

Noncoding RNA genes

Sean R. Eddy

Howard Hughes Medical Institute & Department of Genetics,
Washington University School of Medicine
Saint Louis, Missouri 63110 USA
eddy@genetics.wustl.edu

October 20, 2001

Abstract

Noncoding RNA (ncRNA) genes produce functional RNA molecules rather than encoding proteins. Almost all means of gene identification assume that genes encode proteins, so even in the era of complete genome sequences, ncRNA genes have been effectively invisible. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes. Noncoding RNAs seem particularly abundant in roles that require highly specific nucleic acid recognition without complex catalysis, such as in directing post-transcriptional regulation of gene expression or in guiding RNA modifications.

Introduction

One goal of genome projects is to systematically identify genes [1]. This past year, two papers announced drafts of the human genome sequence [2, 3], but the estimated number of human genes continues to fluctuate. Current estimates center around 30,000-40,000, with occasional excursions to 100,000 or more [4–6]. One reason for the continuing ambiguity is that genes are neither well-defined nor easily recognizable. The numerology is based on three methods: cDNA cloning and EST (expressed sequence tag) sequencing of polyadenylated messenger RNAs [7, 8], conserved coding exon identification by comparative genome analysis [9], and computational gene prediction [2, 3]. These methods work best for large, highly expressed, evolutionarily conserved protein coding genes, and they almost certainly undercount other genes. They essentially do not work at all for one class of genes – the noncoding RNA (ncRNA) genes, which produce transcripts that function directly as structural, catalytic, or regulatory RNAs, rather than expressing messenger RNAs that encode proteins [10–12]. Knowledge of noncoding RNAs has been limited to biochemically abundant species and anecdotal discoveries. Even after the completion of many genome sequences, the number and diversity of ncRNA genes remains essentially unknown.

Could it be possible that a large class of genes has gone relatively undetected because they don't make proteins? How many ncRNA genes are there? How important are they? What functions does a cell delegate to RNA instead of protein, and why?

To address these questions, one needs to develop new systematic gene discovery approaches specifically aimed at ncRNAs. A pioneering study from Roy Parker's group found a few new RNA genes and small ORFs in the yeast genome by doing Northern blots probing for expressed transcripts in "grey holes" (suspiciously large "intergenic" regions), and by searching for consensus polIII promoters [13]. Recently, several groups have carried out systematic ncRNA gene identification screens along three main lines: cDNA cloning and

sequencing tailored to find new small non-messenger RNAs [14], specially designed cDNA cloning screens for a new regulatory RNA gene family of tiny RNAs called microRNAs (miRNAs) [15–17], and general ncRNA genefinding exercises using computational comparative genomics in *E. coli* [18–20]. The results of these screens are startling. All of them suggest that the prevalence of ncRNA genes has indeed been underestimated.

The idea that a class of genes may have remained essentially undetected is provocative, if not heretical. It is perhaps worth beginning with some historical context of how ncRNAs have been discovered. Gene discovery has been biased towards mRNAs and proteins for a long time.

The lessons of history

RNA's central role in translation. It was clear by the 1950's that DNA was in the eukaryotic nucleus, but proteins were being synthesized in the cytoplasm in the presence of abundant RNA [21, 22]. Most of this cellular RNA could be found in discrete particles in the cytoplasm [23], which were later shown to be the site of protein synthesis and dubbed ribosomes [24]. Watson sketched the Central Dogma as early as 1952 [25, 26], imagining that there must be a coding RNA passed from the DNA to the protein synthetic machinery in the cytoplasm. The prevailing theory was the now-forgotten “one gene, one ribosome, one protein” hypothesis [24, 27] that each gene produced a specialized ribosome composed of a specific messenger RNA associated with general ribosomal proteins that catalyzed translation. Various results undermined this hypothesis, including the simple observation that while genes came in a great variety of sizes and base compositions, ribosomal RNAs had no variety [27]. Finally ribosomes were found to be general purpose RNA/protein machines, composed largely of stable ribosomal RNAs [28], and programmed with a variety of unstable messenger RNAs that are only a small fraction of the total RNA population [27, 29].

The second class of functional RNA was predicted by Crick's “adaptor hypothesis” [24]. Crick predicted the existence of a molecule that mediates between the triplet genetic code and the encoded amino acid. Interestingly, Crick argued that the adaptor would not only be an RNA, but that RNA would be evolutionarily *preferred* over protein as the material for his adaptors, because base-pairing made RNA uniquely suited for a role as a small specific RNA recognition molecule [24]. Crick's adaptors had in fact just been biochemically observed by Hoagland and coworkers [30]. These RNAs later proved to be Crick's adaptors, the transfer RNAs (tRNAs) [31].

RNA therefore went from one flavor (the purely information-carrying intermediate in the Central Dogma) to three flavors, all apparently involved in making protein: rRNA, tRNA, and everything else, which was assumed to be mRNA. Genetics and enzyme biochemistry had already shown links between mutant genes, missing enzymatic activities, and missing or altered proteins. The central intellectual problem was to solve the genetic code. The non-rRNA, non-tRNA fraction was complex, non-abundant, and mostly unstable, and there was little motivation or ability to go any further, and ask whether it contained more than mRNA.

RNA comes in more than three flavors. Nonetheless, a number of other abundant small non-messenger RNAs were soon detected and isolated biochemically, among them the U-rich “U RNAs” [32, 33]. Many of these small RNAs are associated with proteins to form ribonucleoprotein (RNP) complexes [34]. Characterization of small RNPs was aided by the discovery that certain patients with autoimmune diseases like systemic lupus erythematosus produce anti-RNP autoantibodies that could be used to immunoprecipitate small RNPs [35]. Many of the abundant small RNPs precipitated by these antisera turned out to be components of the spliceosome, containing U1, U2, U4, U5, and U6 snRNA [34, 36], involved in splicing messenger RNAs. Other U RNAs – U4atac, U6atac, U11, and U12 – have been found to be components of a second spliceosome species [37, 38].

Many other small RNAs have been isolated biochemically. Sometimes these isolations are deliberate, such as the isolation of numerous small nucleolar RNAs (snoRNAs) from nucleoli [39]). In other cases,

biochemical fractions were unexpectedly found to contain essential RNAs, as in the case of ribonuclease P [40]. One of the best stories of such a surprise resulted in the renaming of signal recognition “protein” to signal recognition “particle” when it was unexpectedly found to contain a 7S RNA now called SRP-RNA [41, 42].

New RNAs continue to be crop up; among the more fascinating stories is the discovery that RNAs play roles in chromatin structure [43]. A canonical example is the human *Xist* RNA, a 17 kb noncoding RNA with a key role in dosage compensation and X chromosome inactivation [44]. *Drosophila* also appears to control dosage compensation using small chromatin-associated *roX* RNAs [45]. Several large noncoding RNAs have been found expressed from imprinted regions of vertebrate chromosomes, including the *IPW* (imprinted in Prader-Willi) and *H19* transcripts [46, 47]. (The imprinted Prader-Willi critical region seems especially rich in noncoding RNAs [48, 49]; it is unclear whether this is peculiar, or simply due to the incredibly intense gene hunting in search of the elusive cause of Prader-Willi.) Many of these other RNAs are cis-antisense RNAs that overlap coding genes on the other genomic strand. Various cis-antisense RNAs have been observed in prokaryotes [50], plants [51], and animals [12], and their roles are unlikely to be limited to roles in imprinting and chromatin structure. Mutations in one cis-antisense RNA in humans, SCA8, are found in patients with spinocerebellar ataxia [52].

Continued flurries of snoRNAs

The nucleolus is rich in small nucleolar RNAs (snoRNAs), most of which are about 70-250 nt long [53, 54]. Some snoRNAs have roles in rRNA processing, but most function in rRNA modification [39]. Based on weak sequence similarities, almost all snoRNAs fall into two families: the “C/D box” snoRNAs and the “H/ACA” snoRNAs [39, 55]. The C/D box snoRNAs use base complementarity to guide site-specific 2'-O-ribose methylations to rRNA [56–58], and the H/ACA snoRNAs used base complementarity to guide site-specific pseudouridylations to rRNA [59, 60] (Figure 1). In both cases, the catalytic function appears to be provided by a protein methylase or pseudo-U synthetase associated with the snoRNA, and the specificity for the target base on rRNA is provided by base complementarity to the snoRNA [61–63].

snoRNAs must be numerous. For many eukaryotes, the approximate number of specific 2-O-ribose methylations and pseudouridylations is known, and for some species, many modified positions have been precisely mapped [64–66]. In human rRNAs, for instance, there are about 100-110 of each type of modification, and in yeast, about 50 of each. If snoRNAs direct most (or all) eukaryotic nuclear rRNA 2'-O-ribose methylations and pseudouridylations, there must be a large number of undiscovered snoRNAs. Indeed, computational screens have revealed 41 new C/D snoRNAs in the yeast genome [67] and over 60 new C/D snoRNAs in the *Arabidopsis* genome [68, 69]. Immunoprecipitation with antibodies against fibrillarin (the putative methyltransferase) revealed 17 new C/D snoRNAs in *Trypanosoma brucei* [70], and cDNA sequencing has found 72 new C/D snoRNAs and 41 new H/ACA snoRNAs in mouse (see below) [14]. Numerous homologues of the C/D snoRNAs have been found in the Archaea [71, 72], where they are presumed to have the same function of guiding specific 2'-O-ribose methylations of target RNAs.

