Noncoding RNA genes
Sean R Eddy

Some genes produce RNAs that are functional instead of encoding proteins. Noncoding RNA genes are surprisingly numerous. Recently, active research areas include small nuclear RNAs, antisense riboregulator RNAs, and RNAs involved in X dosage compensation. Genome sequence data and new algorithms have begun to make systematic computational screens for noncoding RNA genes possible.

Addresses
Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63110, USA; e-mail: eddy@genetics.wustl.edu


Abbreviations
BLAST basic local alignment search tool
Car carbon storage regulation
ncRNAs noncoding RNAs
rRNA ribosomal RNA
SCFG stochastic context-free grammar
snRNA small nuclear RNA
snoRNA small nucleolar RNAs
snRNA small nuclear RNA
UTR untranslated region

Introduction
It may sound like a script from 'The X-Files' but it's all true. There is a class of genes whose final products are not proteins; instead, they have a very different biochemistry. Some of these genes originated before the last common ancestor of life on our planet. Some people believe that based life preceded both DNA and protein in evolution [16•,17]. There are many natural 'antisense' RNAs that act as 'riboregulators' [21].

The usual suspects
Ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and the small nuclear RNA (snRNA) components of the spliceosome are probably the most intensively studied ncRNAs. Their genes are numerous because most (particularly tRNA and tRNA) occur in multiple redundant copies. The yeast Saccharomyces cerevisiae, for example, has well over 700 of these genes: 275 transfer RNA genes, 100–200 copies each of four different ribosomal RNA genes (16S, 25S, 5.8S, and 5S) and perhaps ~40 genes in all for the five splicesome snRNAs (U1, U2, U4, U5, and U6) [6]. In comparison, yeast has ~6000 protein-coding genes [6]. In most model organisms, many representatives of these genes have already been identified (usually biochemically), and additional copies are usually identified readily in genome sequence by standard similarity search- es (e.g. BLAST [basic local alignment search tool]) or, in the case of tRNAs, specialized programs [7].

A rogue's gallery
Numerous other ncRNAs have been identified. They have a surprisingly diverse range of functions. RNase P RNA, which processes transfer RNAs (and some other RNAs), is one of the few natural catalytic RNAs [8]. Signal recognition particle RNA is involved in translocating proteins across the endoplasmic reticulum [9]. A plethora of guanosine RNAs are responsible for RNA editing in trypanosomes [10]. Telomerase RNA functions as the template for adding new telomeres in most eukaryotes [11]. Schizosaccharomyces pombe meiRNA is involved in regulating the onset of meiosis [12]. In bacteria, tmRNA is involved in targeting aberrant partial protein products of truncated mRNAs for rapid degradation [13,14]. A 120 nucleotide bacteriophage ϕ29 RNA forms a hexameric structure that is essential (at least in vitro) for packaging of DNA into the phage head [15].

One very active area of study is the role of RNAs in X dosage compensation. The mammalian Xist RNA coats the inactivated X chromosome and is necessary for its inactivation [16•,17]. Xist may itself be regulated by an antisense ncRNA transcript, Fox [18]. X dosage compensation in Drosophila also involves two small ncRNAs called mael and mael [17,19,20].

There are many natural 'antisense' RNAs that act as 'riboregulators' [21]. In Escherichia coli, OxyS RNA [22,23], DsrA RNA [24,25], and MicF RNA [21] all act at the level of translational initiation, by base pairing upstream of the 3′ untranslated region of messenger RNA.
initiator AUG and either blocking translational initiation or (in the case of MinE) competing with a secondary structure to free the initiation site and thereby activate translation. In eukaryotes, the best-studied small riboregulator is the 3′ UTR of Lin-4, which was named CsrB [51]. The CsrB RNA-binding protein that probably acts as a translational repressor [48•], was defined genetically by transposon-insertion mutagenesis [49]. Several insertions fell inside and just upstream of a small 47 nucleotide RNA, which was named CsrB [51]. The CsrB protein is conserved in a variety of species and in different lineages [52].

