlying DNA sequences to transcription factors.[153,154] Many (presumed) cis-regulatory elements encompassing different developmental 'enhancers' and 'response elements' for the epigenetic regulators

[j] That is, transcribed from genomic sequences between annotated protein-coding genes, a description reflecting the deep bias that genes = proteins.

Polycomb and Trithorax[155,156] have been characterized in the loci that express these RNAs,[157] but their mode of action is only now being elucidated, especially in the context of RNA-directed chromatin modifications that control the expression of the clusters (Chapters 14 and 16).

Another early example is the *93D* locus, one of the largest of the genomic regions in *Drosophila* that 'puff' (i.e., become transcriptionally active) after heat shock. In the early 1980s, the groups of Subhash Lakhotia and Mary Lou Pardue found that the *93D* locus does not specify a protein, but rather a set of rapidly evolving non-coding transcripts (although an intron is highly conserved[158]), known as *hsr omega* RNAs: long repeat-containing transcripts with at least three different isoforms of ~1–10 kb that are differentially expressed in different tissues and developmental stages.[159–164]

Transcription of repeat-containing spliced and polyadenylated long non-coding RNAs was also reported in loci involved in immunoglobulin class switching and recombination in the 1980s, with evidence that such transcripts are associated with 'enhancer' action and alteration of chromatin architecture, including V(D)J recombination at antigen receptor loci,[165–168] whose roles in these processes are now being understood.[169–172]

Many long 'nontranslatable' antisense, sense and intergenic RNAs from higher eukaryotes were also cloned during the 1980s and 1990s. One of the first was an interspersed maternally derived 3.7 kb non-coding RNA called ISp1, characterized by Britten and Davidson's group in 1988, studying the sea urchin *Strongylocentrotus purpuratus*. This transcript, whose function is unknown, is polyadenylated and apparently processed into shorter (400–600nt) RNAs stored in the cytoplasm of sea urchin eggs.[173]

Large 'transcription units' of unknown function that appeared to lack protein-coding potential were being reported in humans as early as 1985, notably from the *PVT-1* (plasmacytoma variant translocation) locus that activates expression of the *MYC* oncogene and is heavily implicated in cancers through amplifications, retroviral insertions and translocations. The *PVT-1* locus spans at least 200 kb and expresses large multi-exonic non-coding RNAs that initiate 57 kb downstream of *MYC*.[174–181]

Other cancer-associated long non-coding RNAs (lncRNAs) identified in following years, such as BIC,[182–185] TP53TG1 ('TP53 target gene 1')[186,187] and DD3 (Differential Display Code 3, also known as PCA3, prostate cancer antigen 3),[188,189] turned out, like PVT1,[190] to be, at least in part, microRNA precursors (Chapter 12), as did, for example, the developmentally regulated and highly conserved 7H4 RNAs, found by subtractive hybridization to be highly enriched in synaptic nuclei of rat skeletal neuromuscular junctions.[191,192]

However, the first mammalian long non-coding RNA to be well recognized was H19. It was cloned in 1990 by Shirley Tilghman and colleagues by differential hybridization, and corresponded to a transcript that was originally identified by the same group in 1984 as an abundant RNA in a screen of a mouse fetal liver cDNA library (named after the cDNA clone designated pH19).[193] It was also found to be expressed in rat skeletal muscle (where it was called ASM).[194] H19 is transcribed by RNA polymerase II, spliced and polyadenylated, with a number of short open reading frames that are not conserved between mouse and human.[k] It also did not seem to be associated with ribosomes. *H19* was, therefore, proposed to represent an "unusual gene" whose product differs from a "classical mRNA" in that it may act as an RNA, not an intermediate to a protein.[193] It was subsequently found to be part of an imprinted locus that encompasses the *Igf2* (insulin-like growth factor) gene, to have tumor suppressor capacity[196–201] and to be lethal when ectopically expressed.[202]

In the following year, Carolyn Brown, Hunt Willard and colleagues cloned an RNA expressed exclusively from the 'X-inactivation center' in the inactive X chromosome in females. This alternatively spliced transcript was a candidate for the factor responsible for dosage compensation and was named Xist ('X-inactive specific transcript').[203] They found no evidence of an encoded protein, and characterized human XIST[l] as a 17-kb RNA containing conserved tandem repeats, localized within the nucleus and coating the inactive X chromosome in female cells in a position "indistinguishable from the X inactivation-associated Barr body", leading them to propose that Xist functions as a "structural RNA".[204]

Xist functions partially by recruiting chromatin-repressive complexes to promote heterochromatin formation and transcriptional silencing of the chromosome[205–209] (Figure 9.4), although how it selects

[k] Later proteomic analysis indicated that H19 produces a short protein in humans but not mice,[195] suggesting that it was either gained or lost in one or other lineage.

