Annotating Noncoding RNA Genes

Sam Griffiths-Jones

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton
CAMBS CB10 1SA, United Kingdom

Present address: Faculty of Life Sciences, University of Manchester, Manchester
M13 9PT, United Kingdom; email: sam.griffiths-jones@manchester.ac.uk

Annu. Rev. Genomics Hum. Genet. 2007. 8:279–98

First published online as a Review in Advance on
May 21, 2007.

The *Annual Review of Genomics and Human Genetics*
is online at genom.annualreviews.org

This article's doi:
10.1146/annurev.genom.8.080706.092419

1527-8204/07/0922-0279$20.00

Key Words

transcriptome, microRNA, snoRNA, gene prediction, genome annotation

Abstract

Noncoding RNA genes produce a functional RNA product rather than a translated protein. More than 1500 homologs of known "classical" RNA genes can be annotated in the human genome sequence, and automatic homology-based methods predict up to 5000 related sequences. Methods to predict novel RNA genes on a whole-genome scale are immature at present, but their use hints at tens of thousands of such genes in the human genome. Messenger RNA-like transcripts with no protein-coding potential are routinely discovered by high-throughput transcriptome analyses. Meanwhile, various experimental studies have suggested that the vast majority of the human genome is transcribed, although the proportion of the detected RNAs that is functional remains unknown.

INTRODUCTION

The human genome is about 30 times larger than that of the worm *Caenorhabditis elegans* or the fly *Drosophila melanogaster*, and 250 times larger than that of the yeast *Saccharomyces cerevisiae*. However, the human genome contains only a few thousand more protein-coding genes [23,244 according to the EnsEMBL database (**http://www.ensembl.org/**)] than worm (20,060) or fly (14,039), and only around 3.5 times as many as yeast (6680). The protein-coding gene count is thus not well correlated with organism complexity (at least from our human-centric viewpoint). The nonprotein-coding portion of the human genome is often considered "junk." However, although only around 1.2% of bases reside in protein-coding exons, estimates put conserved (and by implication functional) sequence at 5–10% of the genome (60, 99). Further, it now appears that most of the human genome may be transcribed (reviewed in 35), and that the number of genes that do not code for protein is much larger than expected. A subset of these genes encodes functional RNA products; these are called RNA genes or noncoding RNAs (ncRNAs). Some ncRNAs have well-understood and essential functions [for example, transfer RNA (tRNA) and ribosomal RNA (rRNA)]; other large classes have been described more recently [such as small nucleolar RNAs (snoRNAs) and microRNAs (miRNAs)]; and a rapidly increasing number have no known function. Finally, thousands of mRNA-like transcripts with little or no protein-coding potential have been identified in high-throughput studies of the transcriptome (13).

This review discusses the state of RNA annotation, with a focus on the human genome. The most useful computational methods are highlighted, and whole-genome analyses of families of "classical" RNA genes are discussed. Finally, the functional annotation of RNA sequences is considered.

THE HUMAN TRANSCRIPTOME

The cell's transcriptional output includes protein-coding RNA and non-protein-coding RNA, both of which can have many alternative splice variants, transcription start sites, and termination signals. Only around 1.2% of the human genome is contained in protein-coding exons, but 30–40% is transcribed as protein-coding mRNAs, the vast majority of which is intronic sequence. However, studies using genome-tiling arrays identify substantially more transcription than can be explained by current gene annotations (reviewed in 42, 100). The use of a variety of experimental techniques is adding weight to the idea that the majority of the mammalian genome is transcribed, including massively parallel signature sequencing (MPSS), mapping 5′ ends by cap analysis of gene expression (CAGE) and 3′ ends by serial analysis of gene expression (SAGE), and high-throughput cDNA sequencing technologies (reviewed in 31). The Functional Annotation of the Mouse Genome (FANTOM) project has sequenced 102,281 full-length cDNAs (13), of which 32,129 are annotated as protein-coding transcripts (2222 encode previously unknown proteins) and 34,030 as ncRNAs. Although these approaches focus on cytosolic polyadenylated RNA, there is evidence that as much as half of all transcriptional output might never leave the nucleus, and that a significant proportion is not polyadenylated (15).

CLASSICAL AND mRNA-LIKE ncRNAs

ncRNA transcripts fall into at least two main types, which necessarily require very different annotation approaches. "Classical" ncRNA genes encode small (typically tens to hundreds of bases) RNA products, which are expected to be highly structured through the adoption of a base-paired secondary structure. This group includes most well-understood RNA families, including tRNAs, rRNAs (the largest of which is several kilobases in length), miRNAs, and others discussed here. These are distinct from a possibly much-larger class of RNAs, which have all the properties of messenger RNA, but lack protein-coding potential. These RNAs have been termed mRNA-like ncRNAs (mlncRNAs) (83). mlncRNAs include alternative transcripts emanating from protein-coding genes, the products of antisense transcription, and a growing number of annotated transcriptionally-active loci of unknown function [e.g., as discovered by the FANTOM project (13)]. mlncRNAs are long and likely transcribed by RNA polymerase II, and may be spliced, capped, and polyadenylated. mlncRNAs probably do not conserve a base-paired secondary structure and compact three-dimensional form.

Well-studied mlncRNAs include the 17-kb mammalian Xist transcript, which has a vital role in X-dosage compensation. Xist is selectively expressed from one X chromosome in the female, causing its inactivatation (reviewed in 84); Xist was recently suggested to have evolved from a protein-coding pseudogene (20). 7H4 RNA is expressed at the neuromuscular junction (95), and may be a primary miRNA transcript (83). H19 RNA is maternally expressed in eutherian mammals, likely plays an as-yet-unknown role in the imprinting process (78), and expresses a miRNA gene (11). Growth arrest-specific gene 5 (GAS5) and U22 host gene (U22HG) are snoRNA host genes (88). The functions of almost all mlncRNAs are unknown.

Some mRNA transcripts contain predicted base-paired secondary structures in their untranslated regions. Many of these function by diverse mechanisms to regulate transcription and translation. For example, metazoan histone mRNAs do not have a 3′ poly-A tail; instead, their processing depends on a short hairpin structure that is bound by a stem-loop binding protein (SLBP) and a sequence motif that is bound by U7 RNA (19). The selenocysteine-insertion sequence (SECIS element) is a short hairpin that causes an in-frame TGA codon to be recognizezd by a selenocysteinyl tRNA rather than as a stop codon (96). Riboswitch elements are mainly located in 5′ untranslated regions (UTRs) of bacterial genes to regulate expression of the mRNA in response to metabolite binding (reviewed in 92). The presence of the thiamine riboswitch in eukaryotes highlights the possibility of a more widespread mechanism. The Rfam database of RNA families contains more than 120 known *cis*-regulatory UTR structures, including more than 20 internal ribosome-entry sites (IRES), 12 riboswitches, and a handful of frame-shift elements and thermo-regulators (27, 29).

The following discussion concentrates on the annotation of "classical" ncRNAs in mammalian genomes. These approaches are equally appropriate for the study of *cis*-regulatory mRNA structures.

ANNOTATING CLASSICAL ncRNAs IN SEQUENCED GENOMES

Until recently, many genome-annotation resources have ignored the gene complement that does not code for protein. This is not surprising, as only recently have reliable approaches for annotating RNA gene families become available, and an accurate RNA gene finding method is still lacking. De novo computational prediction of structured RNA genes is difficult. For example, ncRNAs lack the signatures that make protein-coding gene prediction possible (e.g., codons, splice signals, and other features such as length and sequence bias). Any ncRNA-gene finding method must therefore use information such as potential to form base-paired secondary structure and sequence and structural homology with known RNA families.

De Novo ncRNA Prediction

The discovery of novel functional RNAs by de novo gene prediction is a primary focus of computational RNA research. To date, no method is sufficiently mature to reliably identify ncRNAs in a genome-wide fashion. Two studies used sequence composition to identify relatively G+C (GC)-rich ncRNA genes in the A+T (AT)-rich genomes of thermophilic bacteria (47, 86). Otherwise, ncRNA prediction has relied largely on the potential of a sequence to adopt a secondary structure. Unfortunately, it has been demonstrated that predicted structures of RNA genes are not significantly more stable than predicted structures of random RNA sequences, at least for reliable discrimination on a genome-wide scale (80). Thus, a predicted stable RNA structure does not indicate functional significance. There are exceptions to this, for example, miRNA precursors do appear to adopt significantly more stable structures than does random RNA sequence (10). However, the accuracy of single-sequence structure prediction is limited in practice by incomplete models and ad hoc scoring schemes, even when the sequence is known to be a structured RNA (reviewed in 63).

