NON-CODING RNA GENES AND THE MODERN RNA WORLD

Sean R. Eddy

Non-coding RNA (ncRNA) genes produce functional RNA molecules rather than encoding proteins. However, 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. Non-coding RNAs seem to be 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.

One goal of genome projects is to systematically identify genes[1]. In the past year, two papers have announced drafts of the human genome sequence[2,3], but the estimated number of human genes continues to fluctuate. Current estimates centre on 30,000–40,000 genes, 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 expressed sequence tag (EST) sequencing of polyadenylated mRNAs[7,8]; identification of conserved coding exons 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 underestimate the number of other genes. They essentially do not work at all for one class of genes — the non-coding RNA (ncRNA) genes, which produce transcripts that function directly as structural, catalytic or regulatory RNAs, rather than expressing mRNAs that encode proteins[10–12] (see BOX 1 for a list of abbreviations that are used to describe classes of RNA). Knowledge of ncRNAs has been limited to biochemically abundant species and anecdotal discoveries. Even after the completion of many genome sequences, both the number and diversity of ncRNA genes remain largely unknown.

Could it be possible that a large class of genes has gone relatively undetected because they do not 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, new systematic gene-discovery approaches need to be developed that are specifically aimed at ncRNAs. A pioneering study by Roy Parker's group found a few new RNA genes and small open reading frames (ORFs) in the yeast genome by doing northern blots that probed for expressed transcripts in 'grey holes' (suspiciously large intergenic regions), and by searching for consensus RNA polymerase III 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-mRNAs[14]; specially designed cDNA cloning screens for a new regulatory RNA gene family of tiny RNAs called microRNAs (miRNAs)[15–17]; and general ncRNA gene-finding exercises using computational comparative genomics in *Escherichia coli*[18–20]. The results of these screens are startling. All of them indicate that the prevalence of ncRNA genes has indeed been underestimated.

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

Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. e-mail: eddy@genetics.wustl.edu

Box 1 | Abbreviations for different classes of non-coding RNA

- **fRNA**
  Functional RNA — essentially synonymous with non-coding RNA[104]
- **miRNA**
  MicroRNA — putative translational regulatory gene family
- **ncRNA**
  Non-coding RNA — all RNAs other than mRNA[13]
- **rRNA**
  Ribosomal RNA
- **siRNA**
  Small interfering RNA — active molecules in RNA interference
- **snRNA**
  Small nuclear RNA — includes spliceosomal RNAs
- **snmRNA**
  Small non-mRNA — essentially synonymous with small ncRNAs[14]
- **snoRNA**
  Small nucleolar RNA — most known snoRNAs are involved in rRNA modification
- **stRNA**
  Small temporal RNA — for example, *lin-4* and *let-7* in *Caenorhabditis elegans*
- **tRNA**
  Transfer RNA

The lessons of history

The central role of RNA in translation. It was clear by the 1950s that although DNA was located in the eukaryotic nucleus, 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 called ribosomes[24]. James Watson sketched the "central dogma" as early as 1952 (REFS 25,26), imagining that there must be a coding RNA that is 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 mRNA that was associated with general ribosomal proteins that catalysed translation. Various results undermined this hypothesis, including the simple observation that although 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 rRNAs[28], and programmed with various unstable mRNAs that are only a small fraction of the total RNA population[27,29].

The second class of functional RNA was predicted by Francis 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 not only that the adaptor would be an RNA, but also 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 Mahlon Hoagland and co-workers[30]. These RNAs later proved to be Crick's adaptors — the transfer RNAs[31].

RNA therefore changed from being thought of as having one 'flavour' (the purely information-carrying intermediate in the "central dogma") to having three flavours, 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 flavours. Several abundant, small non-mRNAs, other than rRNA and tRNA, were detected and isolated biochemically, among them the uridine (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, such as 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, namely U1, U2, U4, U5 and U6 small nuclear RNA (snRNA), turned out to be components of the spliceosome, involved in splicing mRNAs[34,36]. 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 have been 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 examples of such a surprise resulted in the renaming of the signal recognition 'protein' to the SIGNAL RECOGNITION 'PARTICLE' (SRP), when it was unexpectedly found to contain a 7S RNA that is now called SRP-RNA[41,42].

New RNAs continue to appear; among the more fascinating stories is the discovery that RNAs have roles in chromatin structure[43]. A canonical example is the human *XIST* (X(inactive)-specific transcript) RNA, a 17-kb ncRNA with a key role in dosage compensation and X-chromosome inactivation[44]. *Drosophila melanogaster* also seems to control dosage compensation using small chromatin-associated *roX* (RNA on the X) RNAs[45]. Several large ncRNAs have been found to be expressed from imprinted regions of vertebrate chromosomes, including the *IPW* (imprinted in Prader–Willi syndrome) and *H19* (H19, imprinted maternally expressed untranslated mRNA) transcripts[46,47]. (The imprinted Prader–Willi crucial region seems to be especially rich in ncRNAs[48,49], although it is unclear whether this is peculiar, or simply due to the incredibly intense gene hunting in search of the elusive cause of

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

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

RIBONUCLEASE P
A universally conserved enzyme that cleaves a leader sequence from tRNA precursors.