snoRNAs modify RNAs other than rRNA. Besides ribosomal RNA, other structural RNAs such as tRNAs and snRNAs are known to be extensively modified [33, 73, 74]. Now it appears that some, if not many, of these modifications are snoRNA-guided as well. At least one of the 2'-O-ribose methylations of *Xenopus* U6 snRNA is guided by a C/D snoRNA, mgU6-77 [73]. Human U85 is a chimeric C/D, H/ACA snoRNA (a “Siamese snoRNA”) that guides both a methylation and a pseudouridylation of U5 snRNA [75]. Sequencing of snoRNA-enriched cDNA libraries has revealed several “orphan” snoRNAs with no obvious rRNA target [49, 76, 77], as have the computational screens for Archaeal snoRNAs [71] where a few such cases have been putatively assigned to known tRNA 2'-O-ribose methylations. One puzzling aspect of these discoveries is that one has to wonder how non-ribosomal RNAs are trafficking through the nucleolus, or

whether perhaps there is at least one more site of RNA modification in the cell; recent evidence suggests that the snRNA modifications are associated with Cajal bodies (coiled bodies) in the nucleus [78].

An EST screen for small non-mRNAs. Hüttenhofer *et al.* undertook a general screen for novel small non-messenger RNAs, using an EST sequencing approach [14]. The RNA population used was total (not cytoplasmic) mouse brain RNA that was cloned by tailing (not poly-A selection and dT-priming) and size-selected in two small RNA fractions, 50-100 nt and 110-500 nt. High-throughput filter hybridization was used to screen out clones corresponding to tRNA, rRNA fragments, and other known ncRNAs, increasing the fraction of novel ncRNA sequences from about 3-7% in an unscreened library to about 20-22% after screening. A total of about 5000 clones were sequenced; after accounting for multiple sequences of the same RNA species, 201 different novel RNA sequences were identified.

A little over half of these appear to be new snoRNAs – 72 new C/D snoRNAs and 41 new H/ACA snoRNAs. Of these, several are orphans that do not have obvious rRNA or snRNA targets. Some of these snoRNAs showed brain-specific expression, which would not be predicted for molecules involved in ubiquitous rRNA modification. The human homologues of three of these snoRNA genes mapped to the critical region for Prader-Willi syndrome, two of which (HBII-52 and HBII-85) are C/D snoRNAs found in multi-copy tandem arrays, unlike most vertebrate snoRNAs which are found in single copies in introns of other genes, and both are expressed as imprinted genes only from the paternal chromosome, as expected for a Prader-Willi candidate gene. The HBII-85 array, located just to the left of (centromeric to) the noncoding IPW gene, was also detected as an imprinted noncoding RNA gene array by other studies [48, 79]. The HBII-52 snoRNA has a perfect 18 bp complementarity to serotonin 2C receptor mRNA, and is predicted on that basis to methylate a site of known mRNA editing; this suggests a complex scenario in which a snoRNA may be regulating the editing of an mRNA transcript [14, 80].

Of the 88 sequences that did not appear to be snoRNAs, 20 that did not correspond to known mRNAs or repetitive elements were confirmed as expressed small discrete RNAs by Northern blots, with sizes ranging from 65 to 500 nt. The functions of these 20 novel small RNAs are unknown. Hüttenhofer *et al.* are now analyzing similar libraries from *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*.

miRNAs: one, two... infinity?

One: *lin-4*. A canonical example of the identification of a noncoding RNA gene by genetics is the story of the *lin-4* regulatory RNA in the nematode *Caenorhabditis elegans*. The *lin-4* locus was identified in a screen for mutations that affect the timing and sequence of postembryonic development (heterochronic mutations) in *Caenorhabditis elegans* [81]. Mutant animals reiterate the L1 larval stage rather than progressing to later stages of development. The gene was positionally cloned by isolating a 693 bp DNA fragment that could rescue the phenotype of mutant animals [82]. The Lee *et al.* paper dryly recounts a careful detective story, as the Ambros lab gradually realized that they were dealing not with a protein coding gene, but with a tiny noncoding RNA. The *lin-4* gene product is a 22 nt RNA, processed from a 61 nt precursor RNA with a putative stem-loop structure.

Genetically, *lin-4* acts as a negative regulator of heterochronic protein-coding genes such as *lin-14* and *lin-28*. The 3' UTRs of the target genes have short stretches of complementarity to *lin-4* [82–84] (Figure 2). Deletion of these apparent *lin-4* target sequences cause an unregulated gain-of-function phenotype [83, 84]. The *lin-4* RNA inhibits accumulation of the LIN-14 and LIN-28 proteins by an unknown mechanism. The target mRNA remains stable, fully polyadenylated, and polysome-associated [85].

Two: *let-7*. The *lin-4* gene remained an oddity until a second heterochronic gene, *let-7*, also mapped to a noncoding RNA gene with a 21 nt product [86]. The small *let-7* RNA is also thought to be a posttranscriptional negative regulator, possibly targeting the protein-coding mRNAs for *lin-41* and *lin-42*, based on

phenotypic analysis and plausible complementary sequences in the 3' UTR of these genes.

Surprisingly, Pasquinelli *et al.* showed that *let-7* is almost 100% conserved and expressed as a small 21 nt RNA in all bilaterally symmetric animals tested – including human, mouse, chicken, polychaete worms, and flies, but not in cnidarians (jellyfish) or poriferans (sponges) [87]. The function of these *let-7* homologues is unknown, but because they show temporal regulation generally similar to *let-7*'s developmental pattern in the worm, one presumes they also function in post-transcriptional regulation of developmental genes. Pasquinelli *et al.* proposed the name *small temporal RNAs* (stRNAs) for genes like *lin-4* and *let-7*, and suggested that others might be found.

A surprising link to RNA interference. Meanwhile, the increasingly baroque phenomenology of double-stranded RNA interference (RNAi) was being elucidated [88–91]. Introduction of exogenous double-stranded RNA (dsRNA) into nematodes, by direct injection or even by feeding, leads to the specific rapid degradation of homologous mRNA(s), and a loss of function phenotype. RNAi also works in many other organisms, including plants, where the effect has been called cosuppression or post-transcriptional gene silencing [89, 91]. The input dsRNA is cleaved to form the active agents of the RNAi effect, tiny 21–25 nt “small interfering RNAs” (siRNAs) [92–94]. Several proteins important in the RNAi pathway have been identified, including the putative processing nuclease Dicer and a large family of homologous proteins including *Caenorhabditis Rde-1*, *Arabidopsis ARGONAUTE*, and *Drosophila Piwi*. RNAi has been suggested to function as a primitive immune system against RNA viruses and retrotransposons [90, 91].

Many people noted with suspicion that the sizes of the active *lin-4* and *let-7* stRNAs (22 and 21 nt) are the same as the sizes of siRNAs [87, 90, 95]. Indeed, the RNAi processing pathway shares components with the stRNA processing pathway. Knocking down Dicer function in human cultured cells leads to accumulation of the 72 nt unprocessed human *let-7* precursor [93]. Knocking down either the function of the *C. elegans* Dicer homolog or two of the 23 worm homologs of the *rde-1/Argonaute/piwi* gene family, *alg-1* and *alg-2*, results in accumulation of unprocessed *lin-4* and *let-7* precursor [96]. In the course of cloning and analyzing the small RNAs produced from an exogenous dsRNA, Tuschl's lab noted in passing that *Drosophila* appeared to contain endogenous 21- and 22-mers [94], and suggested that perhaps there were naturally occurring siRNAs.

Introducing the miRNAs. Now, a trio of three papers shows that, indeed, *lin-4* and *let-7* are not alone – they belong to a potentially very large family of small RNAs in nematode, fly, and human (and presumably other organisms) that are being called the microRNAs (miRNAs). Lau *et al.* produced and sequenced a *C. elegans* cDNA library cleverly enriched for tiny RNAs with 5'-phosphate, 3'-hydroxyl termini, and obtained 55 new miRNAs [16]. Lee and Ambros used a size-selected *C. elegans* cDNA library and, to a lesser extent, a computational approach looking for sequences conserved with *C. briggsae* that can be folded into a stem similar to the *lin-4* and *let-7* precursors, to find 15 miRNAs [15]. Lagos-Quintana *et al.* used size-selected cDNA libraries in human and *Drosophila* to isolate 33 miRNAs, 14 in human and 19 in *Drosophila* [17].

In total, so far 91 different miRNAs have been identified in the three species. Some of these are highly conserved in evolution, like *let-7*; homologs of eleven miRNAs are found in more than one of the three species. Northern analyses have been done for many of these RNAs, and generally show both a 21–24 nt form (presumably the active miRNA) and, often, a less abundant ~70 nt form (presumably the precursor stem-loop). The miRNA genes are often clustered in the genome [16, 17] and may be co-expressed in polycistronic precursor transcripts. Many of the miRNAs were identified by single cDNA sequences, so it is clear that none of these screens are near saturation.