More recent and successful approaches exploit the idea that functionally significant RNA structures will be conserved in related species. Computational simulation has shown that a small number of mutations will likely significantly change a secondary structure, thereby giving a conserved structure an implied function (41). In addition, the secondary structure may be maintained without conservation of the primary sequence by compensatory base mutations. These changes can be used as statistical evidence for base pairs at those positions. Predicting consensus secondary structure across a multi-species alignment of sequences is thus far more useful than for a single sequence.

A number of software tools attempt to provide a measure of probability that a given alignment of sequences adopts a conserved RNA fold. QRNA analyzes a pairwise alignment using three models of sequence evolution—protein coding, structured RNA, and other—and reports the highest-scoring model (81). RNAz calculates the probability that a multi-sequence alignment represents a conserved structured RNA by predicting the thermodynamic stability of a consensus secondary structure compared with the stability of a shuffled alignment (98). A derived z-score is combined with a structure conservation index (SCI) that measures the average stability of each sequence compared with the consensus in a support vector machine approach. EvoFold uses a probabilistic model of RNA structure and sequence evolution, called a phylogenetic stochastic context-free grammar (phylo-SCFG), to evaluate how well a substitution pattern in an alignment matches a secondary structure annotation (77).

RNAz and EvoFold have been applied independently to identify putative conserved RNA structures in the human genome (77, 97). Both find tens of thousands of candidate RNA-structural regions. Based on estimated false-positive rates, Pedersen et al. (77) predict approximately 10,000 structured RNA transcripts in the human genome (from an initial set of more than 48,000 structured regions), whereas Washietl et al. (97) estimate that more than 35000 structured RNAs are conserved in mammals. RNAz has also been used to identify candidate ncRNAs in the *C. elegans* (68) and *Ciona* (67) genomes.

The false-positive rates (and therefore the specificity) associated with these methods are subject to large and essentially unknown errors. A recent comparison of EvoFold and RNAz predictions shows that the overlap between the predictions made by the two methods is very small—less than 10% (S. Washietl & J.S. Pedersen, unpublished). The authors also estimate high false-discovery rates of 50–70%; note that the proportion of predictions that can be experimentally validated is so far undetermined. In earlier QRNA analyses of the *Escherichia coli* and *S. cerevisiae* genomes, expression could only be demonstrated for 10–20% of the RNA predictions (65, 82). One might expect the prediction specificity (and thus the validation rates) to be lower in vertebrates due to their larger genomes.

There are two alternative explanations for such poor verification rates: a) the predictions include large numbers of false-positive results; and b) many RNAs are expressed in very tight spatial and/or temporal patterns, making experimental verification difficult. Certainly, some RNA-expression profiles are highly specific. *C. elegans* miRNA lsy-6, which is involved in specifying bilateral asymmetry of chemoreceptor gene expression, is expressed in just a few neuronal cells at a specific developmental time (43). However, such low expression levels are extremely rare among known ncRNAs. In fact, some ncRNAs are among the most highly expressed RNAs in the cell (101). It is also important to note that these predictions make no distinction between bona fide ncRNA genes and structured regions of mRNA transcripts.

Homology Search

When performed on a genome-wide scale, reliable annotation of ncRNAs is currently restricted to searching for homologs of known RNA families. The nucleotide sequence databases and dedicated resources, such as RNAdb (**http://research.imb.uq.edu.au/rnadb/**) (76) and NONCODE (**http://www.bioinfo.org.cn/NONCODE/**) (58), contain thousands of annotated ncRNA sequences. Some ncRNA homologs are readily detected by sequence similarity alone. For example, rRNA sequences are well conserved across phylogenetic kingdoms, such that a simple BLAST search using the sequence of human 18S rRNA will detect all annotated 16S rRNAs in the bacterium *Bacillus subtilis* and the archaeon *Methanococcus jannaschii*. However, RNase P RNA cannot be identified in the *B. subtilis* genome sequence by BLAST using the sequence of the *E. coli* ortholog, even when using parameters that yield maximum sensitivity; for this, more complex models of RNA sequence and structure are required.

Algorithms for detecting homologs of structured RNAs can be divided into two classes: a) those specific to a particular RNA class [e.g., tRNAscan-SE (59) and ARAGORN (52) for tRNAs, miRscan (57) and miRseeker (49) for miRNAs, and snoscan and snoGPS (87) for snoRNAs]; and b) general approaches applicable to all structured RNAs that use patterns or motifs [e.g., PatSearch (30) and RNAMotif (61)] or profile-based methods [e.g., INFERNAL (21) and ERPIN (50)]. The advantages and disadvantages of each approach are clear. A specific tool uses fast, family-specific heuristics to maximize speed and sensitivity, but requires a new approach for each new RNA class. A general tool can be used to detect any RNA, but is likely to be slower and less accurate than a specific tool.

Currently, the most successful general approach for detecting homologs of known ncRNAs involves the use of statistical models, called profile-stochastic context-free grammars (profile-SCFGs) also known as covariance models (CMs). These models are analogous to profile hidden Markov models for protein sequence. An SCFG can be trained using a multi-sequence alignment of related RNA sequences annotated with the consensus secondary structure, and statistically represents the sequence and structure conservation. The model can then be used to analyze a sequence for its similarity to the training set; this can include scanning a whole genome sequence. Model building, sequence analysis, searching, and alignment are implemented in the INFERNAL package (21). Note that these tools are extremely computationally intensive. The Rfam database provides profile-SCFGs for more than 500 families of ncRNA sequences, together with precomputed sequence matches in the nucleotide sequence databases, allowing automatic detection of homologs in whole genome sequences (27, 29).

The Pseudogene Problem

The automatic annotation of structured ncRNA homologs in eukaryotic genomes is complicated by the presence of large numbers of repeats and pseudogenes. These presumed nonfunctional sequences largely swamp out the relatively small number of authentic functional ncRNA. All active SINE repeats in both the human (Alu) and rodent genomes (B1, B2, ID, B4) were derived from RNA genes: Alu/B1 repeats were derived from signal recognition particle (SRP) RNA in human/ rodents; B2 evolved from an Ala-tRNA in mouse and rat; ID is related to both the neuronally expressed BC1 RNA and an Ala-tRNA; and B4 represents a fusion of B1 and ID repeats (23, 99). There are over 1 million copies of the Alu repeat in the human genome, accounting for roughly 10% of its bases (51). Other relevant families in the human genome include U6 spliceosomal RNA, 7SK, and Y RNA—all with hundreds of predicted homologs, most of which are likely pseudogenes. Many of the affected classes share characteristics: tRNA, 7SK, Y, and U6 RNAs are all transcribed by RNA polymerase III and have internal promoter sequences, facilitating retention of transcriptional activity in duplicated copies. Some of these are among the most abundant RNA molecules in the cell: there are 10^5 to 10^6 copies of 7SK and U6 RNAs per cell (101). tRNAs, SRP RNAs, and small nuclear RNAs (snRNAs) (particularly U6 RNA) are packaged into retroviral virions (14, 24) and undergo retrotranscription and recombination.

Automated Versus Manual Annotation

The most appropriate automated ncRNA-annotation approach combines the strengths of general and specific ncRNA-homolog detection methods. For example, a reasonable candidate set of human ncRNAs can be generated using tRNAscan-SE (59) to predict tRNAs, BLAST to identify known miRNAs and snoRNAs and find the small and large subunit rRNAs, and the Rfam library of covariance models (29) to annotate homologs of other known ncRNAs. However, the pseudogene problem discussed above necessitates careful use of the resulting data. Several intensive manual efforts have produced the best currently available ncRNA predictions for the human genome, starting from automated homology predictions (23, 51, 89, 99). Note that manual annotation has largely relied on a pragmatic definition of a pseudogene, which is likely to result in conservative annotation of authentic genes. The predicted ncRNA gene sets for various genomes are listed and compared in **Table 1**.

CLASSICAL ncRNA FAMILIES

Table 1 provides the total numbers of classical RNA genes in the human and other genomes. Each of the major families is briefly reviewed below.