SIGNAL RECOGNITION PARTICLE
An RNA–protein complex involved in exporting secreted proteins from the cell.

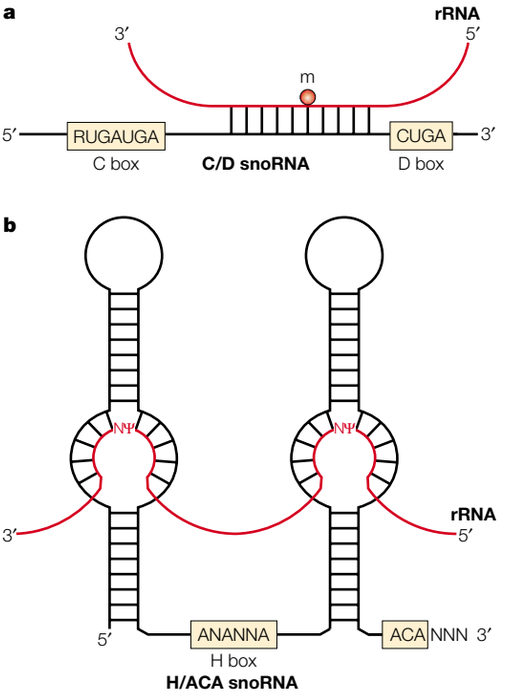

Figure 1 | **Diagrams of snoRNAs guiding modification to target rRNA bases. a** | C/D box small nucleolar RNAs (snoRNAs) use antisense complementarity to target RNA for 2′-*O*-ribose methylation (site marked with 'm' and red dot). R stands for A or G (purine). **b** | H/ACA box snoRNAs use antisense complementarity in an interior loop to target RNA for pseudouridylation (site marked 'NΨ'). Redrawn with permission from REF. 78.

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 those in imprinting and chromatin structure. Mutations in one *cis*-antisense RNA in humans — *SCA8* (spinocerebellar ataxia 8) — are found in patients with spinocerebellar ataxia[52].

Continued flurries of small nucleolar RNAs

The nucleolus is rich in snoRNAs, most of which are ~70–250 nucleotides in length[53,54]. Some snoRNAs have roles in ribosomal RNA PROCESSING, but most function in rRNA modification[39]. On the basis of 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], whereas the H/ACA snoRNAs use base complementarity to guide site-specific pseudouridylations to rRNA[59,60] (FIG. 1). In both cases, the catalytic function seems to be provided by a protein methylase or pseudo-U synthetase associated with the snoRNA, and the specificity for the target base on the rRNA is provided by base complementarity to the snoRNA[61–63].

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 ~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 more than 60 new C/D snoRNAs in the *Arabidopsis thaliana* 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 the mouse (see below)[14]. Numerous homologues of the C/D snoRNAs have been found in the Archaea[71,72], in which they are presumed to have the same function in guiding specific 2′-*O*-ribose methylations of target RNAs.

In addition to rRNA, other structural RNAs — such as tRNAs and snRNAs — are known to be extensively modified[33,73,74], and it now seems that some, if not many, of these modifications are also guided by snoRNA. At least one of the 2′-*O*-ribose methylations of *Xenopus laevis* U6 snRNA is guided by the C/D snoRNA, mgU6-77 (REF. 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]. Furthermore, 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], for which 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-rRNAs are transported through the nucleolus, or whether perhaps there is at least one more site of RNA modification in the cell. Recent evidence indicates that the snRNA modifications are associated with CAJAL BODIES (coiled bodies) in the nucleus[78].

An EST screen for small non-mRNAs. Alexander Hüttenhofer and colleagues[14] undertook a general screen for new, small non-mRNAs, using an EST sequencing approach. The RNA population used was total (not cytoplasmic) mouse brain RNA that was cloned by RNA TAILING (not by poly-A selection and dT priming) and size selected in two small RNA fractions — 50–100 nucleotides and 110–500 nucleotides. High-throughput filter hybridization was used to screen out clones that correspond to tRNA, rRNA fragments and other known ncRNAs, increasing the fraction of new ncRNA sequences from ~3–7% in an unscreened library to ~20–22% after screening. A total of ~5,000 clones were sequenced, and after accounting for several sequences of the same RNA species, 201 new RNA sequences were identified.

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

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

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.