[l] The gene naming convention uses all capitals for human genes, and only the first letter in uppercase (e.g., *Xist*) in mouse.

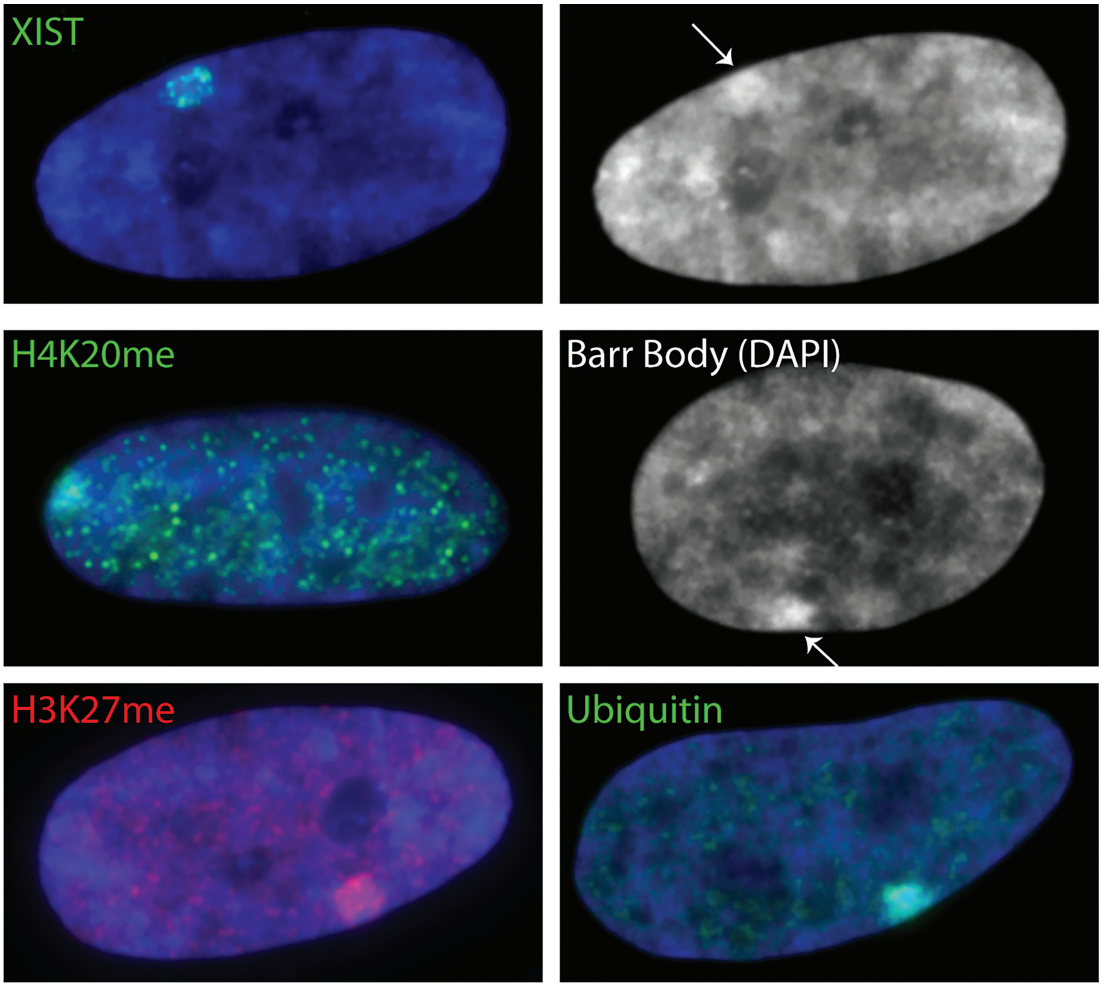

FIGURE 9.4 Localization of Xist on the transcriptionally inactive condensed X chromosome in interphase female human cells (the Barr body, with intense DAPI DNA staining), which is accompanied by repressive histone modifications such as methylation of H4K20 and H3K27, and H2AK119 ubiquitination (Chapter 14). (Image courtesy of Jeanne Lawrence,[205,206] UMass Chan School of Medicine).

just one of the two X-chromosomes[m] is not fully understood. While thought of as a special case at the time, Xist has become emblematic of the extraordinary complexity of non-coding RNA control of chromatin architecture including the nucleation of phase-separated domains (Chapter 16). It emerged later that the *Xist* locus originated by the fusion of a 'pseudogenized' protein-coding gene with a set of transposable elements that are essential to its function.[212,213] And it does explain how the heterochromatic Barr body described by Ohno (Chapters 4 and 7) is formed.