It appears that miRNAs are more likely to function as translational repressors like *lin-4*, not as siRNAs in directing mRNA degradation. Like *lin-4* but unlike siRNAs, miRNAs are produced asymmetrically from the precursor stem; almost invariably, only one strand of the precursor stem can be recovered as a 21–24 nt product (the Bartel lab reports a single exception [16]) (Figure 3). Many miRNAs are produced in a stage- and/or tissue-specific manner, suggesting possible roles in development akin to the stRNAs. Some of the

C. elegans miRNAs are specifically expressed in the germ line and embryo, where translational regulation is particularly prevalent [16]. If the parallels with *lin-4* hold up, the miRNAs should be expected to direct translational repression by binding to one or more sites with imperfect complementarity in 3' UTRs of coding mRNAs (Figure 2).

Another puzzling observation about RNAi now seems to make more sense. Some of the genes implicated in the RNAi processing pathway have lethal phenotypes or show developmental defects when knocked out, which does not make sense if they are functioning solely in RNAi and as an antiviral defense mechanism [88, 90, 91]. In *C. elegans*, knockdowns of some genes in the *rde-1/Argonaute/piwi* gene family, such as *rde-1* itself, produce RNAi-defective phenotypes but no developmental phenotypes, whereas knockdowns of others, such as *alg-1* and *alg-2*, produce developmental phenotypes but are still RNAi sensitive [96]. It appears that in addition to the RNAi effect itself, components of the RNAi processing pathway also function in developmental regulatory processes that may involve a large number of endogenous miRNAs.

Even *E. coli* has been hiding something

The bacterium *Escherichia coli* is, arguably, the best studied organism. The complete genome sequence of *E. coli* K-12 contains an estimated 4200 or so protein coding genes [97]. The small number of known ncRNA genes has continued to climb slowly, as several small stable RNAs have been found anecdotally [98]. Many of these appear to be “riboregulators” [99, 100] that act by using base complementarity to specifically interact with translational start sites and either repress or, more rarely, activate translation (Figure 4). The recent availability of comparative genome sequence information from *Salmonella* and other enterobacteria made it possible to go looking for conserved sequences that might correspond to ncRNAs, instead of coding ORFs.

Argaman *et al.* computationally analyzed intergenic regions to identify loci that have a predicted promoter and terminator spaced 50-400 nt apart and are significantly conserved in other bacterial genomes [18]. They predicted 24 candidate ncRNA genes, 14 of which were shown by Northern blot analysis to produce discrete small transcripts ranging from 70 to 250 nt in size. Many of the RNAs were expressed in conditions other than “normal” exponential growth in rich media; four RNAs were only expressed in stationary phase, five more were preferentially expressed in stationary phase, one was expressed almost exclusively in cold shocked cells, and one was preferentially expressed in minimal media.

Wassarman *et al.* looked for intergenic regions that showed significant conservation to other genomes, then manually prioritized and selected amongst these using additional criteria such as the separation of the conserved region from nearby ORFs, the presence of plausible promoter and terminator signals, and significant RNA expression detected on whole-genome high density oligo probe arrays [20]. They predicted 59 candidate ncRNA loci, 17 of which were shown by Northern analysis to produce discrete small RNA transcripts ranging from 45 to 320 nt in size. Again, several of these were seen to be expressed almost exclusively in stationary phase cells. Seven of these RNAs could be immunoprecipitated with antisera to the abundant RNA-binding protein Hfq, which binds two previously known ncRNAs in *E. coli*, OxyS and DsrA.

Rivas *et al.* have developed a general ncRNA gene-finding algorithm [101]. The algorithm uses comparative genome sequence analysis to detect conserved sequence regions in which the pattern of mutation is more consistent with conservation of a base-paired secondary structure than with conservation of a coding amino acid sequence or with a null hypothesis of uncorrelated position-independent mutation. The approach is therefore limited to detecting only ncRNAs with conserved intramolecular secondary structure. The algorithm was used to screen the *E. coli* genome, and it detected 275 candidate loci [19]. A sample of 49 of these loci were analyzed by Northern blot in a single growth condition (exponential growth in rich media) and 11 were found to produce small RNAs ranging from 40 to 370 nt in size.

In total, three systematic screens have identified 34 new ncRNA transcripts in *E. coli*, of as yet unknown function. There is little overlap in the confirmed transcripts (only 8 were confirmed by more than one of the screens). This indicates that these screens have not saturated the *E. coli* genome for novel ncRNAs. Of the 27 genes confirmed by one or both of the Argaman *et al.* and Wassarman *et al.* screens, 21 are in the Rivas *et al.* candidate list, indicating that the sensitivity of the computational genefinder is fairly high. The experimental characterization done by Argaman *et al.* and Wassarman *et al.* shows that many ncRNAs are being expressed in specific growth conditions, something that had already been seen for known *E. coli* ncRNAs, for instance for the OxyS RNA (expressed in oxidatively stressed cells) [102] or the CsrB RNA (expressed in stationary phase) [103]. This indicates that the examination of a single growth condition by Rivas *et al.* was insufficient, and shows that confirming expression of a candidate ncRNA gene is not necessarily straightforward.

How many new ncRNAs is *E. coli* still hiding? Simulation studies of the false positive rate in the Rivas *et al.* study suggest that 200 or more of the 275 gene predictions should be real ncRNAs (or more precisely, biologically relevant sequences conserving an RNA structure; the approach cannot easily distinguish cis-regulatory RNA structures from independent ncRNA genes) [19]. Gottesman *et al.* intuit that it would be unlikely to find more than 50 new ncRNAs in *E. coli* [20]. Argaman *et al.* (wisely) did not speculate [18]. A fourth screen using a single-sequence neural net based computational genefinding approach in *E. coli* predicted 370 sequence windows as ncRNA genes (because the windows could overlap, this means a somewhat smaller number of RNA gene loci) [104]. These predictions have yet to be experimentally verified, and the amount of overlap with the other screens needs to be examined.

Matters arising

Many genes, little genetics. On one hand, we have genomic screens that are unsaturated and must be just a taste of larger numbers of ncRNA genes to come. On the other hand, if there were many ncRNA genes, one would think they would have been detected sooner in classical genetic screens. (Most biochemical, computational, and molecular biology gene discovery approaches make strong assumptions about seeking proteins, open reading frames, and messenger RNA, so it is easier to rationalize their failure to detect ncRNAs.) There are some biases even in classical genetics, though. Strikingly, few of the *known* ncRNA genes have been identified by genetics. For example, *none* of the known *E. coli* small RNAs have been identified by mutational screens [20, 98]. RNA genes are immune to frameshift or nonsense mutations, and often small and multicopy, which makes them difficult (even impossible) targets for recessive mutational screens.

An interesting additional source of bias is in going from a mapped genetic locus to a cloned gene. Especially in more complex systems, candidate gene identification is often essential for pinpointing a mutant locus, but candidate gene identification is biased towards ORFs and coding genes. Ridanpää *et al.* recently provided an excellent example in human genetics. Cartilage-hair hypoplasia (CHH), a short-limbed dwarfism, was first described by McKusick almost thirty years ago [105]. Positional cloning failed to identify the gene despite straightforward and accurate genetic mapping [106, 107]. Ridanpää *et al.* finally increased the resolution of the genetic map by almost an order of magnitude and sequenced the entire human genomic region. All ten identifiable protein-coding genes were studied with no luck. CHH-associated mutations were at last discovered in the 267 nucleotide *RMRP* ncRNA gene, which produces the essential RNA component of the ribonucleoprotein endoribonuclease MRP (MRP stands for “mitochondrial RNA-processing”) [108]. The only reason *RMRP* was considered as a candidate gene was that human MRP RNA had been previously isolated biochemically [108] and its sequence was in GenBank. Otherwise, Ridanpää *et al.* might still be looking.

One other human genetic disorder has been mapped to a nuclear-encoded ncRNA candidate gene by

positional cloning. Autosomal dominant dyskeratosis congenita patients have mutations in telomerase RNA [109]. Here again, telomerase RNA was already in the database, and moreover, it was an obvious candidate gene; an X-linked dyskeratosis had already been associated with dyskerin, a protein known to interact with telomerase RNA.

The power of comparative analysis. It is difficult to distinguish coding genes with short ORFs from ncRNA genes. Many sequences have long ORFs and are obviously coding, but for others, coding potential is less convincing. Protein-coding regions as small as 7 amino acids are known [110]. Open reading frames of over a hundred amino acids can occur just by chance in completely random sequence; it has been argued that 10-15% of annotated open reading frames in microbial genomes are in fact spurious [111]. Open reading frame length and “coding potential” alone is therefore often insufficient to decide that a gene is coding or noncoding. Errors are being made in both directions. The 360 nt bacterial regulatory ncRNA CsrB [103] was originally misannotated as a 47 amino acid protein, because that was the ORF closest to several mapped mutations [112]; the erroneous “protein” sequence is still in GenBank (*Erwinia carotovora aepH*, AAB32243.1). Conversely, the plant (*Medicago*) *enod40* gene was first thought to be an ncRNA gene based on sequence analysis that revealed “no significant coding potential” [113], but now, based on comparative genome analysis and more detailed directed mutagenesis studies, the *enod40* transcript appears to encode two tiny proteins 13 and 27 amino acids long [114].