Transfer RNAs

Automated annotation of tRNA genes using tRNAscan-SE yields significantly different results among eukaryotic genomes, predicting around 280 tRNAs in chicken and over 175,000 tRNAs and pseudogenes in rat (23, 38). tRNA-prediction programs attempt to distinguish real genes versus pseudogenes, but it is likely that even the most carefully curated rodent tRNA sets contain a large number of pseudogenes and repeat-derived sequences. The pseudogene problem is smaller, but still significant, with other eukaryotic genomes. For example, tRNAscan-SE annotates more than 500 tRNAs in the *C. elegans* genome, and the manually curated annotation of the human genome contains 496 tRNA genes. Both the *Drosophila* and chicken genomes appear to have low numbers of pseudogenes (33, 38) and contain 287 and 280 predicted tRNAs, respectively; both sets contain all tRNA anticodons that are predicted to be required by wobble rules (32) and include a single selenocysteine tRNA. The agreement between these two numbers strongly suggests that the minimal functional tRNA set in animals likely contains around 300 members.

Table 1 Manually annotated whole-genome ncRNA sets in human (51), chicken (38), *C. elegans* (89), and *D. melanogaster* (J. Daub, C. Bergman, D. Ardell & S. Griffiths-Jones, unpublished). A summary of automated analyses of human and mouse ncRNAs is also provided (EnsEMBL)

	H. sapiens (automated)	*H. sapiens (manual)*	*M. musculus (automated)*	*G. gallus (manual)*	*C. elegans (manual)*	*D. melanogaster (manual)*
tRNA	513	496	3278	280	592	297
5S	294	14	147	12	15	99
5.8S	8	0	13	3	2	2
18S	0	0	0	0	3	0
26S	0	0	0	0	1	0
U1	142	14	174	18	12	5
U2	99	14	42	6	19	6
U4	119	2	58	4	5	3
U5	36	4	13	9	13	7
U6	823	49	512	15	23	3
U4atac	0	1	0	1	0	2
U6atac	0	5	0	4	0	1
U11	0	1	0	1	0	1
U12	2	1	6	1	0	1
miRNA	563	474*	549	121	117	78
snoRNA	574	375*	526	83	8	250
SRP	83	3	4	3	5	2
RNase P	2	1	4	1	1	1
RNase MRP	1	1	0	0	0	0
Telomerase	1	1	2	1	1	0
Y	806	32	20	2	1	0
7SK	164	1	22	4	0	0
U7	157	1	41	1	0	0
Vault	5	0	2	0	0	0

*Updated numbers of miRNAs and snoRNAs are derived from the miRBase and snoRNABase databases.

Ribosomal RNAs

The eukaryotic ribosome contains four rRNA species: 5S rRNA, 5.8S rRNA, and small (18S) and large (26S) subunit rRNAs. The atomic structure of prokaryotic ribosome subunits shows that RNA has both major structural and catalytic roles. RNA is responsible for the arrangement of the A- and P-sites and their substrates, and the ribosome itself is a ribozyme, with RNA performing the catalytic function of peptide-bond formation (reviewed in 73).

In higher eukaryotes, 5.8S, 18S, and 26S rRNA genes are usually arranged in tandem arrays that contain many hundreds of copies. The rRNA genes are separated by internally transcribed spacer (ITS) sequences in the arrangement 18S-ITS1-5.8S-ITS2-26S (39). However, complete 18S and 26S rRNA sequences are often missing from assembled genome sequences because such tandem-repeat regions are selected against both during the shotgun sequencing and assembly processes. In human, around 150–200 copies of the 44-kb rDNA repeat reside on the short arms of chromosomes 13, 14, 15, 21, and 22 (51), but the annotated genome sequence has no single intact copy of the repeat. Eight dispersed copies of 5.8S RNA are easily annotated using the Rfam/INFERNAL approach.

5S rRNA is also arranged in tandem repeats, the largest of which is located in the subtelomeric region of human chromosome 1q (51). The Rfam model identifies close to 300 5S rRNA sequences, mostly dispersed throughout the genome and thus likely to be nonfunctional pseudogenes. Only 14 human 5S sequences are classified as authentic rRNAs by manual annotation.

Spliceosomal RNAs

The spliceosome is a large ribonucleoprotein complex containing five RNA species [snRNAs U1, U2, U4, U5, and U6] together with over 200 proteins (94). The functions of the RNAs are well described (101). Interactions between pre-mRNA and snRNA are responsible for splice-site (U1) and branch-site (U2) recognition. Base pairing with a conserved loop of U5 snRNA is partially responsible for aligning the two flanking exons. A di-snRNA complex of U4 and U6 forms prior to activation of the spliceosome, from which U4 is subsequently removed. During the splicing reaction, U6 binds U2, replacing U1. Like rRNAs, snRNAs are heavily modified by pseudouridylation, 2′-O methylation, and base methylation, with such modifications being essential for the proper formation of snRNP complexes and splicing (93). It is suggested that the catalytic steps in the splicing reaction are mediated by spliceosomal RNA (94).

Plants and some animals also have a second, so-called minor or U12-dependent splicesome. The minor spliceosome is involved in the splicing of U12 introns, with characteristic AT/AC splice sites, as opposed to the canonical GT/AG sites. It also utilizes U5 snRNA, along with U4atac, U6atac, U11, and U12 species. The U4/U4atac, U6/U6atac, U1/U11, and U2/U12 pairs are structural and functional homologs (101).

Automated homology-based annotation of spliceosomal RNAs in the human genome identifies well over 800 U6-related sequences. In contrast, similar analyses of the chicken genome sequence suggests a minimal functional set of only 10-20 U6 genes. Thus, the vast majority of human homologs are likely to be pseudogenes. Other spliceosomal RNA classes have similar, albeit less extreme, numbers of pseudogenes in mammals. Rfam identifies U12 snRNA homologs in all analyzed mammalian and fish genomes, and minor splicesomal RNAs have also been annotated in *Drosophila* (74). The absence of minor spliceosomal snRNA homologs in *C. elegans* is consistent with its expected lack of U12 introns.

A recent study confirmed the expression of three U1 human snRNA variants that lack complementarity to the canonical 5′ splice site (48). These sequences can be identified by homology searches with previously known U1 sequences, but are classified as putative pseudogenes in the manually curated set (**Table 1**). These results raise the possibility that a number of other snRNA-related sequences may represent functional RNAs that bind noncanonical recognition sites.

Small Nucleolar RNAs

snoRNAs direct the site-specific modification of ribosomal RNAs and other ncRNAs in Eukaryota and Archaea. Two classes of snoRNAs, called C/D box and H/ACA box, guide 2′-O-ribose methylation and pseudouridylation modifications, respectively (see **Figure 1**). The RNAs act as guides, base pairing with complementary regions of the target RNA, while the catalytic function resides with proteins in the snoRNP complex—specifically, fibrillarin for methylation and dyskerin for pseudouridylation (18).

A third family of guide RNAs is termed the small Cajal body-specific RNAs (scaRNAs). scaRNAs accumulate in Cajal bodies, subnuclear organelles where the final steps of spliceosome assembly take place, where they guide the modification of the RNA-polymerase-II-transcribed spliceosomal RNAs U1, U2, U4, U5, and U12 (17).

scaRNAs are often composed of both H/ACA and C/D box domains, but some have two H/ACA domains or one or two C/D box domains (55).

The snoRNABase database catalogs the presence of 257 C/D box snoRNAs, 94 H/ACA box snoRNAs, and 25 scaRNAs in the human genome. These sequences are often well conserved—about 65 have reported homologs in *S. cerevisiae* (55). Mammalian snoRNAs are usually located in the introns of protein-coding mRNA transcripts. More than 60% of human sno/scaRNAs (230/375) reside within introns associated with transcripts in the EnsEMBL database. A small number of snoRNAs appear to be transcribed independently by RNA polymerase II (93). An increasing number of transcripts that host snoRNAs appear not to encode proteins; some may be functional mlncRNAs, whereas others may be so-called "inside-out" genes, with the intronic snoRNA the sole functional product. snoRNABase contains target sites for 215/375 (57%) of human sno/scaRNAs. A growing number have no obvious complementary sites in known rRNAs/snRNAs, and may guide the modification of other classes of RNAs, including mRNAs. For example, the C/D box snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C mRNA (45). Several clusters of snoRNAs reside within imprinted regions of the human genome, but no imprinting function has yet been defined.

