WASHINGTON UNIVERSITY

Division of Biology and Biomedical Sciences

Molecular Genetics Program

Dissertation Committee:
Sean Eddy, Chairperson
Warren Gish
Kathleen Hall
Steve Johnson
Tim Schedl
Michael Zuker

Combining New Computational and Traditional Experimental Methods to Identify tRNA and snoRNA Gene Families

by

Todd Michael Johnson Lowe

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 1999
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank three important groups of people, without whom this dissertation would not have been possible: my committee, my wonderful lab-mates, and my family.

I would like to first thank the members of my dissertation committee – not only for their time and extreme patience, but for their intellectual contributions to my development as a scientist. I am indebted to Kathleen Hall, who first taught me that RNA genes were "cool" in my favorite section of the Nucleic Acids core graduate course. Without the appreciation and excitement in RNA research inspired by those lectures, I may not have ever pursued this challenging area of biology overlooked by many. To Tim Schedl, I thank for being a supportive, strong guiding force as Chair of my committee. I am particularly appreciative to Tim for agreeing to head a committee dominated by computational biologists; he offered a welcome, balancing perspective as a rigorous experimental geneticist. To Warren Gish, who helped train me as a budding computational biologist even before I arrived at graduate school. My experience working with Warren on dbEST at the National Center for Biotechnology Information was extremely positive and fun (Ultimate and Friday TGIFs sipping margaritas certainly included). To Michael Zuker, for kindly sharing his decades of wisdom in the RNA field. The tRNAscan-SE website was enhanced with his help creating graphic representations of tRNA secondary structures. To Steve Johnson, whom I am most appreciative for agreeing to serve on the committee on short notice, and knowing he would probably have less than two weeks to read my thesis. I earnestly hope to have the chance to contribute to vertebrate genomics in the future through active collaborations with Steve.

Most of all, I would like to thank my thesis advisor, Sean Eddy, a talented teacher and passionate scientist. For a young researcher who had never before taken on a graduate student, Sean seemed to be wise beyond his experience. Sean took me into his lab after I had left my first thesis lab under less-than-favorable conditions, without questions or prejudgement – for that, I am indebted and thankful for the fresh new opportunities he offered. At several points during my thesis work, Sean put my interests as a student ahead of his own – as a young, unestablished faculty member, his ultimate concern for the welfare of his students is noteworthy. I also thank Sean for appreciating my research strengths and patiently encouraging me to improve in my weaker areas. His strong support of my own ideas and research directions, and confidence in my abilities were benefits not all thesis students enjoy (but should). Graduate school can be a difficult, draining experience. I am proud to say my experience in the Eddy lab was intellectually exciting and fun, and has energized me to continue in academic research. I sincerely hope I continue to have opportunities to interact with Sean for the rest of my research career.

To my lab-mates, thanks for the fun and support. My experience in the lab was greatly enhanced as it filled out from just Sean, Mindi, and me. I greatly look forward to having all of you as colleagues in the years ahead. To Mindi and Cheryl, many thanks for help at the bench and great company. Of all the people I have worked with in the "wet-lab" environment, I will easily miss hanging out with you two the most. I only hope my future wet-lab mates have a similarly adventurous taste in music. Lastly, I wish to sincerely thank Linda Lutfiyya. Linda was a best friend, a source of great emotional support, and the best "fun stuff" organizer I knew in grad school. Linda was also critical in the success of my main thesis project. She helped train me in yeast bench technique, sharing her excellent advice, reagents, and protocols *eagerly* through dozens of gene disruptions and tetrad dissections. Without her expertise and other valuable resources from the Johnston lab, my project might not have ever come to fruition. I cannot adequately express how thankful I am.

Finally, but not least, I want to thank my parents and my identical twin brother Robert

(with whom I shared so much growing up, hence the "us" in this section). My parents always encouraged us to ask questions, to be curious about how things work. Thanks for watching endless *Nova*, *Nature*, and *The Body Human* programs on PBS with us when we were little. Thanks for encouraging us to be independent thinkers, and having confidence in our abilities to go after new things that inspired us. Thanks Dad for taking us into lab with you to see those cool pictures you called "electron micrographs" when we were just five or six – and showing us how marvelously exciting biological research can be by your excitement. Thanks for teaching us that it is important to try to leave the world just a little better than when you came into it, and how a career in research can be a worthy part of that pursuit. And, of course, thank you both for your constant support through the ups and downs of my academic career. It has been bumpy at times, but your confidence in me has enhanced my ability to get through it all and succeed in the end.

And my most heartfelt thanks to my brother Robert. Without a doubt, my interest in genetics started when I realized we had the same DNA, down to the base pair, and yet we were still so different in *some* ways. I sometimes consider our lives a life-long experiment, in which chance and the interesting people we interact with split us on different paths. Our intense, yet positive academic and athletic competition up through high school in large part shaped my desire to be first or the best in my endeavors. Actually being the best is of course secondary to always striving for that goal. I cannot imagine being the person I am today without such a great brother through the years. Thanks for *everything* that helped me get to this day.

Chapter 1

Introduction

We are now at the beginning of the "genome age" in biological research. Several dozen genomes have been completed since the first complete bacterial genome was published in 1995 (Fleischmann et al., 1995). Microbial genome sequencing projects have become common, and major genome centers are ramping up production to tackle the gigabase genomes of human and other multicellular eukaryotes. Our capacity to sequence DNA has far outpaced our ability to characterize individual gene function experimentally. Now, many thousands of predicted genes exist in the public databases for which we have little or no understanding of their biological function. For example, no functional information is known for more than 50 % of the 19,099 predicted protein coding genes in the recently completed *Caenorhabditis elegans* genome (*C. elegans* Sequencing Consortium, 1998). Much of the next era of biological research will involve assigning basic function to each of these anonymous components, and fitting them into the massively complex networks of interactions within the cell.

1.1 Genome Annotation and Gene Prediction

Not surprisingly, one of the most urgent tasks of "computational geneticists" today is to identify and infer function of new genes which have not been studied experimentally. Computational functional inference usually involves recognizing sequence similarity between an anonymous query and a characterized matching sequence. For protein-encoding genes, a handful of generalized computational tools such as BLAST (Altschul et al., 1990; Gish, 1998), FASTA (Pearson & Lipman, 1988), and HMMER (Eddy, 1996) are quite adept at recognizing distant evolutionary relationships based on primary sequence conservation. Comparisons of this type yield new information only if a previously studied homolog is present in the database. Dedicated gene-finding programs such as Glimmer (Salzberg et al., 1998), GeneMark (Hayes & Borodovsky, 1998), GEN-SCAN (Burge & Karlin, 1997) and others attempt to identify genes based on sequence features shared by all protein coding genes such as start and stop codons, and the periodicity and non-uniform frequency

of codons. These gene predictions give potential gene boundaries but reveal nothing of function.

Once sequence annotators have performed their analyses on a new stretch of DNA, the inferred information is generally deposited in public or specialized databases for use by experimental biologists. The bulk of sequence in the public databases (Genbank (Benson et al., 1999), EMBL (Rice et al., 1993), DDBJ (Tateno & Gojobori, 1997)) is from the major genome centers which annotate millions of nucleotides of sequence each month. Because of the volume of sequence processed, it is necessary to use computational tools which require limited human supervision. Although some have argued that all annotation should be conducted "on-the-fly" (Wheelan & Boguski, 1998), final inspection by annotation specialists is critical for resolution of conflicting information from different sources, including similarity to expressed sequence tags (ESTs) and homologous genes, or gene boundary predictions from various gene finders. The goal, of course, is to present as many accurate predictions of true DNA/gene function as possible (sensitivity), while limiting the number of false predictions (selectivity).

My first project in the lab involved improving the selectivity of transfer RNA (tRNA) gene detection for large scale, automated genome analysis at the Genome Sequencing Center here at Washington University. The best existing program, tRNAscan 1.3 (Fichant & Burks, 1991), was expected to produce about one false positive for each correctly identified tRNA in the human genome. The new program I developed, tRNAscan-SE, significantly reduces false positives while increasing search sensitivity by combining the strengths of multiple tRNA search methods. This work was published (Lowe & Eddy, 1997) and is detailed in Chapter 2.