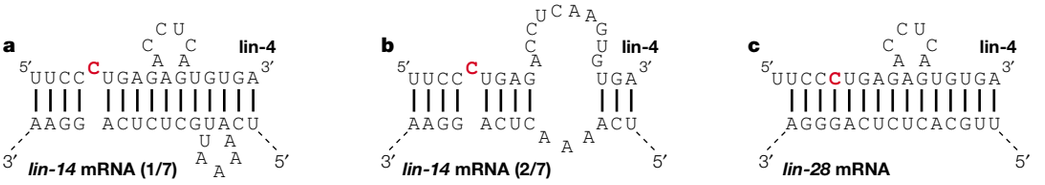

Figure 2 | **Examples of proposed interactions between the *Caenorhabditis elegans lin-4* microRNA and a target mRNA.** *lineage-4* (*lin-4*) is proposed to interact by base pairing with **a**,**b** | seven sites in the 3′ untranslated region (UTR) of *lin-14* mRNA (first two of the seven sites are shown)[129] and **c** | one site in the 3′ UTR of *lin-28* mRNA[84]. A C residue (in red) is predicted to be bulged in four out of the seven *lin-14* interactions, including the two shown; this C is mutated to U in the strong loss-of-function *lin-4 ma161* allele[82].

A few more than half of the new sequences seem 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 crucial region for Prader–Willi syndrome, two of which (*HBII-52* and *HBII-85*) are C/D snoRNAs found in multicopy tandem arrays, unlike most vertebrate snoRNAs, which are found in single copies in the introns of other genes. Both *HBII-52* and *HBII-85* 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 non-coding *IPW* gene, was also detected as an imprinted ncRNA gene array by other studies[48,79]. The *HBII-52* snoRNA has a perfect 18-bp complementarity to 5-hydroxytryptamine 2C (5-HT_{2C}) receptor mRNA, and is predicted on that basis to methylate a site of known mRNA editing; this indicates a complex set of interactions in which a snoRNA might regulate the editing of an mRNA transcript[14,80].

Out of the 88 sequences that did not seem 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 nucleotides. The functions of these 20 new small RNAs are unknown. Hüttenhofer and coworkers are now analysing similar libraries from *Caenorhabditis elegans*, *D. melanogaster* and *A. thaliana*.

MicroRNAs: one, two ... infinity?

One: lin-4. A canonical example of the identification of a ncRNA gene by genetics is the story of the *lin-4* regulatory RNA in the nematode *C. elegans*. The *lin-4* locus was identified in a screen for mutations that affect the timing and sequence of postembryonic development (HETEROCHRONIC MUTATIONS) in *C. elegans*[81]. Mutant animals reiterate the L1 larval stage rather than progress 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 paper by Rosalind Lee and colleagues dryly recounts a careful detective story, as Victor Ambros's lab gradually realized that they were dealing not with a protein-coding gene, but with a tiny ncRNA. The *lin-4* gene product is a 22-nucleotide RNA, processed from a 61-nucleotide 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′ untranslated regions (UTRs) of the target genes have short stretches of complementarity to *lin-4* (REFS 82–84; FIG. 2). Deletion of these apparent *lin-4* target sequences causes 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, *lethal-7* (*let-7*), also mapped to a ncRNA gene with a 21-nucleotide product[86]. The small *let-7* RNA is also thought to be a post-transcriptional 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, Amy Pasquinelli *et al.*[87] showed that *let-7* was almost 100% conserved and expressed as a small 21-nucleotide RNA in all bilaterally symmetrical animals that were tested, including human, mouse, chicken, polychaete worms and flies, but not in cnidarians (jellyfish) or poriferans (sponges). The function of these *let-7* homologues is unknown, but because they show temporal regulation that is generally similar to the developmental pattern of *let-7* in the worm, one presumes that they also function in post-transcriptional regulation of developmental genes. Pasquinelli *et al.* proposed the name "small temporal RNAs" (stRNAs) for genes such as *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]. The 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, in which the effect has been called co-suppression or post-transcriptional gene silencing[89,91].

HETEROCHRONIC MUTATION
A mutation that alters the timing of developmental events, such as the sequence of larval moults in nematodes.

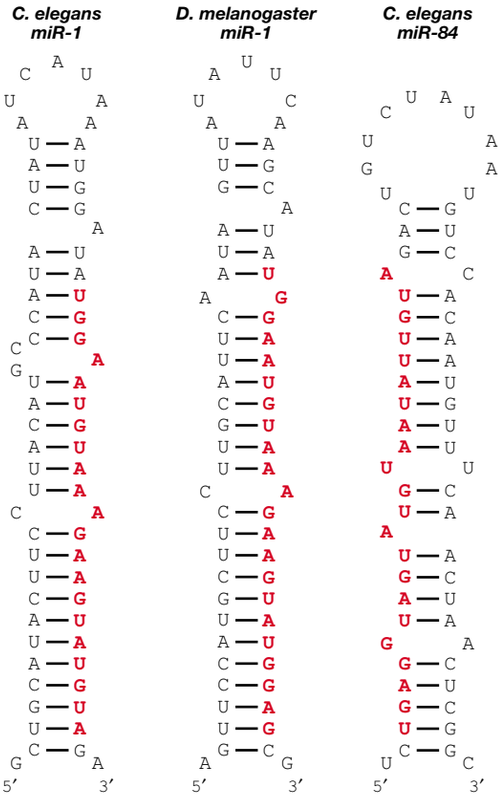

Figure 3 | **Three examples of microRNAs.** Proposed structure of the precursor stem is shown, with residues in the mature microRNA (miRNA) shown in red. Comparison of *Caenorhabditis elegans miR-1* (REFS 15,16) with *Drosophila melanogaster miR-1* (REF. 17) shows perfect conservation of the mature miRNA (except for length variability at the 3′ end). Comparison of *miR-1* with *miR-84* (REF. 16) shows an example of how mature miRNAs are produced asymmetrically from either side of the precursor stem.