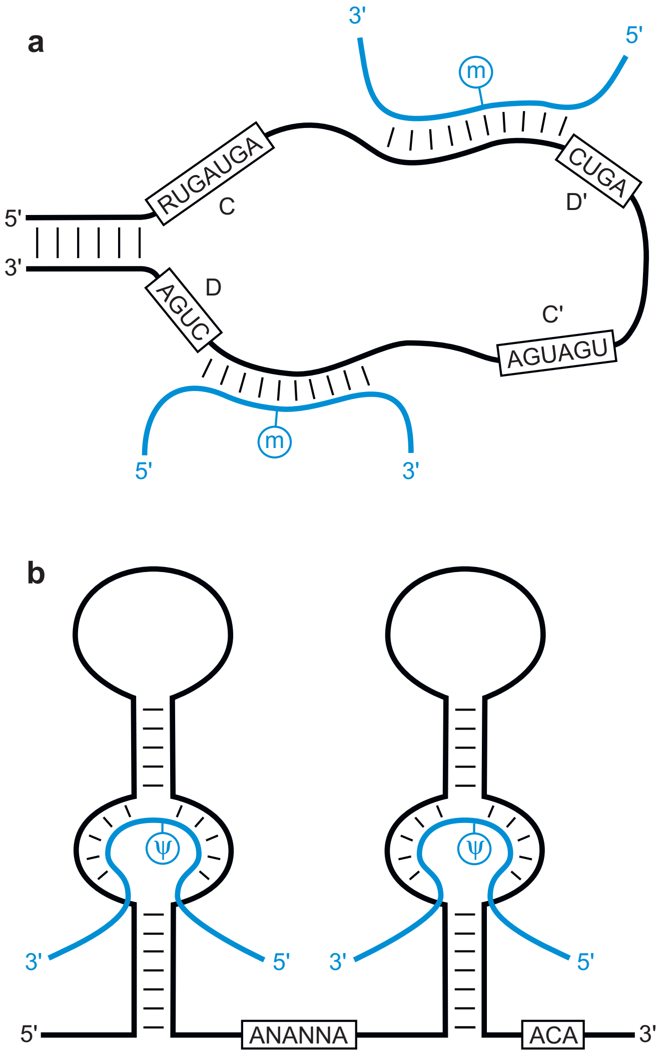


Figure 1

Secondary structures of (*a*) C/D box and (*b*) H/ACA box snoRNAs (*black*). Complementary binding to rRNA (*blue*) is shown, just 5′ of the D (and sometimes D′) boxes in C/D small nucleolar RNAs (snoRNAs), and in the bulge loops of H/ACA snoRNAs. Characteristic sequence motifs are boxed, and the position of ribosomal RNA (rRNA) modifications circled.

microRNAs

The most startling recent development in the ncRNA field has been the widespread importance of miRNAs. miRNAs are short ~22-nt sequences that inhibit the expression of target genes by binding to complementary regions in the 3′ UTRs of transcripts, triggering translational repression or transcript degradation. The history of their discovery and details of their biogenesis pathways are reviewed elsewhere (44, 53). In summary, the mature ~22-nt miRNA is excised from a stem loop called the precursor miRNA (pre-miRNA) by the ribonuclease enzyme DICER. In mammals, the pre-miRNA is around 70 nt long, and is in turn processed from the primary transcript (pri-miRNA) by the DROSHA enzyme. The duplex of mature miRNA sequence and the sequence from the opposite arm of the precursor hairpin (called the miR* sequence) is recruited into the RNA-induced silencing complex (RISC). Identical mature miRNA products can be processed from more than one hairpin precursor, expressed from multiple genomic loci. A scheme for miRNA-gene nomenclature is described elsewhere (1, 26).

The miRBase database contains 474 human precursor miRNAs, which give rise to 443 distinct mature products (26, 28). The genomic context of these sequences is varied and interesting (see **Figure 2**), with 212 (45%) appearing to be expressed on the same strand within introns of EnsEMBL protein-coding genes. An additional 22 have evidence of intronic expression based on expressed sequence tag (EST) data. The available evidence suggests that these human intronic miRNAs are processed from host transcripts, rather than as autonomous transcription units (4, 83). Twenty one overlap exons of annotated transcripts, and 14 map to UTRs, although in many cases, alternative splicing serves to locate these miRNAs in introns. Forty miRNAs overlap transcripts in the opposite orientation, with over two thirds of these having an intron at the same position on the opposite strand.

The nature of intergenic miRNA primary transcripts is poorly understood. Only a very small number have been experimentally characterized (12), with the available evidence suggesting that the pri-miRNA may be long (perhaps 10 kb or more), is transcribed by RNA polymerase II (54), is capped and polyadenylated, and may contain more than one pre-miRNA. Roughly 46% (216/474) of human pre-miRNAs are located within 5000 bases of another miRNA sequence, making up 52 clusters. The largest of these contains 36 closely related pre-miRNAs (the mir-512 family) in a region of human chromosome 19 that spans over 70 kb. A well-studied cluster of six miRNAs on human chromosome 13, with paralogous copies on chromosomes 7 and X (**Figure 2*a***) (90), was implicated as a human oncogene, with established roles in B-cell lymphomas (36) as well as lung (34) and breast cancers (40).

Other Small ncRNA Families

Both manual and automated analyses of the human genome sequence identify tens to hundreds of homologs of a further eight classes of classical ncRNA genes. The related RNase P and RNase MRP sequences are enzymatically active RNAs involved in processing precursor tRNAs and rRNAs, respectively (102). SRP RNA is an essential component of the signal-recognition particle, and the telomerase RNA carries the telomere template, together with an H/ACA box snoRNA domain. 7SK RNA inhibits the kinase activity of the CDK9/cyclin T complex, leading to reduced transcription by RNA polymerase II (71). The vault is a huge ribonucleoprotein complex of unknown function.

Two recent findings add to the complexity of the human RNA repertoire. Pollard et al. (79) reported a computationally predicted, novel RNA gene that is very rapidly evolving in the human lineage (based on human/chimpanzee genome-sequence comparisons) and is expressed during cortical development. This finding suggests an intriguing role for RNA in the expansion of brain function in humans. A class of short RNAs, called piwi-interacting RNAs (piRNAs), was recently discovered to be associated with piwi-domain proteins in the mammalian germline (2, 25). Tens of thousands of candidate piRNAs reside in large clusters in the mouse genome. Orthologous regions are conserved in the human genome, but appear to give rise to distinct piRNA sequences. Their function is currently unknown, but a role in gametogenesis has been proposed (2, 25).

HOW MANY ncRNAs RESIDE IN THE HUMAN GENOME?

Conservative manual approaches identify around 1700 classical ncRNAs in the human genome. Automated homology search methods predict over 4000 ncRNAs, with many of these likely reflecting pseudogenes