1.2 Computational Detection of RNA genes

Most biologists and genome researchers concentrate solely on protein coding genes, thus are not aware of the special issues involved in detecting RNA genes. The variety of RNA

genes known today is fairly small relative to protein coding genes, although the number of members within a single RNA gene family can be substantial. For example, the yeast *S. cerevisiae* contains 274 transfer RNAs (Lowe & Eddy, 1997), and to date, 65 small nucleolar RNAs (snoRNAs) (Samarsky & Fournier, 1999). Taken together, these two RNA families comprise more than 5% of the estimated 6000 total protein coding genes in the yeast genome (Goffeau et al., 1996). Thus, computational methods are certainly needed to identify these and other RNA genes which are otherwise hidden between and sometimes within protein coding regions (*e.g.*, within introns).

RNA gene prediction presents a particularly challenging problem. Unlike for protein-coding genes, there are no generalized computational methods for identifying new classes of RNA genes. Even for well-known RNAs with homologs present in the database, detection via similarity search methods often fails since these methods only detect primary sequence conservation. Homologous RNA genes predominantly preserve secondary structure, which allows for base-paired nucleotides to change as long as a compensatory change in the partner maintains pairing (*e.g.*, a C-G pair can change to G-C, A-T, or T-A pair). This property of RNA genes often precludes detection of other family members within the same genome or within other species' genomes.

Two brief examples illustrate this point. Transfer RNAs all share the same basic "clover-leaf" secondary structure and biological function. The *Haemophilus influenzae* genome has 58 annotated transfer RNAs (Fleischmann et al., 1995). A WU-BLAST search (Gish, 1998) of the *H. influenzae* tRNA-Ser-3 gene against its own genome identifies only 2 other tRNAs with significant P-values (<0.05). The ribonuclease P (RNaseP) RNA, involved in the 5' end maturation of tRNA precursors, is a phylogenetically ubiquitous RNA with homologs from more than 250 species spanning all three domains of life (Brown, 1999). The telomerase RNA, involved in maintaining eukaryotic chromosomal telomeres, has been identified in ciliates, yeast, and mammals. Neither the RNaseP RNA nor the telomerase RNA homologs have been identified by current computational methods in the completed *C. elegans* genome (*C. elegans* Sequencing Consortium, 1998), in spite of the fact that *C. elegans* is expected

to require both.

Currently, the most effective methods for identifying RNA genes use primary *and* secondary structure information specific to each RNA gene family (Gautheret et al., 1990; Fichant & Burks, 1991; Sakakibara et al., 1994a; Eddy & Durbin, 1994). The most accurate of these employ probabilistic RNA structural profiles, or "covariance models". Covariance models are able to capture both primary consensus and secondary structure information through the use of stochastic context-free grammars (SCFGs) (Grate, 1995; Sakakibara et al., 1994a; Eddy & Durbin, 1994). Much like sequence profiles (Gribskov et al., 1990; Krogh et al., 1994), covariance models are constructed from multiple sequence alignments of family members. These SCFG-based methods have practical limitations due to the complexity of their exhaustive calculations, limiting the length of the target RNAs or the size of genome sequences that can be searched in a reasonable amount of time (Sakakibara et al., 1994a; Eddy & Durbin, 1994). The success of tRNAscan-SE (Chapter 2) is due in large part to harnessing the power of covariance models while reducing their genome search space (thus time) by about ten-fold.

Aside from computational complexity issues, covariance models are not well-suited to represent a certain class RNA genes known as "antisense RNAs". This type of RNA interacts with other RNA molecules via short stretches of complementary bases. One example is the small nucleolar RNA (snoRNA) gene family. SnoRNAs direct highly specific nucleotide modifications via their antisense regions that pair with a target ribosomal RNA sequence (reviewed below). An alignment of snoRNAs for SCFG-based profile training does not capture the information contained within the rRNA complementary region, as these sequences change for each snoRNA and appear non-conserved. In fact, the ability for these regions to base pair to other RNAs is their most important, information-rich quality. For these reasons, SCFGs fail to detect snoRNAs and likely other antisense RNA gene families.

SCFG-based profile search methods are also championed because they are general. Instead of creating a completely new search program for each new type of RNA, profile SCFGs only require an alignment from which to create a new RNA gene search model. This quality

can also be seen as limitation. RNA genes are different from proteins in that prediction of RNA gene function can be relatively simple based on examination of specific sequence characteristics. Prediction of tRNA identity and function is a good example. Covariance models are excellent at detecting tRNAs, but because they are general, they only produce gene boundaries and have no concept of anticodon sequence or gene function. Specialized programs like tRNAscan-SE can predict function automatically based on recognition of the anticodon sequence.

1.3 Application of New Tools for Biological Research

tRNAscan-SE continues to be a commonly used tool for genome analysis and has now become a standard tool for analysis of newly completed genomes at the major genome sequencing centers. In Chapter 3, I use the program to search the *C. elegans* genome to identify and analyze the first complete tRNA family from a multicellular eukaryote. A classic prediction (Guthrie & Abelson, 1982) regarding tRNA species representation and application of the "wobble rule" to eukaryotes was confirmed. A correlation between tRNA genome copy number and intracellular tRNA levels was supported. And finally, over 200 tRNA-like pseudogenes were identified and classified, including the first example of a high-copy number SINE-like repetitive element in *C. elegans*.

Creation and application of tools like tRNAscan-SE are of value to the scientific community, although tRNA research has reached a mature, linear growth phase. tRNAs have been intensively studied for at least three decades, thus few "unexpected" biological findings resulted from this project. Upon completion of the tRNA work, I became interested in an active, recently rejuvenated area of RNA research, the small nucleolar RNAs. A landmark study had recently been published (Kiss-Laszlo et al., 1996) showing the link between one type of snoRNA gene and placement of ribose methylations within rRNA. The study also implied that dozens of snoRNAs were yet to be discovered in both yeast and mammals. The same study gave a detailed profile of snoRNA sequence characteristics which could be

used to train a probabilistic search program.

For reasons already discussed, covariance models are not able to model snoRNAs adequately. Instead, I created a new, specialized program employing probabilistic scoring methods, tailored specifically to snoRNA gene features (Chapter 4). My goal was to identify all snoRNAs of this type in the recently completed yeast genome. The project had a definable goal of associating at least one snoRNA with each of 55 ribose methylation sites in yeast rRNA. Once implemented, I carried out multiple rounds of snoRNA gene prediction, experimental gene disruption, and assay for loss of the linked ribose methylation. As newly identified snoRNAs were proven experimentally, I incorporated them into my training data, thus improving search sensitivity and selectivity for subsequent rounds of prediction. In the end, I was able to identify and verify 22 new snoRNA genes, and assign snoRNAs to 51 of the 55 methylation sites (Lowe & Eddy, 1999). Combining a new theoretical method with unambiguous experimental verification was key to success of the project.

The snoRNA search program was then modified and applied to seven archaeal genomes, resulting in the identification of over 200 new snoRNA genes in the first report of snoRNAs in the domain Archaea (Chapter 5). The work was made possible by a collaboration with an experimental lab which provided a "seed" alignment of 18 experimentally verified archaeal snoRNAs. Again, the combination of theoretical and experimental methods produced results surpassing what either method could have achieved independently.

In the following sections, I review the two research areas that formed the foundation of the methylation guide snoRNA work described in Chapters 4 & 5.

1.4 The Small Nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are named for their subcellular localization within the nucleolus. The nucleolus is a dark-staining structure within the nucleus that is the site of ribosomal RNA transcription and maturation (Hadjiolov, 1985; Woolford, 1991). After transcription by RNA polymerase I, the primary rRNA transcript undergoes numerous

cleavages resulting in the 18S, 5.8S and 28S rRNAs, as well as dozens of specific nucleotide modifications. In the end, these trimmed and decorated rRNA molecules fold into complex scaffold structures that associate with a huge number of ribosomal proteins (at least 78 in yeast! (Planta & Mager, 1998)) to become a mature ribosome. Small ribonucleoprotein particle complexes (snoRNPs) associate with ribosomal RNA in the course of maturation, and are essential for proper rRNA cleavage and modification (Tollervey et al., 1991; Mattaj, 1993). SnoRNAs are the RNA component of snoRNPs, and represent a large, complex population of small RNAs (Riedel et al., 1986).