Comparative genome analysis is an indispensable means of inferring whether a locus produces a non-coding RNA as opposed to encoding a protein. For a small gene to be called a protein-coding gene, one excellent line of evidence is that the ORF is significantly conserved in another related species. For almost all protein-coding genes (those undergoing purifying selection or neutral drift, but perhaps not genes under positive selection) the pattern of mutation should also favor synonymous and conservative amino acid changes. Comparative analysis has been instrumental in many cases of distinguishing ncRNA genes from small protein-coding genes, including the examples above. It is more difficult to positively corroborate an ncRNA by comparative analysis, but in at least some cases, a noncoding RNA may conserve an intramolecular secondary structure and comparative analysis can show compensatory base substitutions [19,101]. With comparative genome sequence data now accumulating in the public domain for most if not all major genetic systems, comparative analysis can (and should) become routine.

Discovering new ncRNA genes. There are now three main lines of attack for systematically identifying new ncRNA genes.

First, computational comparative genome analysis seems to be a very powerful approach. All three ncRNA screens in *E. coli* exploited comparative analysis [18–20], as did one of the screens for new miRNAs in *C. elegans* [15]. These approaches range in complexity from BLASTN screens that identify conserved regions that do not correspond to apparent ORFs, to identifying regions that conserve some particular type of RNA structure (such as the miRNA precursor stem), to a general ncRNA genefinding program looking for any significant conserved intramolecular secondary structure [101]. Previous attempts to develop ncRNA genefinders that work on a single genome sequence have been stymied by the apparent lack of much significant statistical signal in ncRNAs [115, 116], compared to the strong ORF and codon bias signals exploited by protein coding genefinders. However, an apparently successful single sequence RNA genefinder using a neural network approach has recently been reported [104], and it may also be possible to identify untranslated spliced ncRNAs just from computational identification of clustered splice site signals [117].

Second, cDNA cloning strategies specifically designed to enrich for non-messenger ncRNAs have been very fruitful. The most obvious enrichment strategy is simply to clone and sequence small RNAs from total RNA (as opposed to the usual selection of large, cytoplasmic, polyadenylated mRNA for cDNA cloning and EST sequencing) [14]. Enrichment by immunoprecipitation with antisera against proteins that associate with specific families of ncRNAs is another strategy that has been used for decades; examples include the isolation of snRNAs using anti-Sm autoantibodies [35] and isolation of C/D snoRNAs using anti-fibrillar

sera [71]. Some ncRNAs can be enriched by virtue of 5' ends that differ from the "normal" mRNA cap; intronic snoRNAs and miRNAs, for example, have simple 5' phosphates that are substrates for RNA ligase [16, 56]. Enrichment by exploiting the subcellular localization of ncRNAs can also be useful, as in the isolation of snoRNAs from cDNA libraries made from purified nucleoli [56]. There must be other clever enrichment schemes. Unenriched public EST and cDNA sequence libraries can also be mined for transcripts that lack significant ORFs, although at some danger of being confused by small ORFs, frameshift sequencing errors, or long untranslated regions of mRNAs.

Third, it should in principle be possible to detect novel transcripts (both ncRNA and protein-coding) using high density oligonucleotide microarrays that systematically probe an entire genome, rather than just probing expression of known and predicted protein-coding genes. However, experience with *E. coli* whole-genome chips has seemed mixed. Successful detection of some known ncRNAs has been shown anecdotally [118], but in systematic use, such data have proved more useful as corroboration rather than a primary screen [20]. I would expect these data to become more useful as microarray technology continues to improve.

The modern RNA world

The discovery of RNA catalysis [119, 120] and the "RNA World" hypothesis for the origin of life [26, 121] provide a seductive explanation for why rRNA and tRNA are at the core of the translation machinery: perhaps they are the frozen evolutionary relic of the invention of the ribosome by an RNA-based "riboorganism" [122]. Other known ncRNAs have also been proposed to be ancient relics of the last riboorganisms [123–125]. The romantic notion of uncovering molecular fossils of a lost RNA world has motivated searches for new ncRNAs. However, as these searches start to succeed, more and more ncRNAs are being found to play apparently well-adapted, specialized biological roles. The idea that ncRNAs are a small and ragged band of relics looks increasingly untenable. The tiny stRNAs and miRNAs, for example, seem highly adapted for a world in which RNAi processing and developmentally regulated messenger RNA targets exist.

Therefore consider an alternative idea, the "modern RNA world". Many of the ncRNAs we see are in fact playing roles where RNA is a more optimal material than protein. Noncoding RNAs are often (though not always) found in roles that involve sequence-specific recognition of another nucleic acid. (The choice of examples in Figures 1, 2, and 4 is deliberate, showing how snoRNAs, miRNAs, and *E. coli* riboregulatory RNAs all function by sequence-specific base complementarity.) RNA, by its very nature, is an ideal material for this role. Base complementarity allows a very small RNA to be exquisitely sequence specific. Evolution of a small specific complementary RNA can be achieved in a single step, just by a partial duplication of a fragment of the target gene into an appropriate context for expression of the new ncRNA.

Many functional roles do not require the more sophisticated catalytic prowess of proteins and could be played by simple RNAs. Posttranscriptional regulation, in particular, can be achieved simply by steric occlusion of sites on a target pre-mRNA or mature RNA. In cases requiring more sophistication than simple steric blockage, necessary catalytic functions can be delegated to a small number of shared proteins, while specific sequence recognition functions are played by a horde of individual small RNAs that interact with those proteins. Morrissey and Tollervey have proposed that modification guide small nucleolar RNAs arose this way, as a more modular system that replaced a smaller number of site-specific protein methylases and pseudouridylases [126].

The idea that ncRNA would be well-adapted for regulatory roles is not new [35, 50, 127]. In the process of defining many of the concepts of molecular genetics, including messenger RNA and operons, François Jacob and Jacques Monod distinguished "structural genes" (like *lacZ*) from "regulatory genes" (like *lacI*) [128]. At that time, regulators like *lacI* had only been defined genetically, and they were known to specifically interact with cis-acting sequences (like *lacO*) either at the DNA or messenger level. Jacob and Monod reasoned that base complementarity would allow RNA to interact highly specifically with other nucleic acid

sequences. They proposed that while structural genes encoded proteins, regulatory genes produced ncRNAs (Figure 5). Forty years later, their proposal is looking more relevant than ever.

Acknowledgements

I thank Tom Tuschl, Victor Ambros, Dave Bartel, Steve Holbrook, and Chris Burge for generously sharing pre-publication results.

References

1. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. New goals for the U.S. human genome project: 1998-2003. *Science* **282**, 682–689 (1998).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Venter, J. C., *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
4. Aparicio, S. A. J. R. How to count ... human genes. *Nat Genet* **25**, 129–130 (2000).
5. Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P., Yang, H. Y., Baer, T., Stredney, D., Spitzner, J., Stutz, A., Krahe, R., and Yuan, B. A draft annotation and overview of the human genome. *Genome Biol.* **2**, research0025.1–0025.18 (2001).
6. Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
7. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239–240 (2000).
8. Ewing, B. and Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25**, 232–234 (2000).
9. Roest Crollius, H., Jaillon, O., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brotier, P., Quétier, F., Saurin, W., and Weissenbach, J. Estimate of human gene number provided by genomewide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genetics* **25**, 235–238 (2000).
10. Eddy, S. R. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699 (1999).
11. Erdmann, V. A., Barciszewska, M. Z., Symanski, M., Hochberg, A., de Groot, N., and Barciszewski, J. The non-coding RNAs as riboregulators. *Nucl. Acids Res.* **29**, 189–193 (2001).
12. Erdmann, V. A., Barciszewska, M. Z., Hochberg, A., de Groot, N., and Barciszewski, J. Regulatory RNAs. *Cell. Mol. Life Sci.* **58**, 960–977 (2001).
13. Olivas, W. M., Muhlrads, D., and Parker, R. Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucl. Acids Res.* **25**, 4619–4625 (1997).

14. Hüttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J. P., and Brosius, J. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20**, 2943–2953 (2001). A screen for novel small non-messenger RNAs, by EST sequencing of size-selected mouse cDNA libraries.
15. Lee, R. C. and Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, in press (2001).
16. Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, in press (2001).
17. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science*, in press (2001).
18. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H., and Altuvia, S. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* **11**, 941–950 (2001). Describes the use of computational prediction of transcriptional promoters and terminators, combined with comparative analysis, to predict putative ncRNA genes in *E. coli*, 14 of which were shown experimentally to express small RNAs.
19. Rivas, E., Klein, R. J., Jones, T. A., and Eddy, S. R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373 (2001). Describes the use of a new general ncRNA genefinding program, which uses comparative genome analysis, to predict structural ncRNA genes in *E. coli*; 11 of these were experimentally shown to produce small noncoding RNA transcripts.
20. Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* **15**, 1637–1651 (2001). Describes the use of comparative genome analysis and microarray expression studies to predict putative ncRNA genes in *E. coli*, 17 of which were experimentally shown to produce small noncoding RNA transcripts.
21. Caspersson, T. Studien über den Eiweißumsatz der Zelle. *Naturwissenschaften* **29**, 33–43 (1941).
22. Brachet, J. and Chantrenne, H. The function of the nucleus in the synthesis of cytoplasmic proteins. *Cold Spring Harbor Symp. Quant. Biol.* **21**, 329–337 (1956).
23. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1**, 59–67 (1955).
24. Crick, F. H. C. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
25. Judson, H. F. *The Eighth Day of Creation: Makers of the Revolution in Biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, (1996).
26. Gesteland, R. F., Cech, T. R., and Atkins, J. F., editors. *The RNA World, Second Edition*. Cold Spring Harbor Laboratory Press, New York, (1999).
27. Brenner, S., Jacob, F., and Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576–581 (1961).
28. Zimmermann, R. A. and Dahlberg, A. E., editors. *Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis*. CRC Press, Boca Raton, (1996).

29. Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Risebrough, R. W., and Watson, J. D. Unstable ribonucleic acid revealed by pulse labeling of *Escherichia coli*. *Nature* **190**, 581–585 (1961).
30. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I., and Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–257 (1958).
31. Soll, D. and RajBhandary, U. L. *tRNA: Structure, Biosynthesis, and Function*. ASM Press, Washington DC, (1995).
32. Zieve, G. W. Two groups of small stable RNAs. *Cell* **25**, 296–297 (1981).
33. Busch, H., Reddy, R., Rothblum, L., and Choi, Y. C. SnRNAs, SnRNPs, and RNA processing. *Ann. Rev. Biochem.* **5**, 617–654 (1982).
34. Yu, Y. T., Scharl, E. C., Smith, C. M., and Steitz, J. A. The growing world of small nuclear ribonucleoproteins. In *The RNA World*, Second Edition, Gesteland, R. F., Cech, T. R., and Atkins, J. F., editors, 487–524. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1999).
35. Lerner, M. R. and Steitz, J. A. Snurps and scyrps. *Cell* **25**, 298–300 (1981).
36. Burge, C. B., Tuschl, T., and Sharp, P. A. Splicing of precursors to mRNAs by the spliceosomes. In *The RNA World*, Second Edition, Gesteland, R. F., Cech, T. R., and Atkins, J. F., editors, 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1999).
37. Tarn, W. Y. and Steitz, J. A. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273**, 1824–1832 (1996).
38. Sharp, P. A. and Burge, C. B. Classification of introns: U2-type or U12-type. *Cell* **91**, 875–879 (1997).
39. Eliceiri, G. L. Small nucleolar RNAs. *Cell Mol. Life Sci.* **56**, 22–31 (1999).
40. Stark, B. C., Kole, R., Bowman, E. J., and Altman, S. Ribonuclease P: an enzyme with an essential RNA component. *Proc. Natl. Acad. Sci. USA* **75**, 3717–3721 (1978).
41. Lewin, R. Surprising discovery with a small RNA. *Science* **218**, 777–778 (1982).
42. Walter, P. and Blobel, G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691–698 (1982).
43. Kelley, R. L. and Kuroda, M. L. Noncoding RNA genes in dosage compensation and imprinting. *Cell* **103**, 9–12 (2000).
44. Avner, P. and Heard, E. X-chromosome inactivation: Counting, choice and initiation. *Nat. Rev. Genet.* **2**, 59–67 (2001).
45. Franke, A. and Baker, B. S. Dosage compensation rox! *Curr. Opin. Cell Biol.* **12**, 351–354 (2000).
46. Tilghman, S. M. The sins of the fathers and mothers: Genomic imprinting in mammalian development. *Cell* **96**, 185–193 (1999).
47. Brannan, C. I. and Bartolomei, M. S. Mechanisms of genomic imprinting. *Curr. Opin. Genet. Dev.* **9**, 164–170 (1999).

48. Meguro, M., Mitsuya, K., Nomura, N., Kohda, M., Kashiwagi, A., Nishigaki, R., Yoshioka, H., Nakao, M., Oishi, M., and Oshimura, M. Large-scale evaluation of imprinting status in the Prader-Willi syndrome region: An imprinted direct repeat cluster resembling small nucleolar RNA genes. *Hum. Mol. Genet.* **10**, 383–394 (2001). Describes an unusual tandem array of a C/D small nucleolar RNA gene in the imprinted Prader-Willi syndrome region of human.
49. Cavaille, J., Bulting, K., Kieffmann, M., Lalonde, M., Brannan, C. I., Horsthemke, B., Bachellerie, J. P., Brosius, J., and Huttenhofer, A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *PNAS* **97**, 14311–14316 (2000). Describes new snoRNAs that apparently do not modify rRNA, show brain-specific expression, and are imprinted genes in the Prader-Willi syndrome region of human, including two different multicopy arrays of snoRNAs HBII-52 and HBII-85.
50. Simons, R. W. and Kleckner, N. Biological regulation by antisense RNA in prokaryotes. *Ann. Rev. Genet.* **22**, 567–600 (1988).
51. Terryn, N. and Rouzé, P. The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci.* **5**, 394–396 (2000).
52. Nemes, J. P., Benzow, K. A., Moseley, M. L., Ranum, L. P., and Koob, M. D. The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). *Hum. Mol. Genet.* **9**, 1543–1551 (2000).
53. Fournier, M. J. and Maxwell, E. S. The nucleolar snRNAs: Catching up with the spliceosomal snRNAs. *Trends Biochem. Sci.* **18**, 131–135 (1993).
54. Maxwell, E. S. and Fournier, M. J. The small nucleolar RNAs. *Ann. Rev. Biochem.* **64**, 897–934 (1995).
55. Balakin, A. G., Smith, L., and Fournier, M. J. The RNA world of the nucleolus: Two major families of small RNAs defined by different box elements with related functions. *Cell* **86**, 823–834 (1996).
56. Kiss-Laszlo, Z., Henry, Y., Bachellerie, J. P., Caizergues-Ferrer, M., and Kiss, T. Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* **85**, 1077–1088 (1996).
57. Nicoloso, M., Qu, L. H., Michot, B., and Bachellerie, J. P. Intron-encoded, antisense small nucleolar RNAs: The characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. *J. Mol. Biol.* **260**, 178–195 (1996).
58. Tycowski, K. T., Smith, C. M., Shu, M. D., and Steitz, J. A. A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in *Xenopus*. *Proc. Natl. Acad. Sci. USA* **93**, 14480–14485 (1996).
59. Ganot, P., Bortolin, M. L., and Kiss, T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell* **89**, 799–809 (1997).
60. Ni, J., Tien, A. L., and Fournier, M. J. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell* **89**, 565–573 (1997).
61. Bachellerie, J. P. and Cavaille, J. Small nucleolar RNAs guide the ribose methylations of eukaryotic rRNAs. In *Modification and Editing of RNA*, Grosjean, H. and Benne, R., editors, 255–272. ASM Press, Washington DC (1998).

62. Lafontaine, D. L. J. and Tollervey, D. Birth of the snoRNPs: The evolution of the modification guide snoRNAs. *Trends Biochem. Sci.* **23**, 383–388 (1998).
63. Weinstein, L. B. and Steitz, J. A. Guided tours: From precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.* **11**, 378–384 (1999).
64. Maden, B. E. H. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucl. Acids Res. Mol. Biol.* **39**, 241–303 (1990).
65. Maden, B. E. H., Corbett, M. E., Heeney, P. A., Pugh, K., and Ajuh, P. M. Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA. *Biochimie* **77**, 22–29 (1995).
66. Ofengand, J. and Bakin, A. Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaeobacteria, mitochondria and chloroplasts. *J. Mol. Biol.* **266**, 246–268 (1997).
67. Lowe, T. M. and Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171 (1998).
68. Barneche, F., Gaspin, C., Guyot, R., and Echeverria, M. Identification of 66 box C/D snoRNAs in *Arabidopsis thaliana*: Extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2'-O-methylation sites. *J. Mol. Biol.* **311**, 57–73 (2001).
69. Liang-Hu, Q., Qing, M., Hui, Z., and Yue-Qin, C. Identification of 10 novel snoRNA gene clusters from *Arabidopsis thaliana*. *Nucl. Acids Res.* **29**, 1623–1630 (2001).
70. Dunbar, D. A., Wormsley, S., Lowe, T. M., and Baserga, S. J. Fibrillarin-associated box C/D small nucleolar RNAs in *Trypanosoma brucei*. Sequence conservation and implications for 2'-O-ribose methylation of rRNA. *J. Biol. Chem.* **275**, 14767–14776 (2000).
71. Omer, A. D., Lowe, T. M., Russell, A. G., Ebhardt, H., Eddy, S. R., and Dennis, P. P. Homologs of small nucleolar RNAs in Archaea. *Science* **288**, 517–522 (2000).
72. Gaspin, C., Cavaille, J., Erauso, G., and Bachellerie, J. P. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: Lessons from the *Pyrococcus* genomes. *J. Mol. Biol.* **297**, 895–906 (2000).
73. Tycowski, K. T., Yao, Z. H., Graham, P. J., and Steitz, J. A. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell* **2**, 629–638 (1998).
74. Yao, Y. T., Shu, M. D., and Steitz, J. A. Modification of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J.* **17**, 5783–5795 (1998).
75. Jády, B. E. and Kiss, T. A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J.* **20**, 541–551 (2001).
76. Ganot, P., Jady, B. E., Bortolin, M. L., Darzacq, X., and Kiss, T. Nucleolar factors direct the 2'-O-ribose methylation and pseudouridylation of U6 spliceosomal RNA. *Mol. Cell. Biol.* **19**, 6906–6917 (1999).