Analogous non-coding RNAs balancing X-chromosome dosage in *Drosophila*, roX1 and roX2 (RNA on the X chromosome), were identified by use of enhancer traps and male-specific hybridization a few years later by Victoria Meller, Richard Axel, Mitzi Kuroda, Ron Davis, Richard Kelley, Asifa Akhtar and colleagues. The roX RNAs (whose activity is modulated by alternative splicing) act not to repress one of the two X-chromosomes in females but to globally upregulate gene expression from the single X chromosome in

[m] There are differences between rodents and primates, associated with distinctions in early development. The paternal allele of *Xist* is silenced in the trophectoderm (placenta) in mice, but not in humans. Seemingly random inactivation of parental alleles occurs in human trophectoderm and in both human and mouse embryos.[210,211]

males via tandem stem-loop structures that bind effector proteins to remodel chromatin and compartmentalize the X chromosome (also involving Phase Separation, Chapter 16).[214–222]

By the end of the 1990s, dozens of lncRNAs with regulatory functions had been identified in a wide variety of eukaryotes.[223,224] Early examples that hinted at the diversity of these non-coding RNAs included: 'meiRNA' in fission yeast, essential for the pairing of homologous chromosomes in meiosis,[225,226] involving phase separation[227] (Chapter 16); the dutA RNA in *Dictyostelium*, induced in development during mold aggregation;[228–230] an "unusual family" of 1.8–2.4 kb transcripts in the malarial protozoan parasite *Plasmodium falciparum* involved in the expression or rearrangements of virulence genes;[231] the bifunctional enod40 RNA in lucerne, expressed during nodule organogenesis, wherein the RNA structures are more highly conserved than the encoded peptides;[232–234] the pseudogene-derived antisense transcript[n] pseudoNOS (or antiNOS) suppressing the expression of the cognate nitric oxide synthase (*NOS*) gene in neurons of the snail *Lymnaea stagnalis*;[236,237] Xlsirt RNAs in frogs, "interspersed repeat transcripts" localized and playing structural roles in the vegetal pole cytoskeleton of *Xenopus* oocytes;[238,239] the 'yellow crescent RNA' in the ascidian *Styela clava*, a maternal transcript localized in the zygotic myoplasm;[240,241] the mammalian non-coding multi-exonic alternatively spliced 'growth arrest-specific' *gas5* gene that hosts several snoRNAs (Chapter 8)[242–244] and is also active as a mitochondrially localized long non-coding RNA;[245] and SRA, a highly conserved steroid receptor coactivator (Figure 9.5), which was found accidentally in a protein-binding screen and later described as being "different from eukaryotic transcriptional coactivators in its ability to function as an RNA transcript to selectively regulate the activity of a family of transcriptional activators".[246–249] Such screens also detected many other RNAs that act as transcriptional activators "when tethered to DNA".[250–252]

As in *Drosophila*, several lncRNAs in vertebrates identified during the 1990s were found to originate from developmental loci, often showing coordinated expression and functional relationships with their associated protein-coding genes. These included RNAs antisense to homeobox-containing genes, such as *HoxA11*,[253,254] *HoxD3*,[255] and *Dlx1* and *Dlx6*.[256,257] *Xist* was also found to be overlapped by a gene specifying a 40 kb unstable antisense lncRNA, Tsix, which was identified by RNA fluorescence *in situ* hybridization and negatively regulates *Xist* expression during the early steps of X inactivation,[258,259] an early indication of the highly intricate regulatory networks involving lncRNAs (Chapter 16).

From the end of the 1990s, it emerged that the imprinted *Igf2/H19* locus and many other imprinted loci differentially express overlapping sense and antisense transcripts,[198,260–264] including the *Igf2 receptor* (*Igf2r*) locus, where an astoundingly long (~108 kb) antisense RNA was serendipitously found to be transcribed from a promoter located in an intron of the *Igf2r* gene.[265,266] In contrast to the protein-coding gene *Igf2r*, which is expressed from the maternal allele, the non-coding RNA, termed Air (Antisense Igf2r RNA), is expressed only from the paternal allele.[265–267]