There are some early indications that snoRNA-guided modification systems operate on targets other than the ribosomal RNAs. Vertebrate spliceosomal snRNAs are heavily modified by specific methylations and pseudouridylations, and there is a corresponding abundance of distinct snoRNA species. Vertebrate genomes are thought to contain several hundred snoRNA genes. A computational search in the yeast genome brought the total number of known yeast C/D box guide snoRNAs up to 41 [37].

There are some early indications that snoRNA-guided modification systems operate on targets other than the ribosomal RNAs. Vertebrate spliceosomal snRNAs are heavily modified by specific methylations and pseudouridylations, and there is a corresponding abundance of distinct snoRNA species. Vertebrate genomes are thought to contain several hundred snoRNA genes. A computational search in the yeast genome brought the total number of known yeast C/D box guide snoRNAs up to 41 [37].

Some snoRNAs have been identified by a computational search [42•]. Others include the abundant Drosha-like hok-omega transcript induced by heat shock [30], and the human H19 transcript [31]. A database that collates the sequences of many of these transcripts has been established ([32•]; http://www.man.poznan.pl/SSData/ncRNA/index.html).

More sno storms predicted
Eukaryotic nucleoli are at present the richest source for discovering new ncRNAs — the small nuclear RNAs (snRNAs) [33••,34–36]. snoRNAs are involved in post-transcriptional processing and modification of ribosomal RNA. Almost all snoRNAs fall into two classes on the basis of sequence features: the C/D box snoRNAs and the H/ACA box snoRNAs. Most C/D box snoRNAs guide 2′-O-ribose methylation at specific sites in ribosomal RNA, and most H/ACA box snoRNAs guide pseudouridylation. Both snoRNA families use complementary base-pairing to a target rRNA site to specify the position of a modification. Eukaryotic ribosomal RNAs are heavily modified (24 pseudouridylations and 30 ribose methylations in the five rat snRNAs). Two new guide snoRNAs have been identified by a computational search and shown to direct two methylations on U6 snRNA [38•]. A target sequence artificially expressed in an mRNA context can also be modified (albeit inefficiently) [39]. As most of the known snoRNAs were originally isolated biochemically [40], the known snoRNAs could easily be biased towards abundant guide RNAs for abundant targets (e.g., tRNAs similarly, computational snoRNA screens have required antisense complementarity to specific targets as part of their search criteria [37•,38••,41]. Genes for snoRNAs that direct modifications at non-RNA targets probably still lurk in eukaryotic genomes.

Inside-out genes
In vertebrates, all known snoRNAs except U3, U8, U13, and mitochondrial RNA processing [MRP] RNA are found in introns of pre-mRNAs [33••,35]. The snoRNAs are released from their host transcripts by nucleolytic processing. Even more surprisingly, at least four mammalian snoRNA host genes seem to be ncRNAs themselves. The sole function of these host genes appears to be to carry a payload of intron-encoded snoRNA genes. The gas5, UHG, U17HG, and U18H genes produce RNAs that have no apparent open reading frame, and their exons are not recognizably conserved between mouse and man [42•,43,44,45]. In each gene, one or more introns of their pre-mRNAs has a highly conserved region corresponding to a snoRNA gene: for example, human gas5 hosts ten snoRNAs, and human UHG hosts eight snoRNAs.

The gas5 story is interesting from the standpoint of genome analysis. The gene was identified in a subtractive hybridization screen many years ago [46]. It was assumed to be a protein-coding gene [47] until the 1380 snoRNA was serendipitously found in gas5 genomic sequence by a similarity search [42•].

Cautionary tales
Indeed, there are other examples of ncRNA genes being initially overlooked because only protein-coding genes are expected. An interesting case appeared in the recent literature. In E. coli and the plant pathogen Erwinia carotovora, there is a global posttranscriptional regulatory pathway called Csr (carbon storage regulation) [48•]. A key gene in the Erwinia pathway, aepH, was defined genetically by transposon-insertion mutagenesis [49]. Several insertions fell inside and just upstream of a small 47 amino acid open reading frame (ORF), which was then assumed to be aepH. The predicted product is still available in the public protein sequence database (Genbank g691744).