SnoRNAs have been subdivided into two main classes, named for their conserved sequence motifs: the C/D box snoRNAs, and the H/ACA box snoRNAs (Balakin et al., 1996). A single snoRNA, the RNA component of the ribonuclease for mitochondrial RNA processing (MRP RNA), fails to fit into either classification. SnoRNAs are technically a subgroup of the small nuclear RNAs (snRNAs), but should not be confused with the snRNAs involved in messenger RNA splicing, the spliceosomal RNAs (*e.g.*, U1, U2,U4, U5, U6 (Guthrie & Patterson, 1988)).

1.4.1 Common Characteristics

SnoRNAs are short molecules, generally between 60-400 nucleotides (nt) in length (yeast snR30 is exceptional at 608 nt). They are transcribed by RNA polymerase II (except MRP and plant U3, transcribed by RNA polymerase III), and can be found in several different genomic contexts. These loci include the introns of protein-coding genes (on the mRNA coding strand), polycistronic arrays of multiple snoRNAs, or single, independent transcription units. Most mammalian snoRNAs occur in introns (Smith & Steitz, 1997), and most yeast snoRNAs occur independently or within arrays. SnoRNAs that are independently transcribed have a 5' trimethylguanosine cap structure which protects them from degradation. Special nucleolytic pathways are required to free intronic and polycistronic snoRNAs for processing to mature lengths (Petfalski et al., 1998; Chanfreau et al., 1998a).

1.4.2 Phylogenetic Range

Phylogenetically, snoRNAs have been found throughout the eukaryotes, including specific examples from mammals (humans, rodents, pigs), other vertebrates (chickens, *Xenopus*, fish, snakes), metazoan invertebrates (*Drosophila melanogaster, Caenorhabditis elegans*), plants (*Arabidopsis thaliana*, rice, corn, potato), yeasts (budding yeast, fission yeast), and protists (trypanosomes, *Euglena gracilis*, *Chlamydomonas reinhardtii*, *Dictyostelium discoideum*). The majority of snoRNA research has been performed in mammalian or yeast systems – of the roughly 400 snoRNAs currently in the Genbank database (Benson et al., 1999), over half are from these two groups. Since the finding of snoRNAs in early branching protists such as *Euglena gracilis* (Greenwood et al., 1996) and trypanosomes (Levitan et al., 1998; Roberts et al., 1998; Dunbar et al., 1999), it is expected that snoRNAs will be found in all eukaryotes. In contrast, snoRNAs have not been found in any bacteria or archaea (prior to work carried out in this thesis).

1.4.3 Associated Proteins

SnoRNAs can be isolated biochemically based on their association with one or more nucleolar proteins. Fibrillarin (NOP1 in yeast) is the best studied snoRNA-associated protein, and is one of the core components of snoRNPs that associate with C/D box snoRNAs. Fibrillarin is highly conserved throughout eukaryotes, and is required for pre-rRNA processing, pre-rRNA methylation, and ribosome assembly (Tollervey et al., 1993). Other C/D box snoRNA-associated proteins include p68 in *Xenopus* (Caffarelli et al., 1998), Nop56p and Nop58p in yeast (Gautier et al., 1997), and a 65 kDa U14-associated protein in mouse (Watkins et al., 1998b). The other major family of snoRNAs, the H/ACA box snoRNAs, were originally recognized as a group based their association with the yeast protein GAR1p. Other proteins associated with H/ACA box snoRNAs include Cbf5p (a pseudouridine synthase), Nhp2p, and Nop10p (Watkins et al., 1998a; Henras et al., 1998). Many other nucleolar proteins are known to be essential for ribosomal RNA processing, yet direct associations

with snoRNAs or snoRNA-containing snoRNPs have not been demonstrated.

1.4.4 Biological Functions

Initial biochemical studies of snoRNAs in *S. cerevisiae* yielded excitement at the large variety of RNA species, but also perplexity at their possible functions (Riedel et al., 1986). In contrast to the abundant (> 200,000 copies / cell), essential U1-U6 snRNAs already being studied in metazoans, these RNAs were of low abundance (100-1000 copies/cell), and most were found to be nonessential (Parker et al., 1988). One strain in which five different snoRNA genes were disrupted showed no change in growth from the wild-type strain (Parker et al., 1988). A small number of yeast snoRNAs were found to be essential (U3, U14, snR30, RNase MRP) or temperature sensitive (snR10), and all these are involved in various pre-rRNA cleavage steps. Their precise molecular roles in cleavage are still unclear, although U3 and U14 contain rRNA complementary regions that are essential for rRNA processing and viability (Beltrame & Tollervey, 1995; Liang & Fournier, 1995).

The main function of all other characterized snoRNAs is mediating ribosomal RNA modifications. Specifically, the H/ACA box snoRNAs are involved in guiding rRNA pseudouridylations (Ni et al., 1997; Gannot et al., 1997), and most C/D box snoRNAs are involved in guiding rRNA ribose methylations (Kiss-Laszlo et al., 1996; Ni, 1998). A chronology of reviews over the past four years detailing these recent discoveries reflects the substantial progress and excitement in the field (Maxwell & Fournier, 1995; Bachellerie et al., 1995; Maden & Hughes, 1997; Smith & Steitz, 1997; Tollervey & Kiss, 1997; Bachellerie & Cavaille, 1997; Bachellerie & Cavaille, 1998; Ofengand & Fournier, 1998; Weinstein & Steitz, 1999).

1.4.5 Methylation Guide snoRNAs

C/D box snoRNAs involved in ribose methylation contain one or two long 10-21 bp stretches of exact complementarity to ribosomal RNA, and four conserved box features: C, C', D and D' boxes (see Figure 1.1). The C and D box sequence motifs are required for snoRNA

nucleolar localization, accumulation, and association with the ribonucleoprotein particle complexes (RNPs). The C' and D' boxes are necessary for methylation guide function (Kiss-Laszlo et al., 1998). The position of 2'-O-methylation of rRNA is within the helix formed by the complementary guide sequence of the snoRNA, and precisely 5 nt upstream of box D or D'.

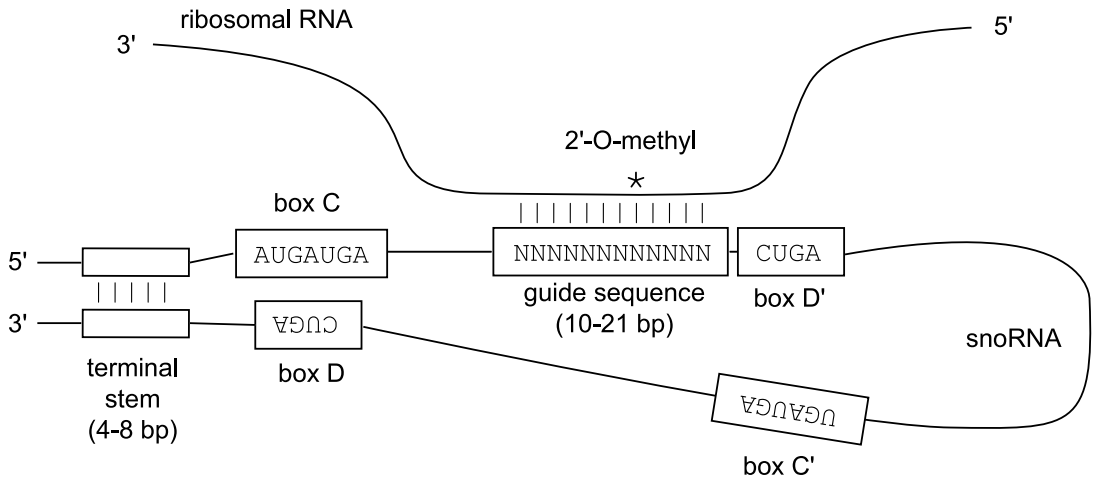

Figure 1.1: **C/D box methylation guide snoRNA.**

Genetic disruption of the U24 snoRNA in *S. cerevisiae* causes loss of the predicted target methyl groups (Kiss-Laszlo et al., 1996). The same study showed that alteration of the rRNA complementary region was sufficient to cause addition of a predictable ectopic methyl at a new position on the rRNA. snoRNA depletion experiments in *Xenopus* oocytes have showed that methylation guide snoRNAs are necessary for specific methylation in vertebrates as well (Tycowski et al., 1996; Dunbar & Baserga, 1998). Methylation guide snoRNAs may also modify other RNAs, including the U6 spliceosomal RNA (Tycowski et al., 1998). A particularly intriguing experiment showed guide snoRNAs could be engineered to modify mRNA by inserting an appropriate complementary region, albeit at low efficiency (Cavaille et al., 1996).