The input dsRNA is cleaved to form the active agents of the RNAi effect — tiny 21–25-nucleotide small interfering RNAs (siRNAs)[92–94]. Several proteins that are 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 nucleotides, respectively) are the same as those of the 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-nucleotide unprocessed human *let-7* precursor[93]. Knocking down either the function of the *C. elegans* Dicer homologue or 2 of the 23 worm homologues of the *rde-1/ARGONAUTE/piwi* gene family — *alg-1* (*argonaute-like gene 1*) and *alg-2* — results in accumulation of unprocessed *lin-4* and *let-7* precursors[96]. In the course of cloning and analysing the small RNAs produced from an exogenous dsRNA, Thomas Tuschl's lab noted in passing that *Drosophila* seemed to contain endogenous 21- and 22-mers[94], and suggested that perhaps there were naturally occurring siRNAs.

Introducing the microRNAs. Now, three papers show 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 miRNAs. Nelson Lau *et al.*[16] produced and sequenced a *C. elegans* cDNA library that was cleverly enriched for tiny RNAs with 5′-phosphate and 3′-hydroxy termini, and obtained 55 new miRNAs. Lee and Ambros used a size-selected *C. elegans* cDNA library and, to a lesser extent, a computational approach, to look for conserved sequences in *Caenorhabditis briggsae* that can be folded into a stem similar to the *lin-4* and *let-7* precursors, and found 15 miRNAs[15]. Mariana Lagos-Quintana *et al.*[17] used size-selected cDNA libraries in human and *Drosophila* to isolate 33 miRNAs — 19 in humans and 14 in *Drosophila*.

In total, 91 different miRNAs have been identified so far in the three species. Some of these are highly conserved in evolution, such as *let-7*, and homologues of 11 miRNAs are found in more than one of the three species. Northern blot analyses have been done for many of these RNAs, and generally show both a 21–24-nucleotide form (presumably the active miRNA) and, often, a less abundant ~70-nucleotide form (presumably the precursor stem–loop). The miRNA genes are often clustered in the genome[16,17] and might 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 seems 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 and, almost invariably, only one strand of the precursor stem can be recovered as a 21–24-nucleotide product, although the Bartel lab reports a single exception[16] (FIG. 3). Many miRNAs are produced in a stage- and/or tissue-specific manner, indicating possible roles in development akin to the stRNAs. Some of the *C. elegans* miRNAs are specifically expressed in the germ line and embryo, in which 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 the 3′ UTRs of coding mRNAs (FIG. 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 anti-virus

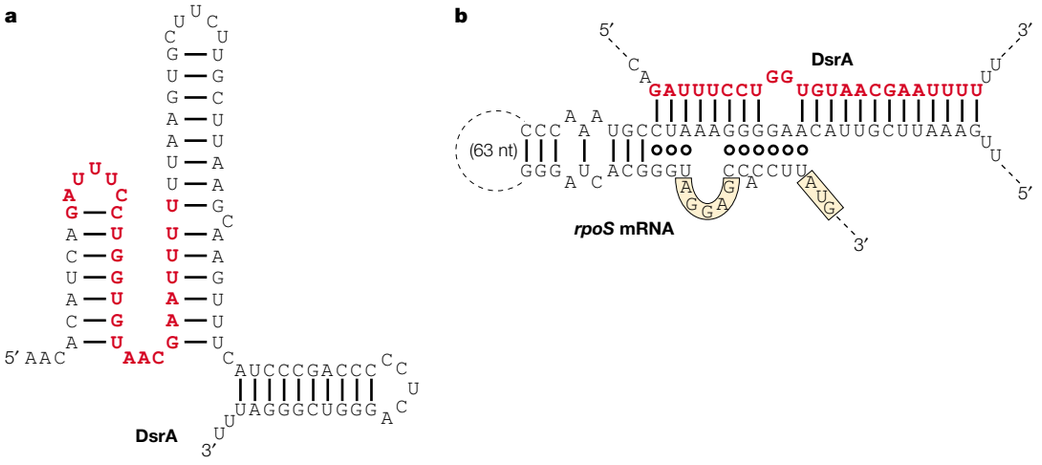

Figure 4 | **Example of an *Escherichia coli* riboregulator. a** | The *dsrA* RNA is proposed to have a three-stem secondary structure on its own, but **b** | unzips to interact by base pairing with the translational start site of several different mRNAs, including *rpoS* (the SHINE/DALGARNO and AUG 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 **b**; base pairs that would be broken by interaction with bases in DsrA (shown in red), freeing the start site, are shown with bold, open circles. nt, nucleotide. Redrawn with permission from REF. 99.

defence 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]. Therefore, it seems that in addition to the RNAi effect itself, components of the RNAi-processing pathway also function in developmental regulatory processes that might involve numerous endogenous miRNAs.