Similar phenomena were observed at other loci,[268–270] including *Xist*,[271] *Meg3*[272] (also known as *Gtl2*, first isolated by a gene trap approach in mice[273,274]) and the *Kncq1* locus. In the latter, transcription of a 91 kb lncRNA named *KvLQT1-AS* (*KvLQT1 antisense*, later known as *Kcnq1ot1*), like Air, initiates in an intron of the paternal allele of the protein-coding *Kncq1* gene from a 'CpG island' (a region of high GC dinucleotide content that is a target for DNA methylation, normally associated with gene repression; Chapter 14) called the imprinting control region (IC2)[275–279] (Figure 9.6). Given their reciprocal expression, these antisense RNAs were proposed to be involved in the silencing of the associated protein-coding genes,[280] demonstrated for *Air* in 2002.[281]

Both *Air* and *Kcnq1ot1* RNAs were later shown by the groups of Peter Fraser and Chandrasekhar Kanduri respectively, to silence transcription by binding to and targeting the histone methylase G9a and DNA methyltransferases to chromatin to alter the epigenetic state of the locus[282,283] (Chapters 14 and 16).

UTR-DERIVED RNAs

The protein-coding portion (the open reading frame) of mRNAs is flanked by 'upstream' and 'downstream' untranslated regulatory regions, referred to as 5′UTRs and 3′UTRs, respectively. 3′UTRs have increased in size with increasing morphological

[n] The first report of an antisense transcript from a pseudogene was that from human topoisomerase I by Bing-Sen Zhou and colleagues in 1992.[235]