Interestingly, one essential yeast snoRNA, U14, functions in both cleavage and ribose rRNA modification (Jarmolowski et al., 1990), using two different rRNA complementarities (one for cleavage, one for guiding methylation). snR10, an H/ACA box snoRNA, is also a dual function snoRNA, involved in cleavage and pseudouridylation. Surprisingly, all others snoRNAs involved in rRNA modification are non-essential, implying the modifications they specify are not essential. So, just what are all these modifications for? Despite over 30 years of research into rRNA modifications, this question still has no definitive answer.

1.5 Ribosomal RNA Modifications

Posttranscriptional ribosomal RNA modification is common in all branches of the tree of life. There are three basic types of modification found in rRNA: base methylation, ribose methylation, and pseudouridylation. Base methylation is the best conserved in total number and position among all species, with bacteria containing slightly more than the 10 commonly found in eukaryotes (see Table 1.1). Base methylation occurs late in ribosome maturation, and occurs only in highly conserved rRNA sequences. Base methylation within small subunit (SSU) rRNA in prokaryotes is not essential (Krzyzosiak et al., 1987), but it is thought to improve protein translation efficiency (Raue et al., 1988).

Species		Base Methyls	2'-O-ribose Methyls	Pseudouridines	Total
H. sapiens	(E)	10	107	$\sim$ 95	212
X. laevis	(E)	10	99	$\sim$ 98	207
S. cerevisiae	(E)	10	55	44	112
E. coli	(B)	22	4	10	36
S. solfataricus	(A)	$\sim$ 8	67	9	88

Table 1.1: **Comparison of Ribosomal RNA Modifications: Species from Three Phylogenetic Domains** (E) = Eukaryote, (B) = Bacterium, (A) = Archaeon. Data from (Bachellerie & Cavaille, 1998; Ofengand & Fournier, 1998; Noon et al., 1998).

Pseudouridine rRNA modifications (Ψ) are numerous in eukaryotes and few in bacteria

and archaea (Table 1.1). Studies of eukaryotic Ψ residues show they are found in the most evolutionarily conserved regions of rRNA. Ψ are spread throughout SSU rRNA with no clear association with particular functional regions. In contrast, to LSU rRNA Ψ residues are clustered in three main regions, all within or structurally associated with the peptidyl transfer center (PTC) of the ribosome. Individual loss of Ψ residues is not lethal (Ni et al., 1997; Gannot et al., 1997), although global loss of pseudouridylation due to mutations in the putative pseudouridine synthase, Cbf5p, causes temperature-sensitive growth impairment. It is thought Ψ residues play a variety of roles in the ribosome, some improving translational efficiency, others with undetermined function.

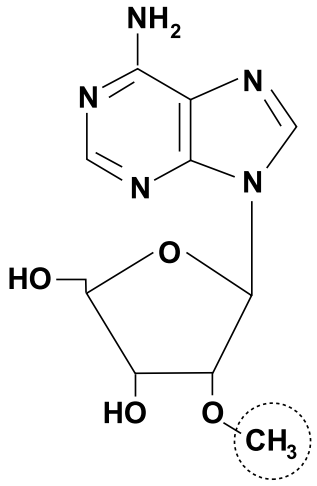

Figure 1.2: **2'-O-methyladenosine**

Ribose methylation, always occurring at the 2' hydroxyl position on the sugar backbone (see Figure 1.2), is frequent in eukaryotes and very limited in bacteria (Table 1.1). Interestingly, ribose methylation in *Sulfolobus solfataricus*, an archaea, is on the order found in eukaryotes (Noon et al., 1998). This contrasts with *S. solfataricus*' bacterial-like paucity of Ψ residues. Ribose methyls occur in highly evolutionary conserved regions of rRNA, in many cases co-clustering near Ψ residues (Maden, 1990). Specific positions of methylation among eukaryotes is well conserved; of the 55 ribose methyls in yeast rRNA, roughly 75%

overlap precisely with mammalian ribose methyls at homologous positions (Maden, 1990). Because most ribose methylation takes place early in rRNA processing, it is hypothesized to be important for rRNA folding or association with chaperone proteins that may aid in folding. No single site of ribose methylation has been found to be essential (Weinstein & Steitz, 1999), although global rRNA demethylation caused by mutation in the NOP1 protein severely impairs growth (Tollervey et al., 1993). In hyperthermophiles, ribose methylation may also be important in thermostability of rRNA and other structural RNA molecules (Noon et al., 1998). One of the goals of this thesis was to learn more about the function(s) of rRNA methylation through study and genetic manipulation of the corresponding guide snoRNAs.

Chapter 2

tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence[1]

[1]This chapter was co-written with Sean Eddy, and appears in Lowe & Eddy, *Nucleic Acids Research* **25**: 955-964, 1997.

2.1 Abstract

We describe a program, tRNAscan-SE, which identifies 99-100% of transfer RNA genes in DNA sequence while giving less than one false positive per 15 gigabases. Two previously described tRNA detection programs are used as fast, first-pass prefilters to identify candidate tRNAs, which are then analyzed by a highly selective tRNA covariance model. This work represents a practical application of RNA covariance models, which are general, probabilistic secondary structure profiles based on stochastic context-free grammars. tRNAscan-SE searches at approximately 30,000 bp/second. Additional extensions to tRNAscan-SE detect unusual tRNA homologues such as selenocysteine tRNAs, tRNA-derived repetitive elements, and tRNA pseudogenes.

2.2 Introduction

Transfer RNA (tRNA) genes are the single largest gene family. A typical eukaryotic genome contains hundreds of tRNA genes; the human genome contains an estimated 1,300 (Hatlen & Attardi, 1971). In a time when complete genomes are being sequenced, one would like to have an accurate means of tRNA gene identification. The tRNA repertoire of an organism affects the codon bias seen in highly expressed protein coding genes. In extreme cases, selective pressure for extremely high or low genomic GC content may have caused loss of a tRNA, producing an unassigned codon (Oba et al., 1991; Kano et al., 1993). Suppressor tRNAs are important genetic loci in many model organisms. In addition to authentic tRNA genes, tRNA-derived short interspersed nuclear elements (SINEs) have been identified in rodents and other mammals as likely mobile genetic elements (Daniels & Deininger, 1985; Deininger, 1989). Detection and discrimination of these elements from true tRNAs is a desirable feature of tRNA identification methods.

It is commonly believed that the best RNA gene detection methods are custom-written programs that search for one type of RNA gene exclusively (Dandekar & Hentze, 1995). Numerous tRNA search programs key on primary sequence patterns and/or secondary struc-

ture specific to tRNAs (Staden, 1980; Paolella & Russo, 1985; Shortridge et al., 1986; Marvel, 1986; Wozniak & Makalowski, 1990; Fichant & Burks, 1991; Pavesi et al., 1994; El-Mabrouk & Lisacek, 1996). Why bother with specialized tRNA-detection software instead of using a fast, commonly available similarity search program such as BLAST (Altschul et al., 1990) or FASTA (Pearson & Lipman, 1988)? Since many functional RNA genes tend to conserve a common base-paired secondary structure better than a consensus primary sequence, the accuracy of RNA similarity searching is much improved by including secondary structure elements. A group of generalized RNA gene search tools look for specific combinations of primary and secondary structure motifs specified by the user (Saurin & Marliere, 1987; Staden, 1988; Gautheret et al., 1990; Sibbald et al., 1992; Laferriere et al., 1994; Billoud et al., 1996; Eddy & Durbin, 1994), although tRNA "descriptors" in these pattern-matching languages have typically under-performed custom-written programs.

tRNAscan 1.3 by Fichant & Burks (Fichant & Burks, 1991) is perhaps the most widely used tRNA detection program. It identifies approximately 97.5% of true tRNA genes and gives 0.37 false positives per million base pairs (Mbp) (Fichant & Burks, 1991). The algorithm uses a hierarchical, rule-based system in which each potential tRNA must exceed empirically determined similarity thresholds for two intragenic promoters, plus have the ability to form base pairings present in tRNA stem-loop structures. The false positive rate of tRNAscan has been acceptable for small genomes, but for larger eukaryotic genomes it becomes a significant problem. It will produce around 1100 false positive tRNAs for the human genome (0.37 false pos/Mbp for 3000 Mbp); given that there are about 1300 true tRNAs in the genome, almost half of the tRNAs predicted by tRNAscan will be false positives.