Even *E. coli* has been hiding something

The bacterium *E. coli* is arguably the best-studied organism. The complete genome sequence of *E. coli* K-12 contains an estimated 4,200 protein-coding genes[97]. The small number of known ncRNA genes has continued to climb slowly, as several small, stable RNAs have been reported anecdotally[98]. Many of these seem to be 'riboregulators'[99,100], which act by using base complementarity to specifically interact with translational start sites and either repress or, more rarely, activate translation (FIG. 4). The recent availability of comparative genome sequence information from *Salmonella* spp. and other enterobacteria made it possible to go looking for conserved sequences that might correspond to ncRNAs, instead of coding ORFs.

Liron Argaman and colleagues computationally analysed intergenic regions to identify loci that have a predicted promoter and terminator spaced 50–400 nucleotides 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 nucleotides 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 were preferentially expressed in stationary phase; one was expressed almost exclusively in cold-shocked cells; and one was preferentially expressed in minimal media.

Karen Wassarman *et al.*[20] looked for intergenic regions that showed conservation to other genomes. They ranked these candidate regions using further 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. They predicted 59 candidate ncRNA loci, 17 of which were shown by northern blot analysis to produce discrete, small RNA transcripts ranging from 45 to 320 nucleotides in size. Again, several of these were shown 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*.

Elena Rivas and co-workers 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 an amino-acid coding 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 was analysed by northern blot analysis in a single growth condition (exponential growth in rich media), and 11 were found to produce small RNAs, ranging from 40 to 370 nucleotides in size.

SHINE/DALGARNO SEQUENCE
A consensus sequence recognized during translational initiation by *Escherichia coli* ribosomes.

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 eight were confirmed by more than one of the screens). This indicates that these screens have not saturated the *E. coli* genome for new ncRNAs. Of the 27 genes confirmed by one or both of the screens carried out by Argaman *et al.* and Wassarman *et al.*, 21 are in the candidate list proposed by Rivas and colleagues, indicating that the sensitivity of the computational gene finder 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 cells)[103]. This indicates that the examination of a single growth condition by Rivas *et al.* was insufficient, and shows that confirming the 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 study by Rivas *et al.* indicate 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]. Wassarman and colleagues proposed that it would be unlikely that more than 50 new ncRNAs would be found in *E. coli*[20]. A fourth screen, using a single-sequence neural-net-based computational gene-finding approach in *E. coli*, predicted 370 sequence windows to be 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 the one hand, we have genomic screens that are unsaturated and must provide just a taste of larger numbers of ncRNA genes to come. On the other hand, if there were many ncRNA genes, one would think that they should have been detected sooner in classical genetic screens. (Most biochemical, computational and molecular biology gene-discovery approaches make strong assumptions about finding proteins, ORFs and mRNA, 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 are often small and multicopy, which makes them difficult (even impossible) targets for recessive mutational screens.

An interesting extra source of bias is introduced 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. Maaret Ridanpää and colleagues recently provided an excellent example in human genetics. Cartilage–hair hypoplasia (CHH) — a short-limbed dwarfism — was first described by Victor McKusick almost 30 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 previously been isolated biochemically[108] and its sequence was in GenBank. Otherwise, Ridanpää and co-workers 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 GenBank database and, moreover, it was an obvious candidate gene, as 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 seven amino acids in length are known[110]. ORFs greater than 100 amino acids can occur just by chance in completely random sequence; it has been argued that 10–15% of annotated ORFs in microbial genomes are in fact spurious[111]. ORF length and 'coding potential' alone is therefore often insufficient to decide whether a gene is coding or non-coding. Errors are being made in both directions. The 360-nucleotide 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 on the basis of sequence analysis that showed "no significant coding potential"[113]. Now, on the basis of comparative genome analysis and more detailed, directed mutagenesis studies, the *ENOD40* transcript seems to encode two tiny proteins that are 13 and 27 amino acids long[114].

Comparative genome analysis is an indispensable means of inferring whether a locus produces a ncRNA 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