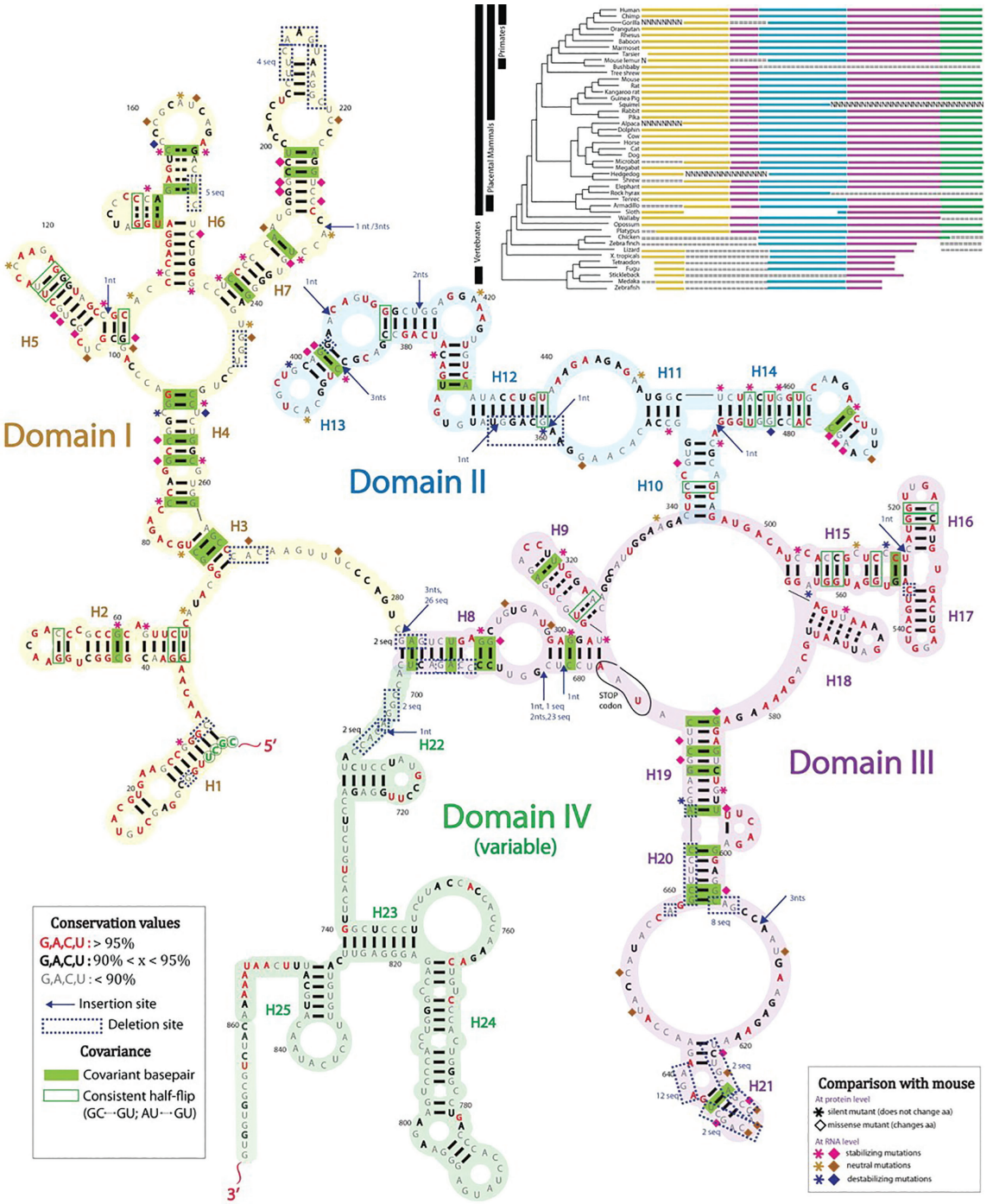

FIGURE 9.5 The structure and evolutionary conservation from fish to mammals of the steroid receptor coactivator RNA (SRA). (Reproduced from Novikova et al.[249] with permission from Oxford University Press.)

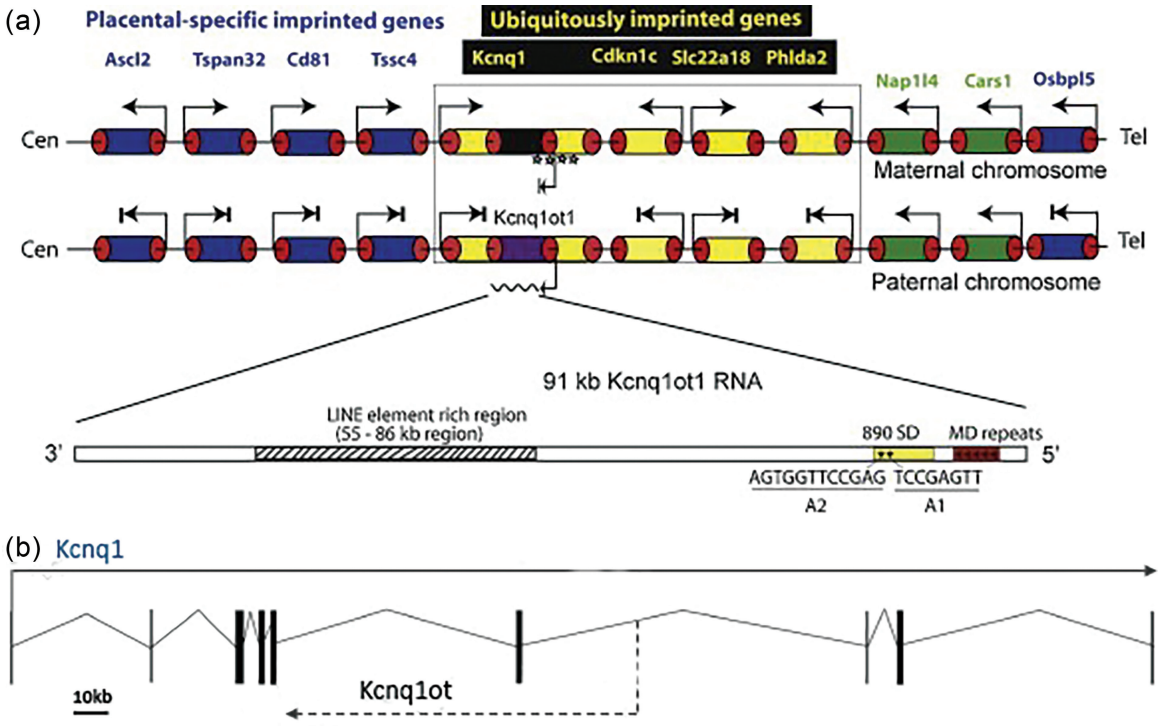

FIGURE 9.6 A composite figure showing (a) the genomic arrangement of the *Kcnq1* imprinted locus in mouse with (b) the exon-intron structure of the *Kncq1* gene and the antisense Kcnq1ot transcript initiated from its 6th intron. (Adapted from Kanduri[279] (a) and Pandey et al.[278] (b) with permission from Elsevier.)

complexity during animal evolution,[o] especially in vertebrates, where they usually occupy as much or more of the mRNA as the coding sequence in mammals and are often highly conserved.[285–287] 3′UTRs contain modules that bind regulatory proteins and small RNAs (Chapter 12) to control the translation, localization and stability of the mRNA.[288,289]

In 1993, Helen Blau and colleagues discovered that the 3′UTRs of three muscle associated genes (troponin I, tropomyosin and α-cardiac actin) could inhibit cell division and suppress malignancy in a myogenic cell line independently (i.e., in the absence) of the normally associated protein-coding sequence.[290,291] Other trans-acting 3′UTR-derived RNAs with similar properties were reported from the genes encoding ribonucleotide reductase[292] and prohibitin.[293] It was also shown that the loss of oogenesis caused by the lack of the *Drosophila* gene *oskar* can be rescued by its 3′UTR alone, indicating that this RNA "acts as a scaffold or regulatory RNA essential for oocyte development".[294]