Pavesi and colleagues have developed a different tRNA detection algorithm (Pavesi et al., 1994) which searches exclusively for linear sequence signals in the form of eukaryotic RNA polymerase III promoters and terminators. The sensitivity and selectivity of this algorithm is roughly comparable to tRNAscan 1.3 in detection of eukaryotic tRNAs. Notably, the Pavesi algorithm identifies tRNAs not detected by tRNAscan 1.3, and vice versa (Pavesi

et al., 1994). The combined sensitivities of these two programs exceed 99%; however, the combined false positive rate is about five times that of tRNAscan alone.

Eddy & Durbin (Eddy & Durbin, 1994) have developed a general RNA structure similarity search method employing probabilistic RNA structural profiles, or "covariance models". Covariance models are able to capture both primary consensus and secondary structure information through the use of stochastic context-free grammars (SCFGs) (Eddy & Durbin, 1994; Grate, 1995; Sakakibara et al., 1994b). Much like sequence profiles (Gribskov, 1994; Krogh et al., 1994), covariance models are constructed from multiple sequence alignments. Sequences are searched against a given covariance model using a three-dimensional dynamic programming algorithm, similar to a Smith-Waterman alignment but including base-pairing terms as well. RNA covariance models have the advantages of high sensitivity, high specificity, and general applicability to any RNA sequence family of interest, obviating the need for custom-written software for each RNA family. However, covariance model dynamic programming algorithms are almost prohibitively CPU-intensive. A tRNA covariance model identifies >99.98% of true tRNAs, with a false positive rate of <0.2/Mbp (Eddy & Durbin, 1994), but searching the human genome with a tRNA covariance model would take about nine and a half CPU-years (based on benchmarks on an SGI Indigo2 R4400/200 CPU, 140 SPECint92).

We describe here a program, tRNAscan-SE, that combines three tRNA search methods to attain the specificity of covariance model analysis with the speed and sensitivities of optimized versions of tRNAscan 1.3 and the Pavesi search algorithm. tRNAscan-SE detects 99-100% of true tRNAs, giving fewer than one false positive per fifteen billion nucleotides of random sequence, at approximately 1,000 to 3,000 times the speed of searching with tRNA covariance models. Additional extensions to tRNAscan-SE allow detection and accurate secondary structure prediction of unusual tRNA species including both prokaryotic and eukaryotic selenocysteine tRNA genes, as well as tRNA-derived repetitive elements and pseudogenes.

2.3 Methods

tRNAscan-SE input consists of DNA or RNA sequences in FASTA format. tRNA predictions are output in tabular, ACeDB, or an extended format including tRNA secondary structure information. tRNAscan-SE does no tRNA detection itself, but instead negotiates the flow of information between three independent tRNA prediction programs, performs some post-processing, and outputs the results (Figure 2.1).

tRNAscan-SE works in three phases. In the first stage, it runs tRNAscan and the Pavesi algorithm on the input sequence. The first of these two programs is an optimized version of tRNAscan 1.3 (Fichant & Burks, 1991). The other is an implementation of the Pavesi search algorithm (Pavesi et al., 1994) which we call EufindtRNA. Results from both programs are merged into one list of candidate tRNAs. Intron information from tRNAscan 1.3 is discarded because its intron predictions are typically unreliable. Analysis with the tRNA covariance model at a later stage (described below) allows non-ambiguous determination of intron boundaries.

In the second stage, tRNAscan-SE extracts the candidate subsequences and passes these segments to the covariance model search program *covels* (Eddy & Durbin, 1994). Seven flanking nucleotides on both sides of the candidate tRNAs are included in the subsequence in case the tRNA was truncated by the initial prediction. The *covels* search program applies a tRNA covariance model (TRNA2.cm) that was made by structurally aligning 1415 tRNAs from the 1993 Sprinzl database (Steinberg et al., 1993). 87 non-canonical "group III" sequences and 509 RNA sequences were removed from the complete 2011 sequence database as described in (Eddy & Durbin, 1994). To improve intron prediction, intron sequences were manually inserted into the Sprinzl alignment for 38 intron-containing tRNAs of known genomic sequence.

Finally, tRNAscan-SE takes predicted tRNAs that have been confirmed with *covels* log odds scores of over 20.0 bits, trims the tRNA bounds to those predicted by *covels*, and runs the covariance model global structure alignment program *coves* (Eddy & Durbin, 1994) to

get a secondary structure prediction. The tRNA isotype is predicted by identifying the anticodon within the *coves* secondary structure output. Introns are identified from this output as runs of five or more consecutive non-consensus nucleotides within the anticodon loop.

tRNAscan-SE uses heuristics to try to distinguish pseudogenes from true tRNAs, primarily on lack of tRNA-like secondary structure. A second tRNA covariance model (TRNA2ns.cm) was created from the same alignment, under the constraint that no secondary structure is conserved (this model is effectively just a sequence profile, or hidden Markov model). By subtracting a tRNA's similarity score to the primary structure-only model from that using the complete tRNA model, a secondary structure-only score is obtained. In Bayesian terms, this difference can be viewed as the evidence for the complete tRNA model, as opposed to a structure-less, sequence-only pseudogene model. We observed that tRNAs with low scores for either component of the total score were often pseudogenes. Thus, tRNAs are marked as likely pseudogenes if they have either a score of less than 10 bits for the primary sequence component of the total score, or a score of less than 5 bits for the secondary structure component of the total score. Selenocysteine tRNAs are not checked by these rules since they have atypical primary and secondary structure. Final tRNA predictions are then saved in tabular, ACeDB, or secondary structure output format.

2.3.1 tRNAscan 1.4

tRNAscan-SE uses an optimized version of tRNAscan 1.3 (Fichant & Burks, 1991) which we refer to as tRNAscan 1.4. The core algorithm is identical to tRNAscan 1.3. tRNAscan versions 1.3 and 1.4 have identical tRNA detection rates except in the case of ambiguous nucleotides occurring within the input sequence. There are implementation errors in tRNAscan 1.3's handling of ambiguous nucleotide codes. tRNAscan 1.4 conservatively calls ambiguous nucleotides as always forming base pairings in stems, and matching the highest scoring choice in consensus promoter matrices. This results in a high false positive rate for sequences containing a large number of ambiguous nucleotides. For our purposes, this

is acceptable because the second stage covariance model analysis eliminates false positives. Several command line options were added to tRNAscan 1.4 for convenience in integration with tRNAscan-SE. Additional code changes were made to increase the robustness and speed of the program. These modifications result in roughly a 650-fold increase in search speed and no upper limit on input sequence size.

2.3.2 Implementation of EufindtRNA

EufindtRNA was implemented from the published algorithm by Pavesi and colleagues (Pavesi et al., 1994). The step-wise algorithm uses four probabilistic profiles for identifying basic tRNA features: 'A box' nucleotide composition, 'B box' composition, nucleotide distance between identified A and B boxes, and distance between identified B boxes and RNA polymerase III termination signals (four or more consecutive thymine nucleotides). In a search, an "intermediate" score is obtained by adding scores from identified A and B boxes to the score for the nucleotide distance between them. A final score is obtained by adding the intermediate score to the score for the distance to the nearest termination signal. If the final score is above a specific cutoff, the tRNA identity and location are saved.