77. Jady, B. E. and Kiss, T. Characterisation of the U83 and U84 small nucleolar RNAs: Two novel 2'-O-ribose methylation guide RNAs that lack complementarities to ribosomal RNAs. *Nucl. Acids Res.* **28**, 1348–1354 (2000).
78. Kiss, T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.* **20**, 3617–3622 (2001).
79. de los Santos, T., Schweizer, J., Rees, C. A., and Francke, U. Small evolutionarily conserved RNA, resembling C/D box small nucleolar RNA, is transcribed from PWCR1, a novel imprinted gene in the Prader-Willi deletion region, which is highly expressed in brain. *Am. J. Hum. Genet.* **67**, 1067–1082 (2000). Describes the identification of the imprinted PWCR1 locus in the human Prader-Willi region, which is an array of about 24 copies of a C/D snoRNA; appears to be the same locus called DR by Meguro *et al.* and HBII-85 by Cavaille *et al.*
80. Filipowicz, W. Imprinted expression of small nucleolar RNAs in brain: Time for RNomics. *Proc. Natl. Acad. Sci. USA* **97**, 14035–14037 (2000).
81. Horvitz, H. R. and Sulston, J. E. Isolation and genetic characterization of cell-lineage mutants of the nematode *Caenorhabditis elegans*. *Genetics* **96**, 435–454 (1980).
82. Lee, R. C., Feinbaum, R. L., and Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993). A case study of the careful positional cloning of a small RNA gene identified by a genetic screen.
83. Wightman, B., Ha, I., and Ruvkun, G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862 (1993).
84. Moss, E. G., Lee, R. C., and Ambros, V. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* **88**, 637–646 (1997).
85. Olsen, P. H. and Ambros, V. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol* **216**, 671–680 (1999).
86. Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
87. Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
88. Hunter, C. P. Gene silencing: Shrinking the black box of RNAi. *Curr. Biol.* **10**, R137–R140 (2000).
89. Carthew, R. W. Gene silencing by double-stranded RNA. *Curr. Opin. Cell Biol.* **13**, 244–248 (2001).
90. Sharp, P. A. RNA interference – 2001. *Genes Dev.* **15**, 485–490 (2001).
91. Vance, V. and Vaucheret, H. RNA silencing in plants—defense and counterdefense. *Science* **292**, 2277–2280 (2001).

92. Hamilton, A. J. and Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999).
93. Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T., and Zamore, P. D. A cellular function for the RNA-interference enzyme dicer in the maturation of the *let-7* small temporal RNA. *Science* **293**, 834–838 (2001). One of the key results that led to the discovery of miRNAs: the enzyme responsible for processing siRNAs is also responsible for processing the endogenous human *let-7* regulatory RNA.
94. Elbashir, S. M., Lendeckel, W., and Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001). An excellent biochemical study, showing that double-stranded RNA is processed by an RNase III like enzyme to make 21-22 nt RNAs (small interfering RNAs, siRNAs) that produce the RNA interference effect.
95. Moss, E. G. Noncoding RNAs: Lightning strikes twice. *Curr. Biol.* **10**, R436–R439 (2000).
96. Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G., and Mello, C. C. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001). Like the Hutvagner *et al.* paper, showed that Dicer is apparently responsible for processing *lin-4* and *let-7* in *C. elegans*; moreover, knockdowns of genes in the large *rde-1/ARGONAUTE/Piwi* family show that developmental defects and defects in RNAi processing are separable, depending on different proteins in this family.
97. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
98. Wassarman, K. M., Zhang, A., and Storz, G. Small RNAs in *Escherichia coli*. *Trends Microbiol.* **7**, 37–45 (1999).
99. Lease, R. A., Cusick, M. E., and Belfort, M. Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl. Acad. Sci. USA* **95**, 12456–12461 (1998).
100. Lease, R. A. and Belfort, M. Riboregulation by DsrA RNA: Trans-actions for global economy. *Mol. Micro.* **38**, 667–672 (2000).
101. Rivas, E. and Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, in press (2001).
102. Altuvia, S., Zhang, A., Argaman, L., Tiwari, A., and Storz, G. The *Escherichia coli* OxyS regulatory RNA represses *fhlA* translation by blocking ribosome binding. *EMBO J.* **17**, 6069–6075 (1998).
103. Romeo, T. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Mol. Microbiol.* **29**, 1321–1330 (1998).
104. Carter, R. J., Dubchak, I., and Holbrook, S. R. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.* **29**, 3928–3938 (2001).
105. McKusick, V. A., Eldridge, R., Hostetler, J. A., Ruangwit, U., and Egeland, J. A. Dwarfism in the Amish. II. cartilage-hair hypoplasia. *Bull. Johns Hopkins Hosp.* **116**, 285–326 (1965).

106. Sulisalo, T., Sistonen, P., Hastbacka, J., Wadelius, C., Makitie, O., de la Chapelle, A., and Kaitila, I. Cartilage-hair hypoplasia gene assigned to chromosome 9 by linkage analysis. *Nat. Genet* **3**, 338–341 (1993).
107. Ridanpää, M., van Eenennaam, H., Pelin, K., Chadwick, R., Johnson, C., Yuan, B., vanVenrooij, W., Pruijn, G., Salmela, R., Rockas, S., Mäkitie, O., Kaitila, I., and de la Chapelle, A. Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell* **104**, 195–203 (2001).
108. Clayton, D. A. A big development for a small RNA. *Nature* **410**, 29–31 (2001).
109. Vulliamy, T., Marrone, A., Goldman, F., Dearlove, A., Bessler, M., Mason, P. J., and Dokal, I. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature* **413**, 432–435 (2001).
110. González-Pastor, J. E., San Millán, J. L., and Moreno, F. The smallest known gene. *Nature* **369**, 281 (1994).
111. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17**, 425–428 (2001).
112. Murata, H., Chatterjee, A., Liu, Y., and Chatterjee, A. K. Regulation of the production of extracellular pectinase, cellulase, and protease in the soft rot bacterium *Erwinia carotovora* subsp. *carotovora*: Evidence that aepH of *E. carotovora* subsp. *carotovora* 71 activates gene expression in *E. carotovora* subsp. *carotovora*, *E. carotovora* subsp. *atroseptica*, and *Escherichia coli*. *Appl. Environ. Microbiol.* **60**, 3150–3159 (1994).
113. Crespi, M. D., Jurkevitch, E., Poiret, M., d’Aubenton Carafa, Y., Petrovics, G., Kondorosi, E., and Kondorosi, A. *enod40*, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *EMBO J.* **13**, 5099–5112 (1994).
114. Sousa, C., Johansson, C., Charon, C., Manyani, H., Sautter, C., Kondorosi, A., and Crespi, M. Translational and structural requirements of the early nodulin gene *enod40*, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. *Mol. Cell. Biol.* **21**, 354–366 (2001).
115. Rivas, E. and Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **6**, 583–605 (2000).
116. Workman, C. and Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.* **27**, 4816–4822 (1999).
117. Lim, L. P. and Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA*, in press (2001).
118. Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J., and Church, G. M. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature Biotech.* **18**, 1262–1268 (2000).
119. Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**, 147–157 (1982).

120. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
121. Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
122. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
123. Benner, S. A., Ellington, A. D., and Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proc. Natl. Acad. Sci. USA* **86**, 7054–7058 (1989).
124. Jeffares, D. C., Poole, A. M., and Penny, D. Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36 (1998).
125. Poole, A. M., Jeffares, D. C., and Penny, D. The path from the RNA world. *J. Mol. Evol.* **46**, 1–17 (1998).
126. Morrissey, J. P. and Tollervey, D. Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system. *Trends Biochem. Sci.* **20**, 78–82 (1995).
127. Caprara, M. G. and Nilsen, T. W. RNA: versatility in form and function. *Nat. Struct. Biol.* **7**, 831–833 (2000).
128. Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
129. Ha, I., Wightman, B., and Ruvkun, G. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes Dev.* **10**, 3041–3050 (1996).

ncRNA	noncoding RNA - all RNAs other than messenger RNA [13]
fRNA	functional RNA - essentially synonymous with ncRNA [104]
snmRNA	small non-messenger RNA - essentially synonymous with small ncRNAs [14]
rRNA	ribosomal RNA
tRNA	transfer RNA
snRNA	small nuclear RNA - includes splicesomal RNAs
snoRNA	small nucleolar RNA - most known ones are involved in rRNA modification
stRNA	small temporal RNA - <i>lin-4</i> and <i>let-7</i> in <i>C. elegans</i>
siRNA	small interfering RNA - active molecules in RNA interference
miRNA	micro RNA - putative translational regulatory gene family

Table 1: Some of the abbreviations used in the literature for different classes and families of noncoding RNA.