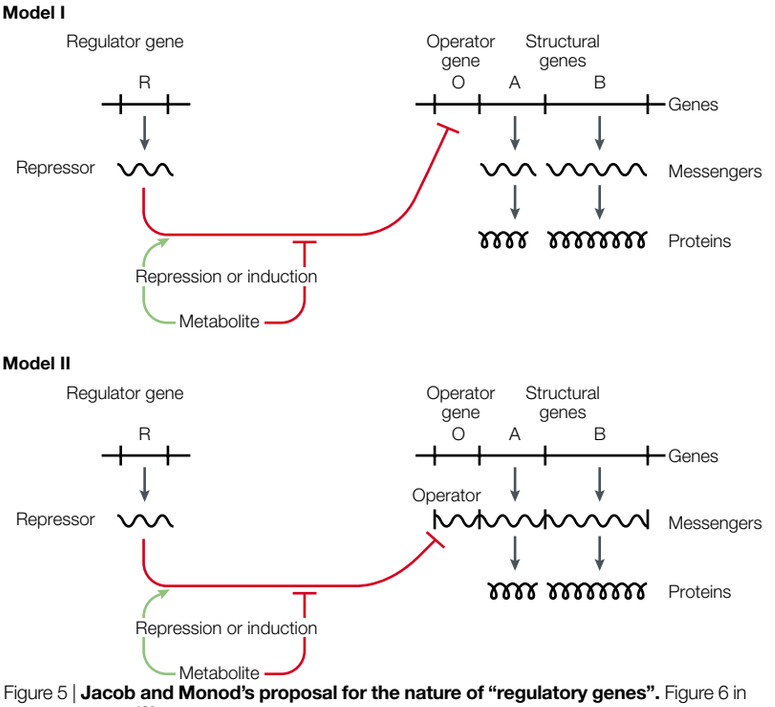

Figure 5 | **Jacob and Monod's proposal for the nature of "regulatory genes".** Figure 6 in their 1961 paper[128] indicated that structural genes might produce mRNAs that code for protein, but regulatory genes produce regulatory RNAs (note the wavy line) that interact by base pairing with operators, either at the transcriptional level (model I) or the post-transcriptional level (model II). Reproduced with permission from REF. 128 © (1961) Academic Press.

another related species. For almost all protein-coding genes (those undergoing PURIFYING SELECTION or NEUTRAL DRIFT, but perhaps not those under POSITIVE SELECTION), the pattern of mutation should also favour 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 a ncRNA by comparative analysis but, in at least some cases, a ncRNA might 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 important genetic systems, comparative analysis can (and should) become routine.

Discovering new non-coding RNA 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 gene-finding program looking for any significant conserved intramolecular secondary structure[101]. Previous attempts to develop ncRNA gene-finders that work on a single genome sequence have been stymied by the apparent lack of much significant statistical signal in ncRNAs[115,116], compared with the strong ORF and codon bias signals exploited by protein-coding gene-finders. However, an apparently successful single-sequence RNA gene-finder, using a NEURAL NETWORK approach, has recently been reported[104], and it might also be possible to identify untranslated, spliced ncRNAs by the computational identification of clustered splice-site signals[117].

Second, cDNA cloning strategies that are specifically designed to enrich for 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-fibrillarin 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 UTRs of mRNAs.

Third, it should be possible, in principle, to detect new transcripts (both ncRNA and protein-coding RNA) 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 been variable. Successful detection of some known ncRNAs has been reported anecdotally[118]; but in systematic use, such data have proved to be 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 idea 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 have apparently well-adapted, specialized biological roles. The idea that ncRNAs are a small

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

NEUTRAL DRIFT
The process by which DNA sequence acquires many mutations over time that have no phenotypic effect, and hence are not acted on by Darwinian selection.

POSITIVE SELECTION
A rare form of evolutionary change in which a mutation seems to be favoured because it is fixed in the population at a rate even greater than predicted by neutral drift.

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.

SM PROTEIN
An RNA-binding protein recognized by antibodies that are produced by people with certain autoimmune diseases; 'Sm' stands for 'Smith', the name of a patient.

and ragged band of relics looks increasingly untenable. The tiny stRNAs and miRNAs, for example, seem to be highly adapted for a world in which RNAi processing and developmentally regulated mRNA targets exist.

Therefore, consider an alternative idea — the "modern RNA world". Many of the ncRNAs we see in fact have roles in which RNA is a more optimal material than protein. Non-coding RNAs are often (though not always) found to have roles that involve sequence-specific recognition of another nucleic acid. (The choice of examples in FIGS 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 carried out by simple RNAs. Post-transcriptional 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, whereas specific sequence recognition functions are carried out by a horde of individual small RNAs that interact with these proteins. John Morrissey and David Tollervey have proposed that modification-guide snoRNAs arose in 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 mRNA and operons, François Jacob and Jacques Monod distinguished "structural genes" (such as *lacZ*) from "regulatory genes" (such as *lacI*)[128]. At that time, regulators such as *lacI* had only been defined genetically, and they were known to specifically interact with *cis*-acting sequences (such as *lacO*), either at the DNA or mRNA level. Jacob and Monod reasoned that base complementarity would allow RNA to interact highly specifically with other nucleic-acid sequences. They proposed that structural genes encoded proteins, and regulatory genes produced ncRNAs (FIG. 5). Forty years later, their proposal is looking more relevant than ever.