Scores from over 30 example tRNAs described in the original publication match our implementation to within 0.1 log odds units. tRNAscan-SE uses a less selective version of the algorithm described above which does not search for transcription termination signals; instead, the intermediate score is used as a final cutoff. Also, the intermediate score cutoff is loosened slightly to -32.10 relative to the intermediate cutoff described in the original algorithm, -31.25. Although the program is designed for eukaryotic tRNA detection, we found EufindtRNA to be effective at identifying prokaryotic tRNAs if the intermediate cutoff score is further adjusted. tRNAscan-SE has a specific option (-P) for scanning prokaryotic sequences which loosens the intermediate cutoff score to -36.0. Also, as with tRNAscan 1.4, ambiguous nucleotides are automatically assigned the best of the four non-ambiguous nucleotide scores at that position in the scoring matrices.

2.3.3 Selenocysteine tRNA Identification

The primary and secondary structure of selenocysteine tRNAs differ from canonical tRNAs in several respects, most notably an eight base pair acceptor stem, a long variable region arm, and substitutions at several well-conserved base positions. These differences make detection and accurate secondary structure prediction difficult using tRNA search programs geared towards canonical tRNAs. tRNAscan 1.3 fails to detect most selenocysteine tRNAs; the Pavesi algorithm incorporates a separate routine specifically for eukaryotic selenocysteines; and the TRNA2.cm covariance model barely detects selenocysteine tRNAs, giving scores just over the minimum cutoff of 20 bits, and in two cases, below the cutoff. tRNAscan-SE addresses this problem in the first-pass stage using EufindtRNA modifications, and in the second stage using selenocysteine tRNA-specific covariance models.

The first-pass scanner EufindtRNA implements a specialized subroutine described by Pavesi *et al.* (Pavesi et al., 1994) for identifying eukaryotic selenocysteine tRNAs (based on a B box score with a value between -2.2 and -3.6, and the motif GGTC(C/T)G(G/T)GGT appearing 36 nucleotides upstream of the B box). To similarly identify prokaryotic selenocysteine tRNAs, a subroutine was added to EufindtRNA which detects tRNAs with B box scores between -2.2 and -4.9, and a conserved sequence motif found in the anticodon loop of all known prokaryotic selenocysteine tRNAs (anticodon in bold): GG(A/T)(C/T)-TTCAAA(A/T)CC. It is unclear if this motif will generalize well for new selenocysteine tRNAs, but it is conserved among the closely related *Escherichia coli* (Y00299), *Proteus vulgaris* (X14255), *Haemophilus influenzae* (U32753), and *Desulfomicrobium baculatus* (X75790) tRNAs, and in the more distant *Clostridium thermoaceticum* (Z26950) tRNA. After EufindtRNA has identified a candidate selenocysteine tRNA, it is passed to a eukaryotic or prokaryotic selenocysteine-specific covariance model. These two covariance models were developed by aligning selenocysteine tRNAs with inferred secondary structure information. Another program in the covariance model program suite, *coveb*, was used to build covariance models from the structure-annotated RNA sequence alignments. The five prokaryotic

tRNAs noted above were used to build the prokaryotic selenocysteine model. Seven selenocysteine tRNAs from *Caenorhabditis elegans*, *Drosophila melanogaster*, *Xenopus laevis*, chicken, mouse, bovine, and human were used to build the eukaryotic model.

2.3.4 Databases Tested

tRNA detection rates were assessed primarily by searching two annotated databases: the 1995 release of the Sprinzl tRNA database (retrieved from ftp://ftp.ebi.ac.uk/pub/databases/trna (Steinberg et al., 1993)), and a tRNA sequence subset of Genbank (retrieved from the National Center for Biotechnology Information on 9/24/96). Genomic DNA was also searched from *Haemophilus influenzae* (v. 1.0, from the Institute for Genome Research (TIGR) (Fleischmann et al., 1995), *Mycoplasma genitalium* (Fraser et al., 1995), *Methanococcus jannaschii* (Bult et al., 1996), *Saccharomyces cerevisiae* (rel. 4/24/96), *Schizosaccharomyces pombe* (completed cosmids retrieved from http://www.sanger.ac.uk/~yeastpub/svw/pombe.html on 9/30/96), *C. elegans* (completed cosmids retrieved 11/13/96 from ftp://ftp.sanger.ac.uk/pub/C.elegans/sequences), and Human (completed cosmids retrieved 8/28/96 from ftp://ftp.sanger.ac.uk/pub/human).

The Sprinzl tRNA database is the most comprehensive tRNA database, containing 2700 entries from a wide variety of organisms (Steinberg et al., 1993). It provides a set of trusted "true positives" for evaluating the sensitivity of a detection method. Since tRNAscan-SE was optimized for analyzing bacterial, archaeal, and eukaryotic genomic DNA, the 1144 tRNAs from species in these groups were chosen for analysis, excluding mitochondrial, chloroplast, and viral tRNA sequences. From this set, tRNAs that were used to train the TRNA2.cm covariance model (553 tRNAs in the 1993 release of the database) were removed to increase the independence between training and testing sequence data. Entries were restored to their correct primary sequence by combining the Sprinzl structural alignment with the atypical insertions that are annotated in a separate file. Introns, not present in the Sprinzl sequences or annotation, were not restored. Two prokaryotic sequences (DI1950, DR1420) were removed which would contain introns over 200 bp long had introns

been included; none of the current tRNA search programs attempt to detect tRNA genes containing long group I or group II introns.

A broad sample of non-viral, non-organellar Genbank sequences indicating at least one tRNA in their feature tables was also analyzed. *C. elegans* and *S. cerevisiae* sequences were excluded since these genomic sequences were tested separately. The sequences were retrieved using the IRX query system at the National Center for Biotechnology Information (NCBI). Incomplete or synthetic tRNA sequences were removed, yielding a total of 1051 in the set. Genbank sequence annotation was not relied upon as a measure of the true number of tRNAs in the set since annotation quality is highly variable. Instead, tRNA detection by covariance model analysis was used to estimate the total number of tRNAs. Sequences with no tRNAs detected by covariance model analysis were manually examined to determine why annotated tRNAs were not detected, and six believed to be tRNAs were added to the covariance model-detected set. This method gave us a reasonable lower bound on the number of true positives in the Genbank subset.

2.3.5 "Random" Sequence Data

Two types of random sequence databases were created to test false positive rates. The first database is generated by a fifth order Markov chain based on six-mer frequencies within the first 54 Mbp of genomic sequence from the *C. elegans* genome project. Two thousand cosmid-sized sequences, 50 kilobases (Kbp) each, were generated based on these frequencies, totaling 100 Mbp of random sequence which is tRNA-free. The second random database was created to roughly simulate the human genome in size and GC content. Not enough human genomic sequence is available to parameterize a fifth order Markov chain model, so human sequence was simulated based on isochore proportion and %GC content. Ten thousand 300 Kbp sequences were generated, each one with a GC content approximating one of the five isochore types (L1 or L2 = 40% GC, H1 = 45% GC, H2 = 49% GC, H3 = 53% GC; (Green & Vold, 1993)). The isochore identities for these random sequences were chosen to approximate the proportion each isochore represents in the human genome (L1 +

L2 60%, H1 20%, H2 10%, H3 5%). The remaining 5% of the human genome attributed to ALU-type repeat elements were not included since ALU sequences were tested separately (the absent 5% was distributed proportionally among the other isochore types).

2.3.6 Implementation & Online Analysis

tRNAscan-SE was written in Perl. The implementation of the Pavesi algorithm (Pavesi et al., 1994), EufindtRNA, was written in C. A single package of the UNIX-based programs used by tRNAscan-SE is available at http://www.genetics.wustl.edu/eddy/software. All analysis times given are for a Silicon Graphics Indigo2 R4400 200 Mhz workstation. A web server is available for on-line tRNA analysis at http://www.genetics.wustl.edu/eddy/tRNAscan-SE/.