Later studies showed that independent expression of 3′UTR sequences is widespread (Chapter 13), occurring in as many as half of all mammalian genes[289,295–299] as well as commonly in plants.[300] These "UTR-associated RNAs" (uaRNAs) or "downstream of genes" (DoGs) are, at least in some cases, nuclear-localized and induce differentiation separately from their usually associated protein-coding sequences (Chapter 13).[290–295,297,301–303] In the testis, for example, the coding sequences of the *Myadm* gene are expressed in the cytoplasm of the interstitial cells, whereas the 3′UTR is not expressed in these cells but highly expressed in the nuclei of germ and Sertoli cells.[295] This phenomenon is particularly pronounced in the brain[295–297] where, for example, the 3′UTR of the *Klhl31* gene but not the coding region is highly expressed in the cerebellum and the hippocampus[295] (Figure 9.7), and cytoplasmic cleavage of the IMPA1 3′UTR is necessary to maintain axon integrity.[303]

[o] Exons specifying 5'UTRs have expanded in humans and are highly alternatively spliced.[284]

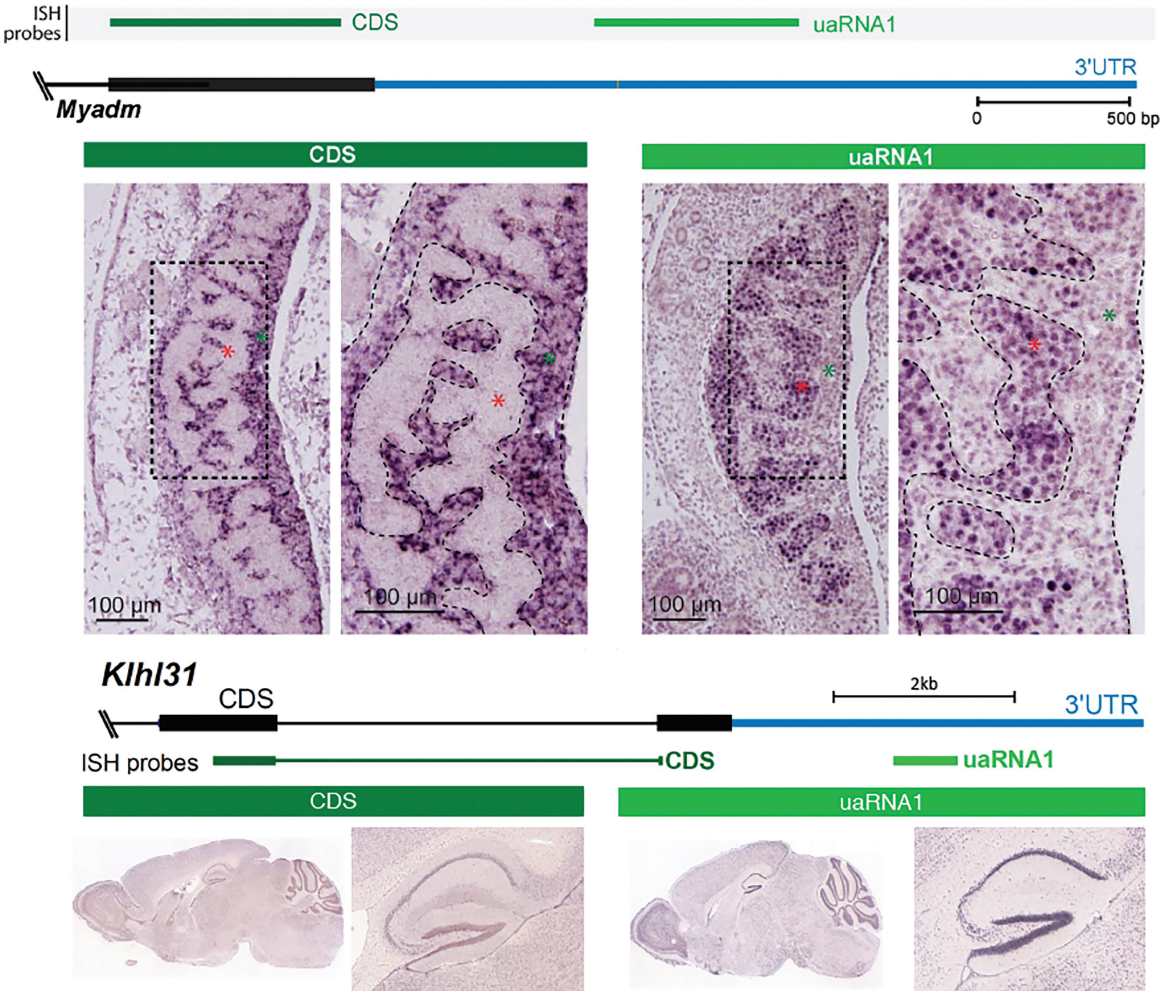

FIGURE 9.7 Top panel: Expression of *Myadm* coding sequences in the interstitial cells of the developing mouse testis and the 3'UTR in the nuclei of Sertoli and germ cells in the testis cords. Bottom panel. Expression of the Klhl31 3'UTR but not coding sequences in the cerebellum and hippocampus, with close-up of the hippocampus showing especially strong expression in the dentate gyrus. (Reproduced from Mercer et al.[295] with permission from Oxford University Press.)

The regulatory and evolutionary logic of having a covalently linked RNA sequence that regulates mRNA activity *in cis* but also acts independently *in trans* is an astounding observation, whose biological *raison d'être* is yet to be satisfactorily explained, but is emblematic of the complexity and mysteries of the emerging world of RNA regulation. It also presaged later findings that non-coding RNAs can act as 'decoys' or 'sponges' for bacterial[47,59] and eukaryotic small RNAs[301] and RNA-binding proteins (Chapters 12, 13 and 16). It is also clear, and in retrospect unsurprising, that the terms messenger and regulatory RNA are not mutually exclusive and that individual RNAs can have multiple functions (Chapter 13).

FIRST EXAMPLES OF SMALL REGULATORY RNAs IN ANIMALS

Also in 1993, two articles published by the groups of Gary Ruvkun and Victor Ambros described a small RNA that played a role in developmental regulation in the nematode worm *Caenorhabiditis elegans*.[304–307] Previous genetic screens by their groups had shown that the product of the *lin-4* gene regulates the expression of *lin-14*, a heterochronic gene