1. Collins, F. S. *et al.* New goals for the US 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. *Nature Genet.* **25**, 129–130 (2000).
5. Wright, F. A. *et al.* A draft annotation and overview of the human genome. *Genome Biol.* **2**, 0025.1–0025.18 (2001).
6. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
7. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
8. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
9. Roest Crollius, H. *et al.* Estimate of human gene number provided by genomewide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
10. Eddy, S. R. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699 (1999).
11. Erdmann, V. A. *et al.* The non-coding RNAs as riboregulators. *Nucleic Acids Res.* **29**, 189–193 (2001).
12. Erdmann, V. A., Barciszewska, M. Z., Hochberg, A., De Groot, N. & Barciszewski, J. Regulatory RNAs. *Cell. Mol. Life Sci.* **58**, 960–977 (2001).
13. Olivas, W. M., Muhlrad, D. & Parker, R. Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.* **25**, 4619–4625 (1997).
14. Hüttenhofer, A. *et al.* 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-mRNAs, by EST sequencing of size-selected mouse cDNA libraries.
15. Lee, R. C. & Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**, 862–864 (2001).
The isolation of 15 microRNAs in *Caenorhabditis elegans* by a combination of cDNA cloning and bioinformatics.
16. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
The isolation of 55 microRNAs in *C. elegans* by cDNA cloning.
17. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
Describes the isolation of 19 human microRNAs and 14 *Drosophila melanogaster* microRNAs by cDNA cloning.
18. Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* **11**, 941–950 (2001).
The use of computational prediction of transcriptional promoters and terminators, combined with comparative analysis, to predict putative non-coding RNA genes in *Escherichia coli*, 14 of which were shown experimentally to express small RNAs.
19. Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373 (2001).
A new general non-coding RNA gene-finding program, which uses comparative genome analysis, is used to predict structural non-coding RNA genes in *Escherichia coli*; 11 of these were experimentally shown to produce small, non-coding RNA transcripts.
20. Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & 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 non-coding RNA genes in *Escherichia coli*, 17 of which were experimentally shown to produce small, non-coding RNA transcripts.
21. Caspersson, T. Studien über den Eiweissumsatz der Zelle. *Naturwissenschaften* **29**, 33–43 (1941).
22. Brachet, J. & Chantrenne, H. The function of the nucleus in the synthesis of cytoplasmic proteins. *Cold Spring Harb. 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, New York, 1996).
26. Gesteland, R. F., Cech, T. R. & Atkins, J. F. *The RNA World* 2nd edn (Cold Spring Harbor Laboratory Press, New York, 1999).
27. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576–581 (1961).
28. Zimmermann, R. A. & Dahlberg, A. E. *Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis* (CRC Press, Boca Raton, 1996).
29. Gros, F. *et al.* 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. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–257 (1958).
31. Soll, D. & RajBhandary, U. L. *tRNA: Structure, Biosynthesis, and Function* (ASM, 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. & Choi, Y. C. SnRNAs, SnRNPs, and RNA processing. *Annu. Rev. Biochem.* **5**, 617–654 (1982).
34. Yu, Y. T., Scharl, E. C., Smith, C. M. & Steitz, J. A. in *The RNA World* 2nd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 487–524 (Cold Spring Harbor Laboratory Press, New York, 1999).
35. Lerner, M. R. & Steitz, J. A. Snurps and scyrps. *Cell* **25**, 298–300 (1981).
36. Burge, C. B., Tuschl, T. & Sharp, P. A. in *The RNA World* 2nd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 525–560 (Cold Spring Harbor Laboratory Press, New York, 1999).
37. Tarn, W. Y. & 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. & 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. & 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. & 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. & Kuroda, M. L. Noncoding RNA genes in dosage compensation and imprinting. *Cell* **103**, 9–12 (2000).