Figure legends

Figure 1.

Diagrams of modification guide snoRNAs targeting rRNA bases. A: C/D snoRNAs use antisense complementarity to target RNA for 2'-O-ribose methylation (site marked with 'm' and black dot). B: H/ACA snoRNAs use antisense complementarity within an interior loop to target RNA for pseudouridylation (site marked Ψ). Redrawn from [78].

Figure 2.

Examples of proposed interactions between the *C. elegans lin-4* miRNA and a target mRNA. *lin-4* is proposed to interact by base pairing with seven sites in the 3' UTR of *lin-14* mRNA (A,B show the first two of the seven sites) [129] and one site in the 3' UTR of *lin-28* mRNA (C) [84]. A C residue (in bold) is predicted to be bulged in 4 of the 7 *lin-14* interactions, including the two shown; this C is mutated to U in the strong loss of function *lin-4 ma161* allele [82].

Figure 3.

Three examples of miRNAs. Proposed structure of the precursor stem is shown, with residues in the mature miRNA shown in red. Comparison of *C. elegans miR-1* [15, 16] with *D. melanogaster miR-1* [17] shows perfect conservation of the mature miRNA (except for length variability at the 3' end). Comparison of *miR-1* with *miR-84* [16] shows an example of how mature miRNAs are produced asymmetrically from either side of the precursor stem.

Figure 4.

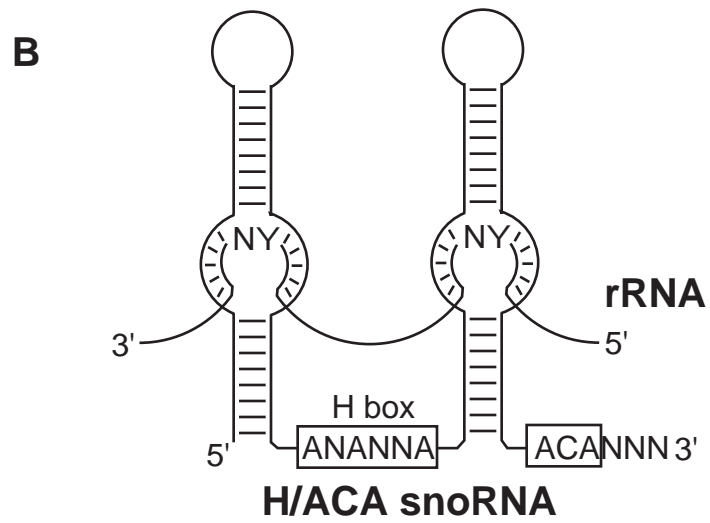
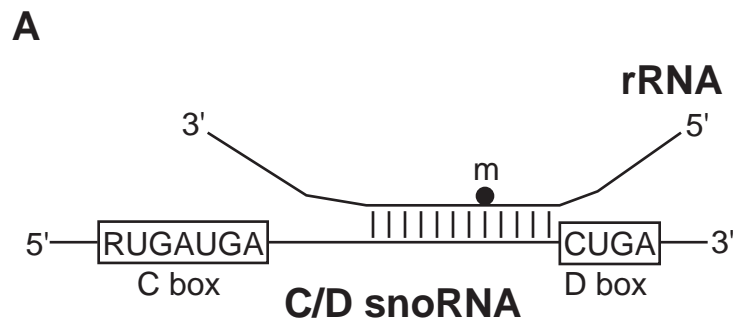
Example of an *E. coli* riboregulator. The DsrA RNA is proposed to have a three-stem secondary structure on its own (A), but unzips to interact by base pairing with the translational start site of several different mRNAs, including *rpoS* (panel B; the Shine/Dalgarno and AUG of the start site are boxed). In some mRNAs, DsrA blocks the ribosome binding site and acts as a translational repressor. For *rpoS*, DsrA acts as a translational activator, which is proposed to happen by competition with an occluding secondary structure as shown in panel B; base pairs that would be broken by interaction with bases in DsrA (shown in bold), freeing the start site, are shown with open circles. Redrawn from [99].

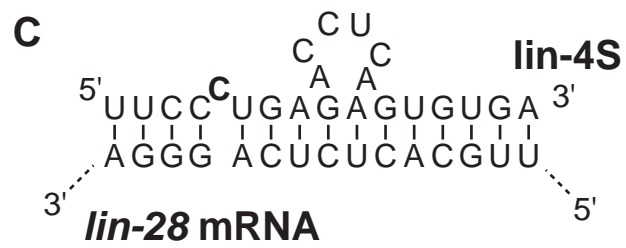
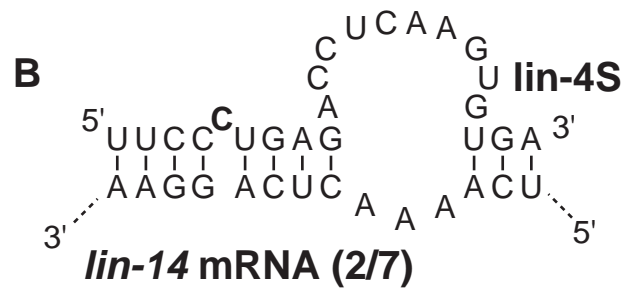
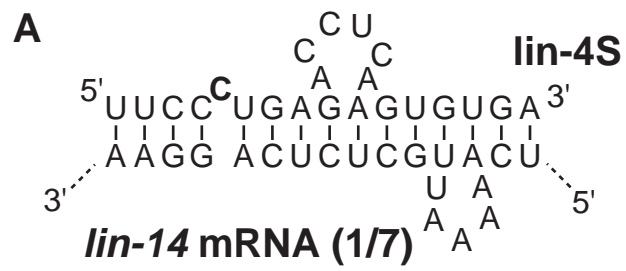
Figure 5.

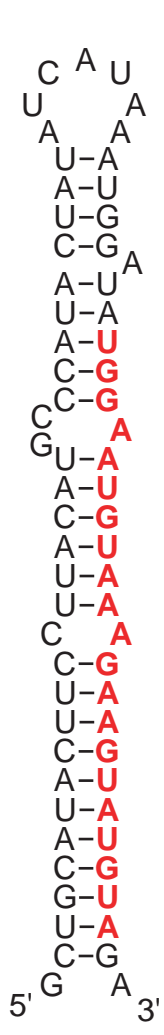
Figure 6 from Jacob and Monod's 1961 paper [128], showing their proposal that structural genes produce messenger RNAs that code for protein, but regulatory genes produce regulatory RNAs (note the wavy line) that interact by base pairing with operators, either at the transcriptional level (Model I) or the posttranscriptional level (Model II).

Summary

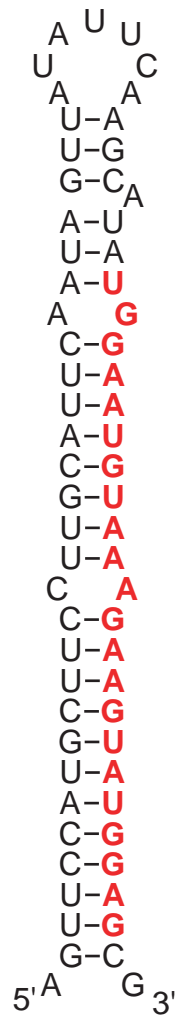
- Although some textbooks still talk about only three types of RNA (ribosomal RNA, transfer RNA, and messenger RNA), many other noncoding RNA species have been isolated anecdotally.
- Systematic screens looking for more noncoding RNA genes have been undertaken recently.
- Two large families of small nucleolar RNAs (snoRNAs) are involved in directing site-specific modifications of ribosomal RNAs and other RNAs, and the number of known genes in these families continues to grow rapidly.
- Three recent papers describe a new large eukaryotic RNA gene family, the microRNAs (miRNAs), tiny 21-24 nt RNAs which are probably acting as translational regulators of protein-coding mRNAs.
- Four recent papers describe screens for new noncoding RNA genes in *E. coli*, leading to the experimental confirmation of over 30 new noncoding transcripts of as yet unknown function, and the computational prediction of many more.
- RNA genes have been thought of as rare relics of a primordial “RNA World” that has mostly been replaced by more efficient proteins. Now, though, it seems that ncRNAs may be numerous and highly adapted in their roles in modern organisms.
- RNA is particularly well suited for the job of specific recognition of other RNAs by complementary base pairing. Evolution may favor ncRNAs instead of proteins in certain roles, for instance as posttranscriptional regulatory molecules that interact with specific mRNAs.