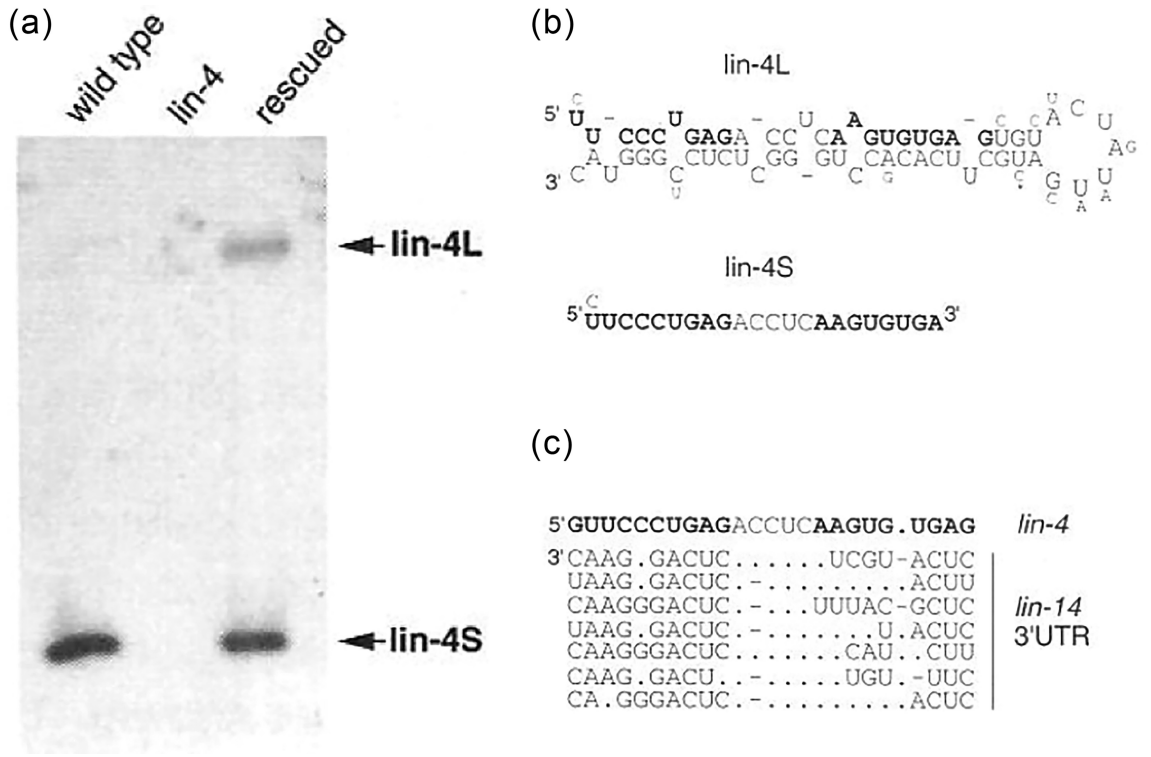

FIGURE 9.8 (a) Northern blot showing the small ~22nt RNA (lin-4S) produced by the *lin-4* gene and its precursor (lin4-L) in wildtype *C. elegans*, absent in a deletion mutant. (b) The sequences lin-4S and lin4-L, the latter showing the predicted secondary stem-loop structure. Sequences complementary to the lin-14 3'UTR are bold. (c) The complementarity between lin-4 and seven copies of a repeated element in the 3'UTR of lin-14 RNA that is conserved in *C. elegans* and *C. briggsae*. (Reproduced from Lee et al.[304] with permission from Elsevier.)

encoding a nuclear protein involved in the temporal control of post-embryonic development, by a mechanism targeting the 3'UTR of *lin-14*.

Both groups were expecting a regulatory protein,[306,307] but the *lin-4* locus mapped to the intron of a long spliced non-coding RNA.[304] The Ambros' group found that the primary *lin-4* transcript was processed into two overlapping RNAs of 61nt and 22nt in length.[304] Similar to the elucidation of the roles of spliceosomal snRNAs and snoRNAs, both groups found that the small RNAs produced from the *lin-4* locus had partial complementary to a number of sequences in the 3'UTR of the *lin-14* mRNA (Figure 9.8). Curiously, while noticing that lin-14 protein levels are reduced in development without changes in the transcript abundance, they proposed that these "small temporal RNAs" formed multiple RNA duplexes that inhibited the translation of *lin-14* mRNA. This inhibition depended on the (partial) base pair complementarity, which was conserved in the homologous sequences of the related species *C. briggsae*.[304,305,308,309] Thus, it was speculated that there may be a "novel kind of antisense translational control mechanism" and that "lin-4 may represent a class of developmental regulatory genes that encode small antisense RNA products".[304]

They were not to realize how prophetic these words would be (Chapter 12).

CURIOSITIES OR EMISSARIES?

The general significance of this finding was, once again, not recognized at the time. tRNAs were for a long time considered the "smallest biologically active nucleic acids known".[310] Given the "incredible" small size of these RNAs and the lack of obvious homologs outside of worms, even the groups of Ruvkun and Ambros saw them as a

"curiosity" of worms, comparable to the "gene regulatory vignettes" of small bacterial regulatory RNAs and the few known eukaryotic non-coding RNAs.[304,306,307]

Not only was it novel that such tiny RNAs could have regulatory properties, it was also surprising that they originated from an intron and targeted non-coding regions in mRNAs, at a time when the regulatory relevance of 3′UTRs was still being established.[311] However, this did not disturb the prevailing conceptual framework. According to Ambros, "there was no theoretical need to explain existing phenomena in terms of new mechanisms or new classes of molecules. Transcription factors-mediated regulation of cell fate was a successful model to account for developmental biology".[307]

A 1994 editorial in *Science* highlighted these emerging findings, the various interpretations and Mattick's hypothesis, remarking on the existence of "too many cases of odd RNAs" and speculating that "there might be a whole family of regulatory RNAs".[312] Similarly, an editorial in *Nature* posed the question: "Are these RNAs all grotesque deviants, one-of-a-kind aberrations, like characters in a Fellini film?" and admonished "But pay close attention to them. [They may] instead have been the first emissaries from an unexplored and vast RNA world".[311]

FURTHER READING

Breaker R.R. (2012) Riboswitches and the RNA world. *Cold Spring Harbor Perspectives in Biology* 4: a003566.

Darnell J.E. (2011) *RNA: Life's Indispensable Molecule* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Lee R., Feinbaum R. and Ambros V. (2004) A short history of a short RNA. *Cell* 116: S89–S92.

Mattick J.S. (2004) RNA regulation: A new genetics? *Nature Reviews Genetics* 5: 316–23.

Morange M. (1998) *A History of Molecular Biology* (Harvard University Press, Cambridge, MA).

Morris K.V. and Mattick J.S. (2014) The rise of regulatory RNA. *Nature Reviews Genetics* 15: 423–37.

Ruvkun G., Wightman B. and Ha I. (2004) The 20 years it took to recognize the importance of tiny RNAs. *Cell* 116: S93–6.