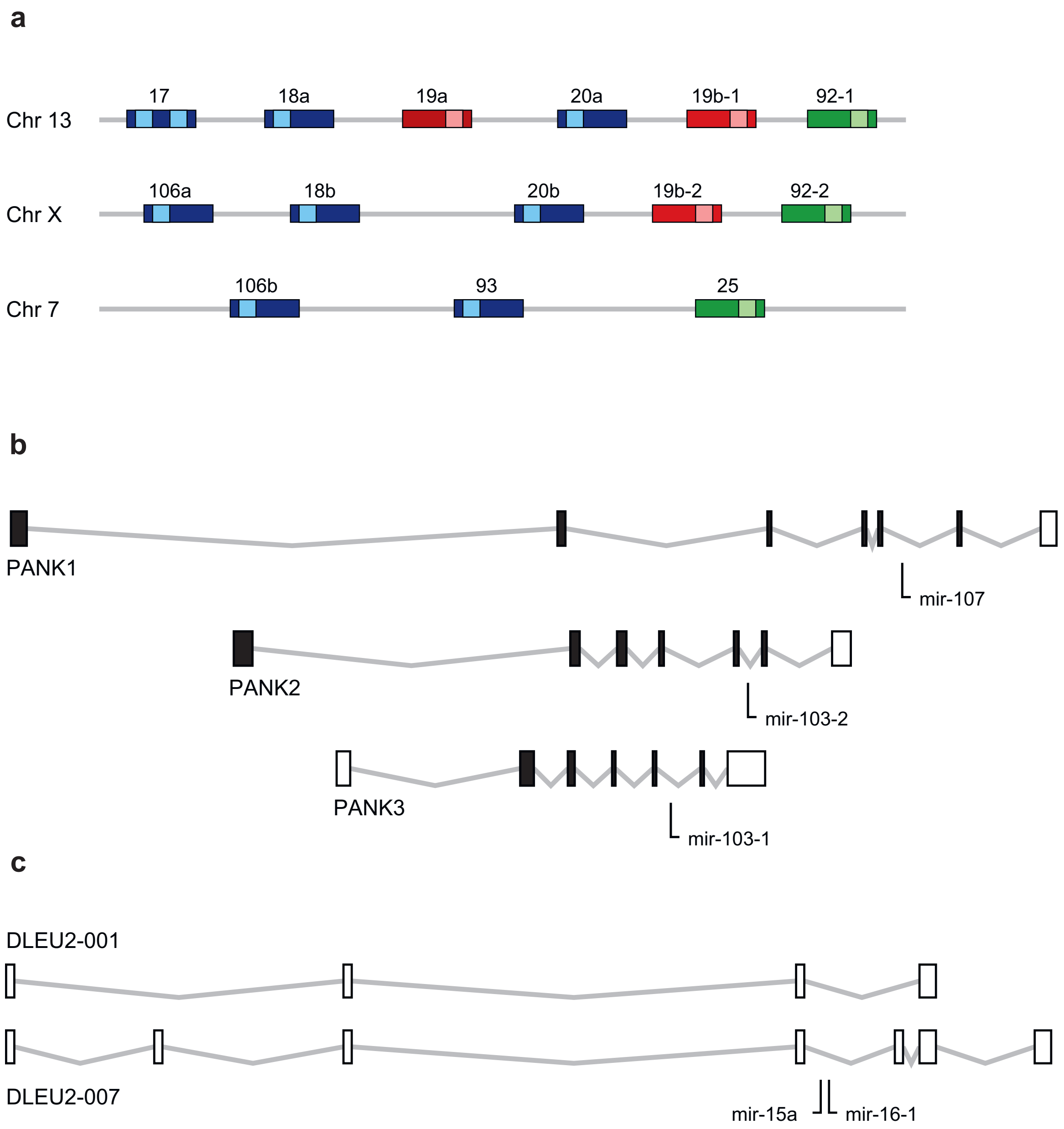


Figure 2

microRNA (miRNA) gene contexts. (*a*) The mir-17 cluster has paralogous copies on human chromosomes 7 and X. Colors depict three families of miRNA precursors, with mature miRNAs indicated by the light bands. (*b*) Paralogous miRNAs mir-107 and mir-103 are expressed from introns of paralogous panthenoate kinase (PANK) genes on human chromosomes 10, 20, and 5 (EnsEMBL accessions ENST00000371774, ENST00000316562, and ENST00000239231). (*c*) The noncoding DLEU2 locus on human chromosome 13 (two representative transcripts shown, VEGA accessions OTTHUMT00000044954 and OTTHUMT00000044960) contains mir-15a and mir-16-1. Filled exons represent open reading frames. (*a*) is adapted from (90) with permission.

and repeat-derived sequences. De novo gene-prediction programs identify upwards of 30,000 genomic sequences whose conservation patterns suggest structured RNA, but the false-positive rates for these predictions are likely to be high. mlncRNAs appear to be prevalent, with over 30,000 non-protein-coding full-length cDNA sequences catalogued, but the proportion that are functional is essentially unknown. Although such numbers are interesting, they are preliminary at best, and any discussion about the total ncRNA count in the human genome is largely speculative.

More informed discussion about specific classes of RNA is possible. The miRBase database contains 474 human miRNA sequences, already far more than the early estimates that put the total number at no more than 255 (56). Necessarily, the current miRNA set is biased toward those that are conserved (and have thus been amenable to prediction by computational approaches) and highly expressed (and thus easily detected experimentally). Recent evidence suggests that significant numbers of miRNAs may be primate specific (5, 7), and have specific temporal and spatial expression patterns. Indeed, second-generation miRNA-detection methods [e.g., amplification cloning protocols (91), microarray assays (5, 8), and deep sequencing] are identifying hundreds of candidate miRNAs. A phylogenetic-profiling method predicts close to 1000 miRNA sequences that are conserved between human and rodent (6). Speculation that the human genome contains tens of thousands of miRNAs is not yet well-supported by the data, but the presence of 1000 human miRNAs is fast becoming a realistic estimate.

FUNCTIONAL PRODUCTS OR NOISY TRANSCRIPTION

Understanding the function of all ncRNAs in the vertebrate genome is a far more pertinent problem. The known functional repertoire of ncRNAs is large. In addition to the functions described above, RNAs have roles in gene silencing, antisense regulation, imprinting, DNA methylation, chromosome maintenance and segregation, and processing of the 3′ ends of transcripts (9, 64, 69, 85). However, some known classical RNAs are still without an assigned function. For example, the vast majority of experimentally annotated miRNAs have unknown targets, and computational target prediction is a difficult problem. An increasing number of snoRNAs are also without identified targets. Y RNAs, components of the Ro ribonucleoprotein complex, were identified in 1981 (37), yet were only found to be involved in chromosomal DNA replication in 2006 (16). The functions of almost all mlncRNAs are unknown.

The question of how much non-protein-coding transcription is functional remains actively debated. Such prolific transcription can be viewed in two ways: that the majority of the human genome is functional through the action of transcribed, mostly noncoding, sequences, or that most noncoding transcription is biological noise. The latter could result from background levels of transcription that may result from initiation by randomly distributed cryptic promoter signals and transcriptional readthrough of termination signals. Only 5–10% of the human genome is conserved at thresholds commonly used to infer evolutionary constraint (51, 60, 99), and only 3–4% of mouse FANTOM cDNAs are conserved in human (13). However, functional RNAs may be conserved at a structural level in the absence of conservation at the primary-sequence level (75). Indeed, the primary sequences of several well-described mlncRNAs are not well conserved among mammals, including Xist and Air (70). It is difficult to rule out the possibility that so-called noncoding transcripts do in fact encode small proteins (22). It is also important to consider that the act of transcription may be a function in itself (for example, keeping a chromosomal region "open" for processes such as replication, other transcription, or epigenetic modification).

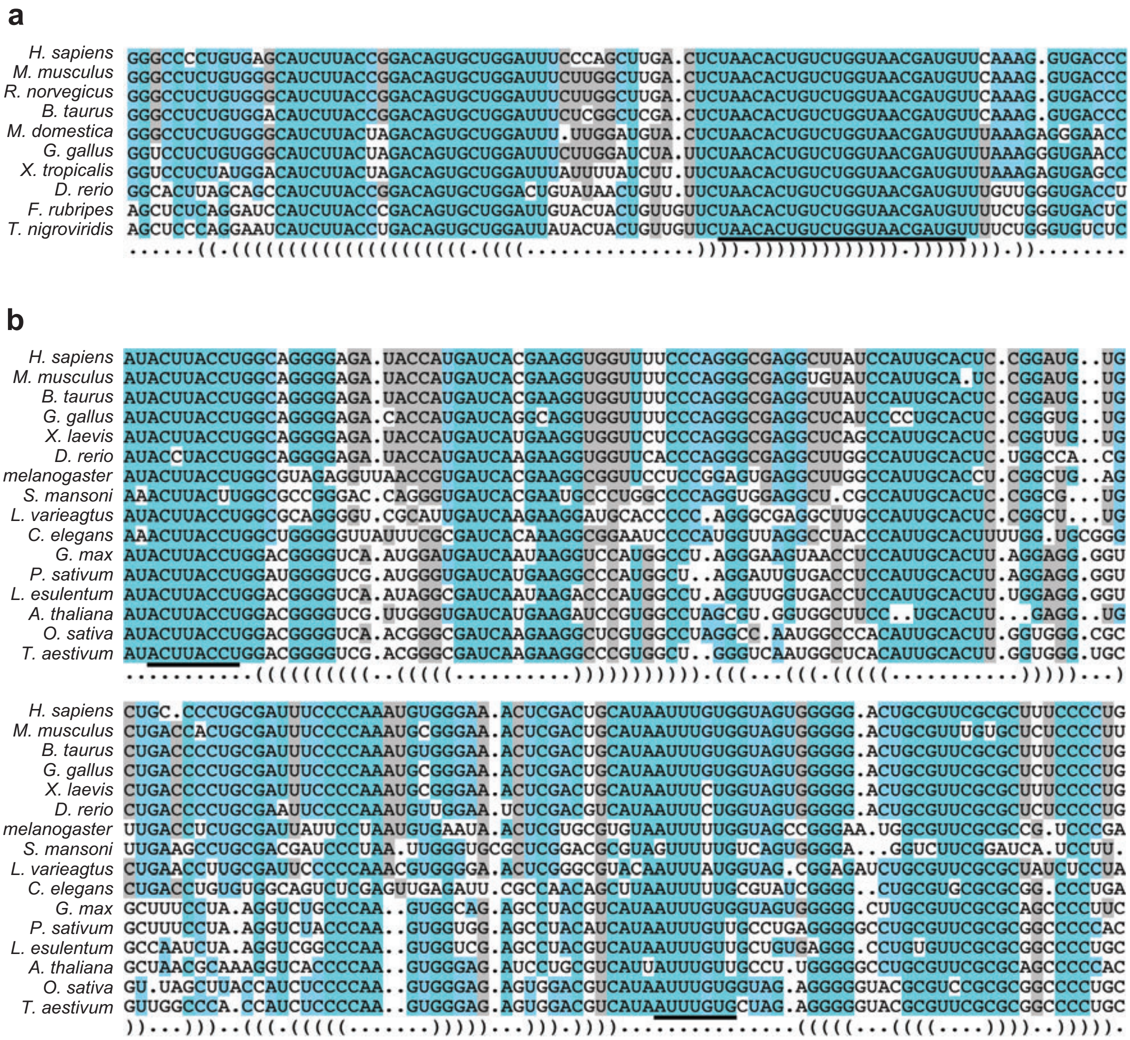
A paradigm shift relating to RNA function is not well supported by genetic screens, which have historically provided many more mutations with functional consequences in protein-coding transcripts than in ncRNAs (66). Possible explanations for this include a bias toward the publishing of protein-coding data and the tolerance of RNA function to point mutations. However, deleting megabase-sized regions of the mouse genome, selected as deserts with respect to protein-coding genes but containing more than 1000 noncoding sequences, resulted in viable mice with no obvious phenotypes (72). Although it is tempting to believe that increasingly sensitive transcription assays will eventually suggest that every base in the (euchromatic) genome is transcribed at some time in some cell (100), the relevance of such widespread transcription to the functional repertoire of RNA is not yet clear.