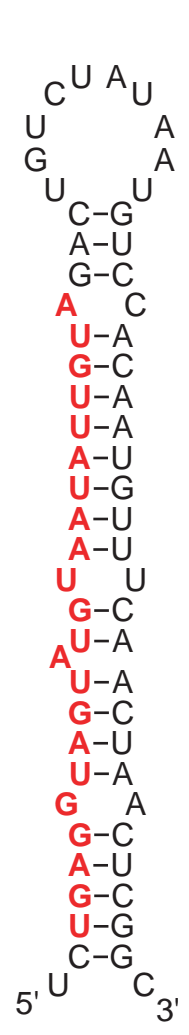




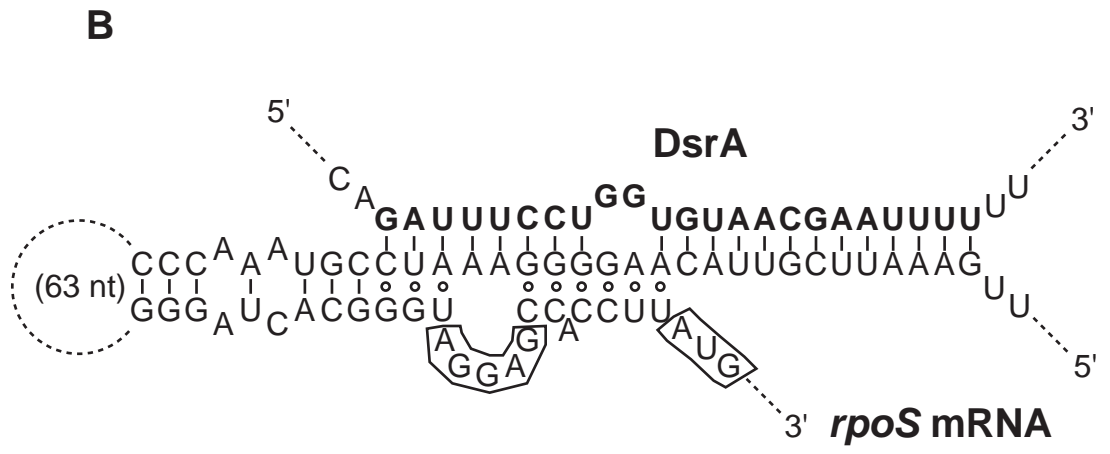
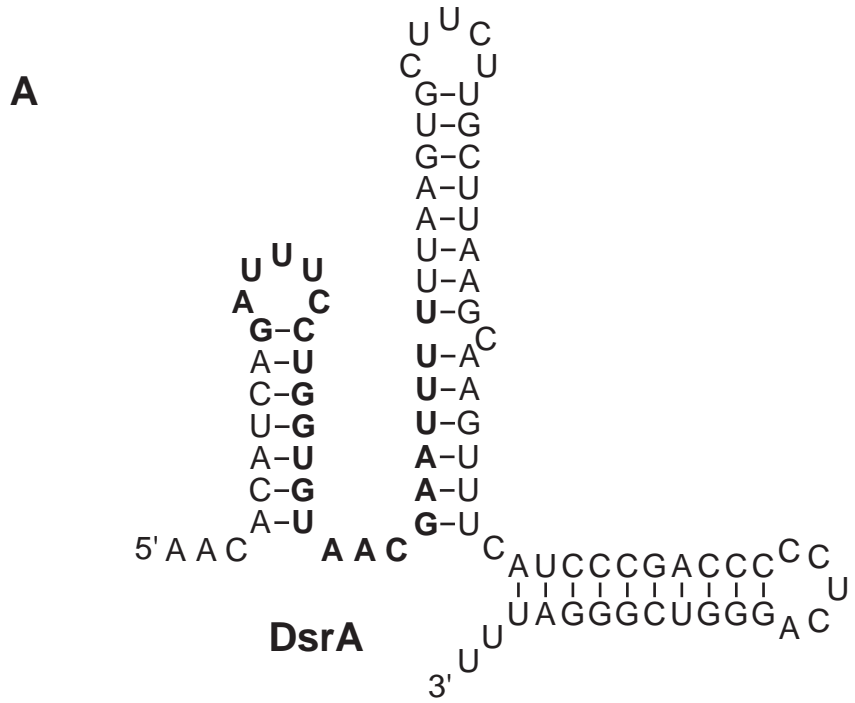
C. elegans miR-1

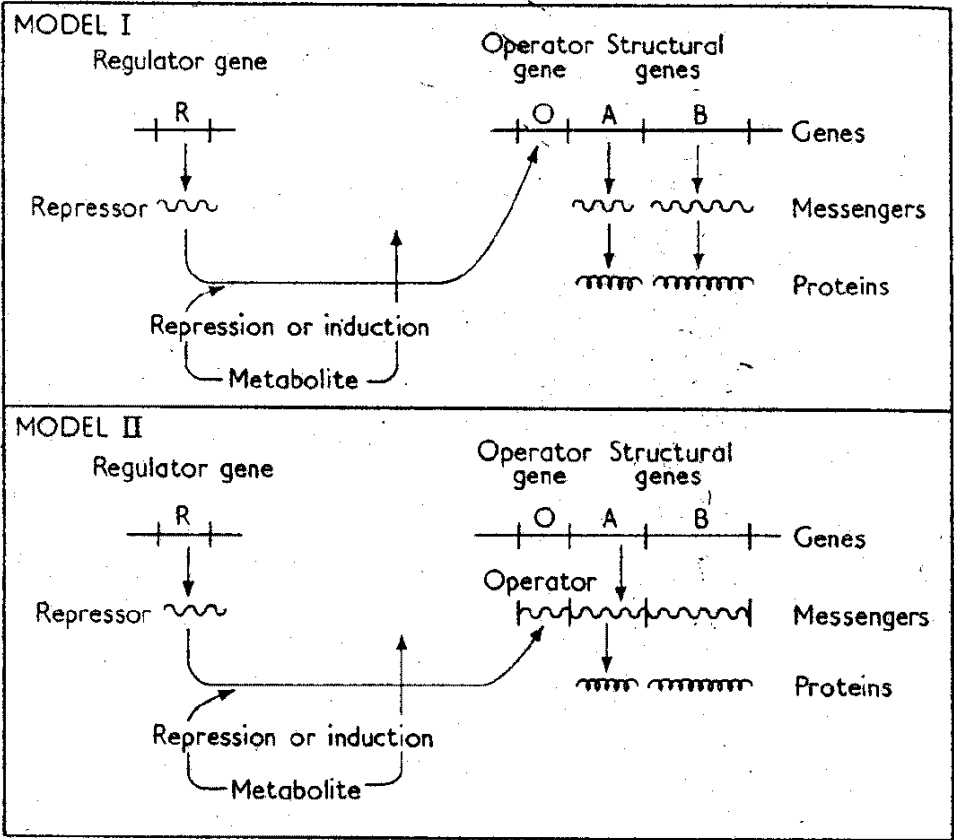


D. melanogaster miR-1



C. elegans miR-84





Author biography

Sean R. Eddy is the Alvin Goldfarb Distinguished Professor of Computational Biology and an HHMI assistant investigator in the Department of Genetics at Washington University School of Medicine in Saint Louis, where he is also affiliated with the WU Genome Sequencing Center. He is the author of the HMMER software for biological sequence analysis, a coauthor of the Pfam protein domain database, and a coauthor of the book *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998). His current research interests are in the development of algorithms for discovering and analyzing noncoding RNA genes.

Glossary

Cajal bodies (also known as coiled bodies). Nuclear organelles of unknown function named in honour of Ramón y Cajal.

Heterochronic mutation A mutation that alters the timing of developmental events, such as the sequence of larval molts in nematodes.

Nucleolus A highly organized nuclear organelle that is the site of ribosomal RNA processing and ribosome assembly.

Neural network A popular machine learning method that is often used for automatic classification of biological sequences, based on “training” on a set of known examples.

Neutral drift The process whereby DNA sequence acquires many mutations over time that have no phenotypic effect, and hence are not acted upon by Darwinian selection.

Positive selection A rare form of evolutionary change in which a mutation appears to be favored because it is fixed in the population at a rate even greater than predicted by neutral drift.

Purifying selection A common form of evolutionary change in which a mutation is harmful, and therefore disappears from the population.

Ribonuclease P A universally conserved enzyme that cleaves a leader sequence off of tRNA precursors.

RNA processing A general for the maturation of a precursor RNA; includes the processes of RNA splicing, RNA modification, RNA editing, and RNA cleavage.

RNA tailing A technique in which an artificial homopolymer sequence is enzymatically added to an RNA to facilitate molecular cloning, as opposed to relying on the presence of a natural poly-A tail.

Shine/Dalgarno sequence A consensus sequence recognized during translational initiation by *E. coli* ribosomes.

Signal recognition particle An RNA/protein complex involved in exporting secreted proteins from the cell.

Sm An RNA-binding protein recognized by antibodies produced by people with certain autoimmune diseases; “Sm” stands for “Smith”, the name of a patient.

U RNAs Small nuclear RNAs in eukaryotes. The first such RNAs that were found were rich in uridine (U), and the name stuck.