2.4 Results

A summary of the overall sensitivity, selectivity, and search speed for the four tRNA search programs tested is shown in Table 2.1. The number of true positives is based on the percentage of tRNAs detected within a test set taken from the Sprinzl tRNA database (see Table 2.2). The false positive rate is based on analysis of randomly generated sequence data (Table 2.4). The search speeds for the various programs are shown for a scan of the current *C. elegans* genomic sequences averaging 30 kilobases per clone. tRNAscan 1.3 search speed decreases approximately linearly with length. Search speed for tRNAscan-SE is approximately constant, but varies based on tRNA density within the sequence.

2.4.1 Sensitivity

tRNAscan-SE was shown to be more sensitive than tRNAscan 1.3 by several measures, the first being a search of the Sprinzl and Genbank databases subsets (Table 2.2). In the Sprinzl test set, tRNAscan-SE detected 586 of 589 known tRNAs (99.5%), versus 560 of 589 (95.1%) for tRNAscan 1.3. Of all 1144 non-organellar tRNAs in the complete Sprinzl

Search Method	True Positives (%)	False Positives (per Mbp)	Search Speed (bp/sec)
tRNAscan 1.3	95.1	0.37	400
EufindtRNA	88.8	0.23	373,000
tRNA covariance model search	99.8	< 0.002	20
tRNAscan-SE	99.5	< 0.00007	30,000

Table 2.1: **Overall detection rates of tRNA search programs.**
True positives are based on detection rates within a non-organellar, non-viral subset of the Sprinzl tRNA database (see Table 2.2). False positive rates are estimates based on searches of randomly generated human sequence (see Table 2.4). Search speeds are from a search of 58.4 Mbp of *C. elegans* cosmid sequences on a Silicon Graphics Indigo2 R4400 200 Mhz workstation.

database, tRNAscan-SE fails to recognize seven. One was a eukaryotic sequence from Trypanosoma brucei (Sprinzl ID DT6050, Genbank TBTRNA3) which has been previously noted by Pavesi *et al.* (1994) as being missed by both tRNAscan 1.3 and the Pavesi search algorithm. The other six tRNAs missed by tRNAscan-SE were from various eubacteria (Sprinzl ID's: DA1543, DE2180, DG1351, DG1482, DS1250, RG1380). Several of these undetected tRNAs appear to be irregular in source or function. DE2180 is derived from DNA from the cyanelle (a photosynthetic organelle) of the unicellular eukaryote *Cyanophora paradoxa* and is thus misclassified as eubacterial in the database. DG1482 and RG1380 both contain substitutions of four highly conserved bases within the TψC loop, an indication that the tRNAs are probably used in synthesis of the peptidoglycan instead of protein translation (29). All seven of these atypical tRNAs were detected using covariance model analysis. The tRNA covariance model search does miss two tRNAs within the 1144-member Sprinzl database subset, both selenocysteine tRNAs (Sprinzl ID DZ1430 & DZ7742) that pass below the 20.0 bit cutoff at 0.60 and 14.19 bits, respectively. EufindtRNA, designed to search eukaryotic sequences exclusively, shows improved sensitivity for eukaryotic tRNAs (98.6%) over tRNAscan 1.3 (95.0%), but is still slightly less sensitive than tRNAscan-SE (100%). Over the three phylogenetic domains, tRNA covariance model analysis appears to

Sequence Source	Literature tRNAs	tRNAscan 1.3 Tot	(%)	EufindtRNA Tot	(%)	tRNA CM Tot	(%)	tRNAscan-SE Tot	(%)
Sprinzl db (Archaea)	70	69	(98.6)	43	(61.4)	70	(100)	70	(100)
Sprinzl db (Eubacteria)	240	226	(94.2)	205	(85.4)1	239	(99.6)	237	(98.7)
Sprinzl db (Eukarya)	279	265	(95.0)	275	(98.6)	279	(100)	279	(100)
Sprinzl db (total)	589	560	(95.1)	523	(88.8)	588	(99.8)	586	(99.5)
Genbank tRNA subset	1462	1366	(93.4)	760	(52.0)	1456	(99.6)	1440	(98.5)

Table 2.2: **tRNA prediction within annotated database subsets.**
The detection rates for the Sprinzl tRNA database are broken down by phylogenetic domain. The Sprinzl subset tested contains only non-organellar, non-viral tRNAs which were not used in training of the tRNA covariance model. For the Sprinzl database subset, numbers in parentheses indicate percentage of correct tRNA identifications relative to total in the literature. The Genbank subset sequences were selected by retrieving non-organellar, non-viral, full-length tRNA sequences with "tRNA" indicated in the feature field of the entry. Since Genbank tRNA annotation is less reliable, the numbers in parentheses for this row are the percentage of correct tRNA identifications relative to all tRNAs verified by either covariance model analysis or visual inspection.

be the most sensitive detection method, yet tRNAscan-SE trails by as little as one third of one percentage point.

Searching the Genbank subset sequences which contain less reliable tRNA annotation, tRNAscan-SE detects 98.5% of the 1462 tRNAs verified by either covariance model analysis or visual inspection, whereas tRNAscan 1.3 has a 93.4 % detection rate (Table 2.2). All prediction discrepancies were visually inspected. Of the 18 tRNAs that covariance model analysis detected but were missed by all three other methods, all had scores over 36 bits, and were annotated in the Genbank entries. The two tRNAs detected by tRNAscan-SE but missed by covariance model analysis were a selenocysteine tRNA (CTTRSEL; same as previously noted Sprinzl DZ1430 tRNA), and a long tRNA from *Haloferax volcanii* (HALTGW) whose 104 bp intron caused the tRNA to exceed the maximum total length limit for normal tRNA covariance model analysis (150 bp). Of the 9 sequences annotated as tRNAs but missed by all four detection methods, four have large group I or group II introns of 241 bp or larger (ANATGL, SSU10482, PHU29955, SYOTRNLUAA), and five appear to have either sequencing errors or modified bases which appear in the Genbank annotation but not

in the sequence (corresponding tRNAs within the Sprinzl database were identified correctly by all four detection methods). Because of sequence discrepancies between the Genbank sequences and corresponding Sprinzl entries, these five Genbank tRNAs were not included in the 1462-member test set.

2.4.2 Genome Analysis

Another measure of sensitivity was derived from searching complete or partial genomic sequence data from eubacterial, archaebacterial, yeast, and *C. elegans* sequencing projects (Table 2.3). For *M. genitalium*, 33 tRNAs were noted in the published (Fraser et al., 1995) and on-line gene identifications (http://www.tigr.org/tdb/mdb/mgdb/mgdb.html), whereas 36 tRNAs were detected by three tRNA detection methods (tRNAscan 1.3, tRNAscan-SE, covariance model analysis). The three tRNAs not appearing in the literature are for Arg (anticodon: CCT, bounds: 306615-306686, upper strand), Leu (anticodon: CAA, bounds: 448783-448861, upper strand), and Leu (anticodon: GAG, bounds: 446265-446181, reverse strand). For the completed *H. influenzae* genome, 56 tRNAs are noted in the literature and on-line gene identifications (Fleischmann et al., 1995). tRNAscan-SE and covariance model analysis both identify the tRNAs noted in the literature, plus two potentially novel tRNAs not noted in the literature: SelCys (anticodon: TCA, bounds 753881-753791), and Leu (anticodon: GAG, bounds 1577041-1576960). The first is a selenocysteine tRNA and the other appears to be either a pseudogene or a true tRNA containing a short intron. [Note: Since publication of these results (Lowe & Eddy, 1997), TIGR has adopted our program for tRNA analysis, and updated their annotation.] The selenocysteine tRNA identification is not unexpected; BLAST searches identify two enzymes in the selenocysteine insertion pathway, as well formate dehydrogenase containing a 'UGA' selenocysteine-insertion codon. The evidence for the other potentially novel tRNA is less certain. The short 12 bp "intron" would presumably require protein-splicing to generate a functional tRNA, a feature that would be novel among eubacterial tRNAs. However, the covariance model score of 36.88 bits for the tRNA is well above the minimum cutoff of 20 bits, indicating that the sequence is likely to