44. Avner, P. & Heard, E. X-chromosome inactivation: counting, choice and initiation. *Nature Rev. Genet.* **2**, 59–67 (2001).
45. Franke, A. & 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. & Bartolomei, M. S. Mechanisms of genomic imprinting. *Curr. Opin. Genet. Dev.* **9**, 164–170 (1999).
48. Meguro, M. *et al.* 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 box small nucleolar RNA gene in the imprinted Prader–Willi syndrome region of humans.
49. Cavaille, J. *et al.* Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA* **97**, 14311–14316 (2000).
The identification of new small nucleolar RNAs (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. & Kleckner, N. Biological regulation by antisense RNA in prokaryotes. *Annu. Rev. Genet.* **22**, 567–600 (1988).
51. Terryn, N. & 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. & 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. & Maxwell, E. S. The nucleolar snRNAs: catching up with the spliceosomal snRNAs. *Trends Biochem. Sci.* **18**, 131–135 (1993).
54. Maxwell, E. S. & Fournier, M. J. The small nucleolar RNAs. *Annu. Rev. Biochem.* **64**, 897–934 (1995).
55. Balakin, A. G., Smith, L. & 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. & 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. & 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. & 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. & 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. & Fournier, M. J. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell* **89**, 565–573 (1997).
61. Bachellerie, J. P. & Cavaille, J. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 255–272 (ASM, Washington DC, 1998).
62. Lafontaine, D. L. J. & Tollervey, D. Birth of the snoRNPs: the evolution of the modification guide snoRNAs. *Trends Biochem. Sci.* **23**, 383–388 (1998).
63. Weinstein, L. B. & 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. Nucleic Acid Res. Mol. Biol.* **39**, 241–303 (1990).
65. Maden, B. E. H., Corbett, M. E., Heeney, P. A., Pugh, K. & 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. & Bakin, A. Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.* **266**, 246–268 (1997).
67. Lowe, T. M. & Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171 (1998).
68. Barneche, F., Gaspin, C., Guyot, R. & 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. & Yue-Qin, C. Identification of 10 novel snoRNA gene clusters from *Arabidopsis thaliana*. *Nucleic Acids Res.* **29**, 1623–1630 (2001).
70. Dunbar, D. A., Wormsley, S., Lowe, T. M. & 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. *et al.* Homologs of small nucleolar RNAs in Archaea. *Science* **288**, 517–522 (2000).
72. Gaspin, C., Cavaille, J., Erauso, G. & 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. & Steitz, J. A. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell* **2**, 629–638 (1998).
74. Yu, Y. T., Shu, M. D. & Steitz, J. A. Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J.* **17**, 5783–5795 (1998).
75. Jády, B. E. & 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. & 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. & 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. *Nucleic 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. & 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 box small nucleolar RNA; seems to be the same as DR by Meguro *et al.* (reference 48) and HBII-85 by Cavaille *et al.* (reference 49).
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. & 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. & 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. & 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. & 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. & 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. *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
87. Pasquinelli, A. E. *et al.* 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. & Vaucheret, H. RNA silencing in plants — defense and counterdefense. *Science* **292**, 2277–2280 (2001).
92. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999).
93. Hutvágner, G. *et al.* 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 microRNAs. The enzyme responsible for processing small interfering RNAs is also responsible for processing the endogenous human *let-7* regulatory RNA.
94. Elbashir, S. M., Lendeckel, W. & 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-nucleotide small interfering RNAs that produce the RNA interference effect.
95. Moss, E. G. Noncoding RNAs: lightning strikes twice. *Curr. Biol.* **10**, R436–R439 (2000).
96. Grishok, A. *et al.* 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 reference 93, this study showed that Dicer is apparently responsible for processing *lin-4* and *let-7* in *Caenorhabditis elegans*; moreover, knockdowns of genes in the large *rde-1/ARGONAUTE/piwi* family show that developmental defects and defects in double-stranded RNA interference processing are separable, depending on different proteins in this family.
97. Blattner, F. R. *et al*. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
98. Wassarman, K. M., Zhang, A. & Storz, G. Small RNAs in *Escherichia coli*. *Trends Microbiol.* **7**, 37–45 (1999).
99. Lease, R. A., Cusick, M. E. & 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. & Belfort, M. Riboregulation by DsrA RNA: transactions for global economy. *Mol. Microbiol.* **38**, 667–672 (2000).
101. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).
102. Altuvia, S., Zhang, A., Argaman, L., Tiwari, A. & 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. & Holbrook, S. R. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.* **29**, 3928–3938 (2001).
105. McKusick, V. A., Eldridge, R., Hostetler, J. A., Ruangwit, U. & Egeland, J. A. Dwarfism in the Amish. II. Cartilage–hair hypoplasia. *Bull. Johns Hopkins Hosp.* **116**, 285–326 (1965).
106. Sulisalo, T. *et al.* Cartilage–hair hypoplasia gene assigned to chromosome 9 by linkage analysis. *Nature Genet.* **3**, 338–341 (1993).
107. Ridanpää, M. *et al.* 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. *et al.* 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. & Moreno, F. The smallest known gene. *Nature* **369**, 281 (1994).
111. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & 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. & 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. *cartovora*, *E. carotovora* subsp. *atroseptica*, and *Escherichia coli*. *Appl. Environ. Microbiol.* **60**, 3150–3159 (1994).
113. Crespi, M. D. *et al. 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. *et al.* 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. & 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. & Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27**, 4816–4822 (1999).
117. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98**, 11193–11198 (2001).
118. Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature Biotechnol.* **18**, 1262–1268 (2000).
119. Kruger, K. *et al.* 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. & 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. & 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. & Penny, D. Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36 (1998).
125. Poole, A. M., Jeffares, D. C. & Penny, D. The path from the RNA world. *J. Mol. Evol.* **46**, 1–17 (1998).
126. Morrissey, J. P. & 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. & Nilsen, T. W. RNA: versatility in form and function. *Nature Struct. Biol.* **7**, 831–833 (2000).
128. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
129. Ha, I., Wightman, B. & 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).

Acknowledgements
I thank T. Tuschl, V. Ambros, D. Bartel, S. Holbrook and C. Burge for generously sharing pre-publication results.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/
Dicer | dyskerin | *H19* | *IPW* | *MRP* | Piwi | *RMRP* | *roX* | *SCA8* | telomerase | *XIST*

OMIM: http://www.ncbi.nlm.nih.gov/Omim/
autosomal-dominant dyskeratosis congenita | cartilage–hair hypoplasia | Prader–Willi syndrome | spinocerebellar ataxia 8 | systemic lupus erythematosus

TAIR: http://www.arabidopsis.org/
ARGONAUTE

WormBase: http://www.wormbase.org/
alg-1 | *alg-2* | *let-7* | *lin-4* | *lin-14* | *lin-28* | *lin-41* | *lin-42* | RDE-1

Access to this interactive links box is free online.