IMPROVING THE FUNCTIONAL ANNOTATION OF RNA SEQUENCES

How then can we improve our ability to predict classical ncRNA function? Better homology-detection methods can directly improve the assignment of function based on relationships among related sequences. Currently, many known ncRNA families are restricted to narrow taxonomic ranges. Despite intensive searches, telomerase RNA is largely unknown outside ciliates and vertebrates, yet is expected to be present in most higher eukaryotes; the *C. elegans* homolog was recently detected by improved computational methods (46). RNase P is ubiquitous and well described, yet the *C. elegans* homolog was only recently identified in the same study, and this gene has yet to be identified in the *Aquifex* genome (62). Many experimentally validated, small bacterial RNAs are known only in *E. coli* and its close relatives. We often do not know whether RNA functions are truly so specific, or whether we are simply unable to recognize remote homologs. For example, the *E. coli* 6S RNA family was, until recently, unknown outside γ-proteobacteria. Computational analyses have now identified homologs in diverse bacterial groups, including previously identified ncRNAs of unknown function in *B. subtilis* and cyanobacteria (3).

RNA gene identification, both homolog detection and de novo predicton, currently relies almost exclusively on RNA sequence and structure. Improved methods would likely utilize a much wider range of available signals. For example, ribosomal-modification maps improve the sensitivity of snoRNA-prediction methods. Prediction of promoter sequences (including internal RNA polymerase III signals) may enable better distinction between real ncRNAs and pseudogenes. Protein-binding signals, such as for Sm proteins in many snRNAs, may also be of use. It is interesting to note that ncRNA genes do not appear to be conserved with respect to gene order and orientation, in marked contrast to protein-coding genes (38). Synteny is therefore not a useful signal for ncRNA detection.