Meanwhile, the homologous pathway was being dissected in E. coli. One component, the E. coli CsrA protein, was identified in a genetic screen for glyogen-overproduction mutants and then shown biochemically to be a small RNA-binding protein that probably acts as a translational repressor [48•,50]. Purification of the CsrA protein from E. coli extracts unexpectedly showed an associated 360 nucleotide RNA, which was named CsrB [51]. The CsrB RNA represses CsrA, probably by sequestering CsrA protein and preventing it from binding its downstream targets. The csrB ncRNA gene is conserved in Erwinia, and overlaps with the aepH ‘protein coding’ region (which is otherwise not conserved); this and other data showed that the Erwinia locus did not encode a protein, but instead encodes a noncoding RNA, now renamed from AepH to CsrB [51].
Finding hairpins in a haystack
From all the evidence above, and with new examples arriving, it seems clear that there are many ncRNAs to be found. How should they be identified in a genome sequence? The most straightforward computational gene-identification approach is database similarity searching. Statistically-based similarity searching (e.g. BLAST and FASTA) is taken for granted in protein gene analysis but is more difficult for ncRNAs. Many noncoding RNAs evolutionarily conserve a consensus secondary structure more than they conserve primary sequence, so primary sequence alignment methods are often not very effective.

If the consensus secondary structure of an RNA is known and has at least one highly conserved region, one can define an exact match pattern and use it to search sequence databases using any of several software packages that are available [32–34]. If the consensus structure tolerates significant structural variations (as most do), statistical similarity search techniques are now available. Mathematical models called ‘stochastic context-free grammars’ (SCFGs) have been introduced as a sound statistical framework for RNA structure/sequence alignment [55]. Probably the first practical SCFG application was our program ‘tRNAscan-SE,’ which detects tRNA genes in DNA sequence [7]. A model used to computationally detect a number of new snoRNA genes in the yeast genome also included SCFG components [37]. SCFG methods are computationally demanding on current computer hardware, but the cited examples used short-cuts to reduce this complexity.

Here’s looking for you, kid
Probably the most significant open problem in computational ncRNA analysis is gene-finding, where one tries to find novel ncRNAs in genome sequence without using similar databases. Powerful information on both structure/sequence and function is available [52–54]. If the consensus structure tolerates significant structural variations (as most do), statistical similarity search techniques are now available. Mathematical models called ‘stochastic context-free grammars’ (SCFGs) have been introduced as a sound statistical framework for RNA structure/sequence alignment [55]. Probably the first practical SCFG application was our program ‘tRNAscan-SE,’ which detects tRNA genes in DNA sequence [7]. A model used to computationally detect a number of new snoRNA genes in the yeast genome also included SCFG components [37]. SCFG methods are computationally demanding on current computer hardware, but the cited examples used short-cuts to reduce this complexity.

Maizel’s group [36,57] has explored methods that look for regions in a genome that are significantly more stable thermodynamically than random sequence of the same base composition. This approach detects a few highly structured ncRNAs — as well as a few cis-regulatory structures — but does not appear to work in general (E. Rivas, SR Eddy, unpublished data).

Olivas et al. [58] searched for polIII promoter consensus sequences in the yeast genome and experimentally confirmed that one of their candidates was indeed a novel ncRNA. However, many RNAs are expressed from polII promoters (which are typically more difficult to predict); also, the snoRNAs provide examples of ncRNAs that do not have their own promoters at all.

In coding-dense genomes, suspicious-looking large regions with little or no coding potential have been dubbed ‘grey holes’ [59]. Olivas et al. [58] examined 59 grey holes of ≥ 2 kb in the S. cerevisiae genome. Northern analysis detected distinct transcripts from 15 of the grey holes. One transcript appears to be an H/ACA snoRNA. The remaining transcripts may either be additional ncRNAs, or may encode short ORFs. One 2.1 kb grey hole in yeast was already known to contain the telomerase mRNA gene tlc1, originally identified in a generic screen for high-copy suppressors of telomeric silencing [60].

Conclusions
A large number of ncRNA genes have been discovered but, on the other hand, most systematic genome screens for new genes are biased against discovering ncRNAs (as just one example, the whole purpose of preparing an oligo dT selected cDNA library is to deplete an RNA population of noncoding RNAs). Genetic screens are also somewhat biased against ncRNAs because they are usually small, sometimes present in multiple redundant copies, and are immune to frameshift or nonsense mutations. There are undoubtedly many more ncRNAs to be found.