It appears that most novel functional information about ncRNAs will come directly from experimental studies. However, increased understanding about how ncRNA structure relates to function may lead to improved functional predictions for novel ncRNAs. Many classical RNAs demonstrate a pattern of sequence conservation with lower mutation rates in base-paired regions and higher rates in loops, as a consequence of selection pressure to maintain the overall structure. However, structure may be conserved without conserving the primary sequence, and the sites in loops may be highly conserved for binding or catalysis (see **Figure 3**). For example, miRNA stems are highly conserved, both because the mature miRNA needs to be heavily sequence-constrained to retain complementarity to its targets and because the structure must be conserved for proper processing. Orthologous snoRNAs have conserved loops for complementary binding to the target RNA, and many other RNAs form an intermolecular base-paired structure. Alternate structures can be



```
a
H. sapiens      GGGCCCCUGUGAGCAUCUUACCGGACAGUGCUGGAUUUCCCAGCUUGA.CUCUAACACUGUCUGGUAACGAUGUUCAAAG.GUGACCC
M. musculus     GGGCCUCUGUGGGCAUCUUACCGGACAGUGCUGGAUUUCUUGGCUUGA.CUCUAACACUGUCUGGUAACGAUGUUCAAAG.GUGACCC
R. norvegicus   GGGCCUCUGUGGGCAUCUUACCGGACAGUGCUGGAUUUCUUGGCUUGA.CUCUAACACUGUCUGGUAACGAUGUUCAAAG.GUGACCC
B. taurus       GGGCCUCUGUGGACAUCUUACCGGACAGUGCUGGAUUUCUCGGCUCGA.CUCUAACACUGUCUGGUAACGAUGUUCAAAG.GUGACCC
M. domestica    GGGCCUCUGUGGGCAUCUUACUAGACAGUGCUGGAUUU.UUGGAUGUA.CUCUAACACUGUCUGGUAACGAUGUUUAAAGAGGGAACC
G. gallus       GGUCCUCUGUGGGCAUCUUACUAGACAGUGCUGGAUUUCUUGGAUCUA.UUCUAACACUGUCUGGUAACGAUGUUUAAAGGGUGAACC
X. tropicalis   GGUCCUCUAUGGACAUCUUACUAGACAGUGCUGGAUUUAUUUUAUCUU.UUCUAACACUGUCUGGUAACGAUGUUUAAAGAGUGAGCC
D. rerio        GGCACUUAGCAGCCAUCUUACCGGACAGUGCUGGACUGUAUAACUGUU.UUCUAACACUGUCUGGUAACGAUGUUUGUUGGGUGACCU
F. rubripes     AGCUCUCAGGAUCCAUCUUACCCGACAGUGCUGGAUUGUACUACUGUUGUUCUAACACUGUCUGGUAACGAUGUUUUCUGGGUGACUC
T. nigroviridis AGCUCCCAGGAAUCAUCUUACCUGACAGUGCUGGAUUAUACUACUGUUGUUCUAACACUGUCUGGUAACGAUGUUUUCUGGGUGUCUC
                ......((.(((((((((((((((((((((.((((...............)))).))))))))))))).)))))))).))........

b
H. sapiens      AUACUUACCUGGCAGGGGAGA.UACCAUGAUCACGAAGGUGGUUUUCCCAGGGCGAGGCUUAUCCAUUGCACUC.CGGAUG..UG
M. musculus     AUACUUACCUGGCAGGGGAGA.UACCAUGAUCACGAAGGUGGUUUUCCCAGGGCGAGGUGUAUCCAUUGCA.UC.CGGAUG..UG
B. taurus       AUACUUACCUGGCAGGGGAGA.UACCAUGAUCACGAAGGUGGUUUUCCCAGGGCGAGGCUUAUCCAUUGCACUC.CGGAUG..UG
G. gallus       AUACUUACCUGGCAGGGGAGA.CACCAUGAUCAGGCAGGUGGUUUUCCCAGGGCGAGGCUCAUCCCCUGCACUC.CGGGUG..UG
X. laevis       AUACUUACCUGGCAGGGGAGA.UACCAUGAUCAUGAAGGUGGUUCUCCCAGGGCGAGGCUCAGCCAUUGCACUC.CGGUUG..UG
D. rerio        AUACCUACCUGGCAGGGGAGA.UACCAUGAUCAAGAAGGUGGUUCACCCAGGGCGAGGCUUGGCCAUUGCACUC.UGGCCA..CG
D. melanogaster AUACUUACCUGGCGUAGAGGUUAACCGUGAUCACGAAGGCGGUUCCUCCGGAGUGAGGCUUGGCCAUUGCACCU.CGGCUG..AG
S. mansoni      AAACUUACUUGGCGCCGGGAC.CAGGGUGAUCACGAAUGCCCUGGCCCCAGGUGGAGGCU.CGCCAUUGCACUC.CGGCG...UG
L. varieagtus   AUACUUACCUGGCGCAGGGGU.CGCAUUGAUCAAGAAGGAUGCACCCCC.AGGGCGAGGCUUGCCAUUGCACUC.CGGCU...UG
C. elegans      AAACUUACCUGGCUGGGGGUUAUUUCGCGAUCACAAAGGCGGAAUCCCCAUGGUUAGGCCUACCCAUUGCACUUUUGG.UGCGGG
G. max          AUACUUACCUGGACGGGGUCA.AUGGAUGAUCAAUAAGGUCCAUGGCCU.AGGGAAGUAACCUCCAUUGCACUU.AGGAGG.GGU
P. sativum      AUACUUACCUGGAUGGGGUCG.AUGGGUGAUCAUGAAGGCCCAUGGCU..AGGAUUGUGACCUCCAUUGCACUU.AGGAGG.GGU
L. esulentum    AUACUUACCUGGACGGGGUCA.AUAGGCGAUCAAUAAGACCCAUGGCCU.AGGUUGGUGACCUCCAUUGCACUU.UGGAGG.GGU
A. thaliana     AUACUUACCUGGACGGGGUCG.UUGGGUGAUCAUGAAGAUCCGUGGCCUAGCGU.GGUGGCUUCC..UGCACUU...GAGG..UG
O. sativa       AUACUUACCUGGACGGGGUCA.ACGGGCGAUCAAGAAGGCUCGUGGCCUAGGCC.AAUGGCCCACAUUGCACUU.GGUGGG.CGC
T. aestivum     AUACUUACCUGGACGGGGUCG.ACGGGCGAUCAAGAAGGCCCGUGGCU..GGGUCAAUGGCUCACAUUGCACUU.GGUGGG.UGC
                ...........((((((((((..(((((..........))))))))))).(((...(((.(((((...........)))))...)

H. sapiens      CUGC.CCCUGCGAUUUCCCCAAAUGUGGGAA.ACUCGACUGCAUAAUUUGUGGUAGUGGGGG.ACUGCGUUCGCGCUUUCCCCUG
M. musculus     CUGACCACUGCGAUUUCCCCAAAUGCGGGAA.ACUCGACUGCAUAAUUUGUGGUAGUGGGGG.ACUGCGUUUGUGCUCUCCCCUU
B. taurus       CUGACCCCUGCGAUUUCCCCAAAUGUGGGAA.ACUCGACUGCAUAAUUUGUGGUAGUGGGGG.ACUGCGUUCGCGCUUUCCCCUG
G. gallus       CUGACCCCUGCGAUUUCCCCAAAUGCGGGAA.ACUCGACUGCAUAAUUUGUGGUAGUGGGGG.ACUGCGUUCGCGCUCUCCCCUG
X. laevis       CUGACCCCUGCGAUUUCCCCAAAUGCGGGAA.ACUCGACUGCAUAAUUUCUGGUAGUGGGGG.ACUGCGUUCGCGCUUUCCCCUG
D. rerio        CUGACCCCUGCGAAUUCCCCAAAUGUUGGAA.UCUCGACGUCAUAAUUUCUGGUAGUGGGGG.ACUGCGUUCGCGCUCUCCCCUG
D. melanogaster UUGACCUCUGCGAUUAUUCCUAAUGUGAAUA.ACUCGUGCGUGUAAUUUUUGGUAGCCGGGAA.UGGCGUUCGCGCCG.UCCCGA
S. mansoni      UUGAAGCCUGCGACGAUCCCUAA.UUGGGUGCGCUCGGACGCGUAGUUUUUGUCAGUGGGGA...GGUCUUCGGAUCA.UCCUU.
L. varieagtus   CUGAACCUUGCGAUUCCCCCAAACGUGGGGA.ACUCGGGCGUACAAUUUAUGGUAG.CGGAGAUCUGCGUUCGCGCUAUCUCCUA
C. elegans      CUGACCUGUGUGGCAGUCUCGAGUUGAGAUU.CGCCAACAGCUUAAUUUUUGCGUAUCGGGG..CUGCGUGCGCGCGG.CCCUGA
G. max          GCUUUCCUA.AGGUCUGCCCAA..GUGGCAG.AGCCUACGUCAUAAUUUGUGGUAGUGGGGG.CUUGCGUUCGCGCAGCCCCUUC
P. sativum      GCUUUCCUA.AGGUCUACCCAA..GUGGUGG.AGCCUACAUCAUAAUUUGUUGCCUGAGGGGGCCUGCGUUCGCGCGGCCCCCAC
L. esulentum    GCCAAUCUA.AGGUCGGCCCAA..GUGGUCG.AGCCUACGUCAUAAUUUGUUGCUGUGAGGG.CCUGUGUUCGCGCGGCCCCUGC
A. thaliana     GCUAACGCAAAGGUCACCCCAA..GUGGGAG.AUCCUGCGUCAUUAUUUGUUGCCU.UGGGGGCCUGCGUUCGCGCAGCCCCCAC
O. sativa       GU.UAGCUUACCAUCUCCCCAA..GUGGGAG.AGUGGACGUCAUAAUUUGUGGUAG.AGGGGGUACGCGUCCGCGCGGCCCCUGC
T. aestivum     GUUGGCCCA.CCAUCUCCCCAA..GUGGGAG.AGUGGACGUCAUAAUUUGUGCUAG.AGGGGGUACGCGUUCGCGCGGCCCCUGC
                ))...)))..(((.(((((.......)))))...))).))))...............(((((...((((....))))..))))).
```

Figure 3

Multi-sequence alignments of (*a*) vertebrate mir-200a precursor-flanking sequences and (*b*) animal and plant U1 spliceosomal RNA sequences. Color represents the degree of sequence conservation, and the consensus secondary structure is shown below in dot-bracket notation. The helical region of the microRNA (miRNA) precursor includes the mature miRNA product (*underlined*) and is highly conserved. The secondary structure of the U1 RNA is conserved without high sequence conservation, whereas the loop regions, including a known 5′ splice site complementary region and Sm protein-binding sequence (*underlined*), are highly conserved.

used as regulatory switches (for example, by riboswitch elements). Sequence- and structure-conservation patterns are therefore likely to be highly family specific (I. Holmes & S. Griffiths-Jones, manuscript in preparation). In summary, although our functional understanding of the non-protein-coding portion of the human transcriptome is rudimentary, there is hope that computational analyses can be effectively combined with experimental studies to improve the annotation of functional RNAs.

SUMMARY POINTS

1. Although the robustness of de novo RNA gene prediction programs is improving rapidly, reliable genome-wide ncRNA annotation is currently restricted to homologs of known structured RNA families.
2. More than 1500 classical ncRNA genes have to date been manually annotated in the human genome, with over 4000 related sequences detected by automated methods.
3. High-throughput cDNA-sequencing efforts have revealed more than 34,000 sequences that are annotated as mlncRNAs. A number of experimental techniques suggest that the majority of the human genome is transcribed, but the functional significance of this transcription remains to be established.

DISCLOSURE STATEMENT

The author is not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I would like to thank Paul Gardner and Alex Bateman for useful comments, and Sean Eddy and Ian Holmes for discussions. S.G.J. was funded by the Wellcome Trust.

LITERATURE CITED

1. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, et al. 2003. A uniform system for microRNA annotation. *RNA* 9:277–79
2. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442:203–7
3. Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR. 2005. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA* 11:774–84
4. Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 11:241–47
5. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* 37:766–70
6. Berezikov E, Plasterk RH. 2005. Camels and zebrafish, viruses and cancer: a microRNA update. *Hum. Mol. Genet.* 14(Suppl. 2):R183–90
7. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, et al. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 38:1375–77
8. Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L, et al. 2006. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* 16:1289–98
9. Bernstein E, Allis CD. 2005. RNA meets chromatin. *Genes Dev.* 19:1635–55
10. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. 2004. Evidence that microRNA precursors, unlike other noncoding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 20:2911–17