The functions of the known ncRNAs are diverse. A common theme is that many are involved in specific recognition of nucleic-acid targets via complementary base pairing. This is a function that RNA is well suited for, more so than protein. For example, it must have been easier to evolve one protein Z-D ribose methylase that interacts with a hundred small guide snoRNAs with different ~15 nucleotide complementary targeting sequences, than to evolve a couple of hundred protein methylases with ~100 amino acid RNA-binding domains having different precise binding specificities. This observation supports an argument against an RNA World origin for many ncRNAs. There is ample reason for evolution to invent new ncRNAs even in a protein-DNA world. From this perspective, the ncRNAs with no apparent nucleic acid target are interesting: what is signals recognition particle RNA doing, for example?

Genome sequence data makes ncRNA discovery easier — but only somewhat. The advent of SCFG search methods means that similarity searching for structurally homologous ncRNAs is headed for firmer ground; however, the lack of useful statistical signals common to all ncRNA genes means that novel ncRNA genefinding will remain difficult. ncRNA gene-finders may have to be specialized for particular subsets of genes (polIII-transcribed ncRNAs, highly structured ncRNAs, and so on).

One powerful computational approach that will become possible is comparative genome analysis. Comparison of two genomes that have diverged just the right distance, so that alignments show functional regions standing out as islands of conservation, reveals regions that are under selective pressure, including ncRNA genes. Most completely sequenced genome sequences are too divergent to permit powerful comparative analysis but the C. briggsae and mouse genomes are slated for sequencing, to enable
comparisons to C. elegans and human, and many closely related microbial genomes are also becoming available. Well-chosen pairs of genomes may be the most fertile hunting grounds for novel ncRNAs.

Acknowledgements
Research in my group is supported by National Institutes of Health Grant "NucleicISBN 14-04, to those whose work I have been unable to cite due to space limitations.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:
* of special interest
** of outstanding interest
16. Altenmuller K, Cech T, Atkins J. The RNA World model doesn't generate many testable hypotheses but here the authors propose one. If modern RNAs descend from the RNA world, phylogenetic analysis might indicate that RNA sequence families diversified before proteins did. This is the most basic test of the RNA World model. The results are compared to RNAs evolved before the modern forms of tRNA-like RNA synthases.
17. Altenmuller K, Cech T, Atkins J. The RNA World model doesn't generate many testable hypotheses but here the authors propose one. If modern RNAs descend from the RNA world, phylogenetic analysis might indicate that RNA sequence families diversified before proteins did. This is the most basic test of the RNA World model. The results are compared to RNAs evolved before the modern forms of tRNA-like RNA synthases.

An inclusive review of the recent and voluminous snoRNA literature.


38. Tycowski K, Yao ZH, Graham P, Steitz JA: Searching genome databases for other RNAs whose sequence and structure are almost all of the C/D snoRNAs that guide ribosomal RNA 2′-O-methylations that hadn’t been found. Our model used a number of probabilistic modeling techniques that have only recently been understood and applied to RNA sequence and structural analysis; the game techniques should apply to searching genome databases for other RNAs whose sequence and structural consensus is reasonably well-understood.


40. Schneider C, King R, Philipson L: Definition of snoRNA guide modification of RNAs other than tRNA. This paper may open the floodgates for an even larger deluge of snoRNA genes.

41. Nicoloso M, Qu LH, Michot B, Bachellerie JP: RNA splicing is a member of the 5′-TOP gene family. It has long been observed that intronic snoRNAs tend to occur predominantly in introns of ribosomal protein genes and other proteins involved in translation but the correlation was not absolute. The correlation with the 5′-TOP family is stronger, as points to the intronic snoRNA genes being ‘chosen’ because they share a special global transcriptional program.


50. Cavaille J, Nicoloso M, Bachellerie JP: Evidence that host genes of the 5′-TOP family is a member of the 5′-TOP gene family. It has long been observed that intronic snoRNAs tend to occur predominantly in introns of ribosomal protein genes and other proteins involved in translation but the correlation was not absolute. The correlation with the 5′-TOP family is stronger, as points to the intronic snoRNA genes being ‘chosen’ because they share a special global transcriptional program.