11. Cai X, Cullen BR. 2007. The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA* 13:313–6
12. Cai X, Hagedorn CH, Cullen BR. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10:1957–66
13. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309:1559–63
14. Chen PJ, Cywinski A, Taylor JM. 1985. Reverse transcription of 7SL RNA by an avian retrovirus. *J. Virol.* 54:278–84
15. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308:1149–54
16. Christov CP, Gardiner TJ, Szüts D, Krude T. 2006. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.* 26:6993–7004
17. Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, et al. 2002. Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. *EMBO J.* 21:2746–56
18. Decatur WA, Fournier MJ. 2003. RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* 278:695–98
19. Dominski Z, Marzluff WF. 1999. Formation of the 3′ end of histone mRNA. *Gene* 239:1–14
20. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312:1653–55
21. Eddy SR. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* 3:18
22. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. 2006. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2:e52
23. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
24. Giles KE, Caputi M, Beemon KL. 2004. Packaging and reverse transcription of snRNAs by retroviruses may generate pseudogenes. *RNA* 10:299–307
25. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442:199–202
26. Griffiths-Jones S. 2004. The microRNA registry. *Nucleic Acids Res.* 32:D109–11
27. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* 31:439–41
28. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34:D140–44
29. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. 2005. Rfam: annotating noncoding RNAs in complete genomes. *Nucleic Acids Res.* 33:D121–24
30. Grillo G, Licciulli F, Liuni S, Sbisa E, Pesole G. 2003. PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.* 31:3608–12
31. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, et al. 2006. The complexity of the mammalian transcriptome. *J. Physiol. (Lond.)* 575:321–32
32. Guthrie C, Abelson J. 1982. Organization and expression of tRNA genes in *Saccharomyces cerevisiae*. In *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression*, ed. J. Strathern, J. Broach, pp. 487–528 New York: Cold Spring Harbor
33. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* 31:1033–37

34. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, et al. 2005. A polycistronic microRNA cluster, miR-17–92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res.* 65:9628–32

35. Hayashizaki Y, Carninci P. 2006. Genome network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet.* 2:e63

36. He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, et al. 2005. A microRNA polycistron as a potential human oncogene. *Nature* 435:828–33

37. Hendrick JP, Wolin SL, Rinke J, Lerner MR, Steitz JA. 1981. Ro small cytoplasmic ribonucleoproteins are a subclass of La ribonucleoproteins: further characterization of the Ro and La small ribonucleoproteins from uninfected mammalian cells. *Mol. Cell. Biol.* 1:1138–49

38. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716

39. Hillis DM, Dixon MT. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.* 66:411–53

40. Hossain A, Kuo MT, Saunders GF. 2006. Mir-17–5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. *Mol. Cell. Biol.* 26:8191–201

41. Huynen M, Gutell R, Konings D. 1997. Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* 267:1104–12

42. Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* 21:93–102

43. Johnston RJ, Hobert O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 426:845–49

44. Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.* 6:376–85

45. Kishore S, Stamm S. 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311:230–32

46. Klein RJ, Eddy SR. 2003. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* 4:44

47. Klein RJ, Misulovin Z, Eddy SR. 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA* 99:7542–47

48. Kyriakopoulou C, Larsson P, Liu L, Schuster J, Söderbom F, et al. 2006. U1-like snRNAs lacking complementarity to canonical 5′ splice sites. *RNA* 12:1603–11

49. Lai EC, Tomancak P, Williams RW, Rubin GM. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* 4:R42

50. Lambert A, Fontaine JF, Legendre M, Leclerc F, Permal E, et al. 2004. The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res.* 32:W160–65

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

52. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32:11–16

53. Lee R, Feinbaum R, Ambros V. 2004. A short history of a short RNA. *Cell* 116:S89–92

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

55. Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* 34:D158–62

56. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. *Science* 299:1540
57. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, et al. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*. 17:991–1008
58. Liu C, Bai B, Skogerbo G, Cai L, Deng W, et al. 2005. NONCODE: an integrated knowledge database of noncoding RNAs. *Nucleic Acids Res*. 33:D112–15
59. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 25:955–64
60. Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2:e5
61. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, et al. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*. 29:4724–35
62. Marszalkowski M, Teune JH, Steger G, Hartmann RK, Willkomm DK. 2006. Thermostable RNase P RNAs lacking P18 identified in the *Aquificales*. *RNA* 12:1915–21
63. Mathews DH. 2006. Revolutions in RNA secondary structure prediction. *J. Mol. Biol.* 359:526–32
64. Mattick JS, Makunin IV. 2006. Non-coding RNA. *Hum. Mol. Genet.* 15(Suppl. 1):R17–29
65. McCutcheon JP, Eddy SR. 2003. Computational identification of noncoding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res*. 31:4119–28
66. Mendes Soares LM, Valcarcel J. 2006. The expanding transcriptome: the genome as the "Book of Sand." *EMBO J*. 25:923–31
67. Missal K, Rose D, Stadler PF. 2005. Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics* 21(Suppl. 2):ii77–78
68. Missal K, Zhu X, Rose D, Deng W, Skogerbø G, et al. 2006. Prediction of structured noncoding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. Exp. Zool. B. Mol. Dev. Evol.* 306:379–92
69. Munroe SH, Zhu J. 2006. Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cell. Mol. Life Sci.* 63:2102–18
70. Nesterova TB, Slobodyanyuk SY, Elisaphenko EA, Shevchenko AI, Johnston C, et al. 2001. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res*. 11:833–49
71. Nguyen VT, Kiss T, Michels AA, Bensaude O. 2001. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* 414:322–25
72. Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* 431:988–93
73. Noller HF. 2005. RNA structure: reading the ribosome. *Science* 309:1508–14
74. Otake LR, Scamborova P, Hashimoto C, Steitz JA. 2002. The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. *Mol. Cell* 9:439–46
75. Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*. 22:1–5
76. Pang KC, Stephen S, Engstrom PG, Tajul-Arifin K, Chen W, et al. 2005. RNAdb: a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res*. 33:D125–30
77. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol*. 2:e33
78. Pfeifer K, Leighton PA, Tilghman SM. 1996. The structural H19 gene is required for transgene imprinting. *Proc. Natl. Acad. Sci. USA* 93:13876–83

79. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–72
80. Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 16:583–605
81. Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2:8
82. Rivas E, Klein RJ, Jones TA, Eddy SR. 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* 11:1369–73
83. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* 14:1902–10
84. Rougeulle C, Avner P. 2004. The role of antisense transcription in the regulation of X-inactivation. *Curr. Top. Dev. Biol.* 63:61–89
85. Royo H, Bortolin ML, Seitz H, Cavaille J. 2006. Small noncoding RNAs and genomic imprinting. *Cytogenet. Genome Res.* 113:99–108
86. Schattner P. 2002. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.* 30:2076–82
87. Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33:W686–89
88. Smith CM, Steitz JA. 1998. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell. Biol.* 18:6897–909
89. Stricklin SL, Griffiths-Jones S, Eddy SR. 2005. *C. elegans* noncoding RNA genes. In *Wormbook*, ed. The *C. elegans* Research Community. **http://www.wormbook.org**
90. Tanzer A, Stadler PF. 2004. Molecular evolution of a microRNA cluster. *J. Mol. Biol.* 339:327–35
91. Takada S, Berezikov E, Yamashita Y, Lagos-Quintana M, Kloosterman WP, et al. 2006. Mouse microRNA profiles determined with a new and sensitive cloning method. *Nucleic Acids Res.* 34:e115
92. Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* 15:342–48
93. Tycowski KT, Aab A, Steitz JA. 2004. Guide RNAs with 5′ caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr. Biol.* 14:1985–95
94. Valadkhan S. 2005. snRNAs as the catalysts of pre-mRNA splicing. *Curr. Opin. Chem. Biol.* 9:603–8
95. Velleca MA, Wallace MC, Merlie JP. 1994. A novel synapse-associated noncoding RNA. *Mol. Cell. Biol.* 14:7095–804
96. Walczak R, Westhof E, Carbon P, Krol A. 1996. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA* 2:367–79
97. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23:1383–90
98. Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* 102:2454–59
99. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
100. Willingham AT, Gingeras TR. 2006. TUF love for "junk" DNA. *Cell* 125:1215–20
101. Yu YT, Scharl EC, Smith CM, Steitz JA. 1999. The growing world of small nuclear riboproteins. In *The RNA World*, ed. RF Gesteland, TR Cech, JF Atkins. New York: Cold Spring Harbor

102. Zhu Y, Stribinskis V, Ramos KS, Li Y. 2006. Sequence analysis of RNase MRP RNA reveals its origination from eukaryotic RNase P RNA. *RNA* 12:699–706

RELATED RESOURCES

C. elegans ncRNA annotations: **http://selab.wustl.edu/people/sls/WBhtml/**
EnsEMBL 41: **http://www.ensembl.org/**
miRBase 9.0: **http://microrna.sanger.ac.uk/**
NONCODE: **http://www.bioinfo.org.cn/NONCODE/**
Rfam 7.0: **http://www.sanger.ac.uk/Software/Rfam/**
RNAdb: **http://research.imb.uq.edu.au/rnadb/**
snoRNABase v3: **http://www-snorna.biotoul.fr/**

Annual Review of
Genomics and
Human Genetics

Volume 8, 2007

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/