Sequence Source	Size (Kbp)	Literature tRNAs	tRNAscan 1.3 Tot	(%)	EufindtRNA Tot	(%)	tRNA CM Tot	(%)	tRNAscan-SE Tot	(%)
M. genitalium	580	33	36	(100)	19 + 1 fp	(52.8)	36	(100)	36	(100)
H. influenzae	1,830	56	55	(98.2)	42 + 2 fp	(73.7)	58	(103.6)	58	(103.6)
M. jannaschii	1,730	37	36	(97.3)	20 +1 fp	(54.0)	37	(100)	37	(100)
S. pombe (through 9/96)	4,176	–	45 +4 fp	(93.7)	46 +1 fp	(95.8)	48		48	(100)
S. cerevisiae	12,057	273	270 +4 fp	(98.5)	274 +10 fp + 1 ps	(100)	274 +1 ps		274 +1 ps	(100)
C. elegans (through 11/13/96)	58,402	– 16 fp	389 +29 fp	(96.5)	400 +355 fp + 19 ps	(99.2)	403 +23 ps		403 +11 ip + 8 ps	(100)
P. anserina mitochondrion	100	27	18	(66.7)	11	(40.7)	27	(100)	22	(81.5)

Table 2.3: **tRNAs identified in genomic databases by various search methods.** "Literature" column represents the published number of tRNAs found within genomes. "Tot" columns indicate total number of tRNAs found in searches for each program. Numbers in parentheses in (%) columns indicate percentage of tRNAs detected relative to literature (*H. influenzae, M. jannaschii, P. anserina*), or when published tRNA annotation is incomplete or uncertain (*M. genitalium, S. pombe, S. cerevisiae, C. elegans*), detection percentages are relative to total tRNAs found by tRNA covariance model analysis and supported by manual inspection. "fp" = false positives determined by covariance model analysis and manual inspection (these do not include pseudogenes that have strong similarity to known tRNAs). "ps" = tRNA identifications which appear to be pseudogenes containing 5' truncations of 3-16 bp, large insertions or deletions elsewhere, or other characteristics of tRNA-derived repetitive elements. "ip" = tRNAs automatically identified by tRNAscan-SE as likely pseudogenes which have qualities similar to manually detected pseudogenes described above.

have evolutionary homology with tRNA. It is possible that it is a pseudogene. tRNAscan 1.3 identifies 55 of the 56 tRNAs noted in the literature (Gly-B, by TIGR nomenclature, is not detected), and does not detect either of the novel tRNAs detected by tRNAscan-SE and covariance model analysis.

The genomic sequence of the archaebacterium *M. jannaschii* was also analyzed. Both tRNAscan-SE and covariance model analysis identified all 37 tRNAs as given in the literature (Bult et al., 1996). tRNAscan 1.3 identified 36 of the 37 tRNAs, missing the single selenocysteine tRNA in the set. We also scanned the recently completed genomic sequence of the budding yeast *S. cerevisiae* (12 Mbp). The covariance model search took 14 days to complete, and produced 275 tRNAs. Based either on inspection for ability to form correct tRNA secondary structure, or exact identity with previously characterized yeast tRNAs, we believe 274 predicted tRNAs are true tRNAs, and one is a pseudogene with an 7 bp 5' truncation. One of these 274 tRNAs was missing from the yeast genome project web site annotation http://www.mips.biochem.mpg.de/), but this is probably an oversight since a tRNA of identical sequence is correctly annotated elsewhere in the genome (tRNA_i_S (GCT)LR2). tRNAscan-SE took 19 minutes and detected the same 275 tRNAs found by covariance model analysis. EufindtRNA found the same 275 tRNAs in just over one minute. tRNAscan 1.3 took about 10 hours to complete, and missed 4 (2 pairs identical in sequence) of the 274 true tRNAs found by the other three methods. The 4 Mbp of available genomic sequence from Schizosaccharomyces pombe (fission yeast) was also analyzed. tRNAscan-SE and covariance model analysis both predict 48 tRNAs. tRNAscan 1.3 identifies 45 of the 48 predicted by covariance model analysis (2 of 3 missed were identical in sequence), whereas EufindtRNA identifies 46 of the 48 total tRNAs.

Finally, we scanned the largest set of genomic sequence currently available, 58.4 Mbp from the *C. elegans* genome project. Since only a handful of the tRNAs detected have been previously published in the literature, we again relied on covariance model detection of tRNAs as our best measure for "true" tRNAs. Conflicts in tRNA predictions between tRNAscan 1.3, tRNAscan-SE and covariance model analysis were all examined manually

for highly conserved primary sequence motifs and proper secondary structure. As most tRNA species are multicopy in eukaryotes, BLAST similarity searches were used to help discern "false positives" from pseudogenes. We define false positives as predicted tRNAs which do not appear to be evolutionarily derived from true tRNAs. These false positives are assessed by failure to form recognizable tRNA secondary structure and the lack of related tRNAs elsewhere in the genome. Pseudogenes, on the other hand, usually have at least partial tRNA secondary structure, plus clear deletions or insertions relative to at least one related, intact tRNA elsewhere in the genome. tRNA-derived mobile elements also have recognizable primary sequence similarity to tRNAs, although most have poor tRNA secondary structure similarity. Of the 403 complete tRNAs detected by covariance model analysis, tRNAscan-SE detected all 403 tRNAs (100%), whereas tRNAscan 1.3 detected 389 (96.5%), and EufindtRNA found 400 (99.2%).

Taken together, the data analyzed from the *M. genitalium*, *H. influenzae*, *M. jannaschii*, *S. cerevisiae*, *S. pombe*, and *C. elegans* genomes, 100% of the 856 tRNAs detected by covariance model analysis were found by tRNAscan-SE. tRNAscan 1.3 detected 831, missing 25 tRNAs identified by covariance models, a 97.1 % detection rate. EufindtRNA detects 93.5% of the 856 tRNA set, but if only eukaryotic genomes are considered, the program finds 720 of 725 (99.3%).

2.4.3 Selectivity

While the "sensitivity" of an algorithm is measured by the proportion of true positives identified in reference sequences, a method's "selectivity" is measured by its ability to avoid misidentifying unrelated sequences as true tRNAs. Increased sensitivity is usually gained at the expense of an increased false positive rate. A rate of one false positive per five to ten million bases of sequence has, in the past, been acceptable since the total amount of uncharacterized or non-protein coding sequence in the databases has been relatively small. However, with the advent of whole-genome sequencing projects on the megabase scale, this false positive rate is of much greater concern.

	Size (Mbp)	tRNAscan 1.3		EufindtRNA		tRNA CM		tRNAscan-SE	
		FP	FP/Mbp	FP	FP/Mbp	FP	FP/Mbp	FP	FP/Mbp
***S.* cerevisiae**									
Actual FP (completed genome)	12.0	4	0.33	10	0.83	0	< 0.08	0	< 0.08
***C.* elegans**									
Actual FP (portion completed)	58.4	29	0.50	355	6.08	0	< 0.03	0	< 0.03
Simulated FP (total genome)	100	42.5	0.42	26	0.26	0	< 0.01	0	< 0.001
Human									
Actual FP (portion completed)	5.32	3	0.56	5	0.94	0	< 0.19	0	< 0.19
Simulated FP (total genome)	3000	1118	0.37	684	0.23	ND	–	0	< 0.00007

Table 2.4: **False positive rates for actual & simulated genomes.**
"Actual FP" rows contain false positives detected in actual genomic sequence. "Simulated FP" rows contain the false positives found in whole-genome scale random sequence simulations (10 trials for *C. elegans*, 5 for human). For tRNA covariance model searches (tRNA CM), only one random *C. elegans* and no human genome simulations were performed due to extreme CPU demands (ND=not done).

Assessing the ability of an algorithm to discriminate between true and false positives using biological sequence data can be difficult. At false positive rates of less than one per million bases, there is not enough well annotated sequence in the public databases to give a reliable indication of an algorithm's true performance. Even for the data that is available, it is uncertain whether or not an accurate prediction has been made in the absence of biochemical experimental evidence. An alternative strategy is to generate random nucleotide sequence which is known to have no biologically-derived genes. An unlimited amount of random sequence can be generated based on a general or species-specific genomic nucleotide frequency. Each identification of a tRNA gene in this random sequence can then be confidently counted as a false positive. False positives due to biologically-derived repetitive elements or pseudogenes are not taken into account in these synthetic test sequences, and must be addressed separately.

We generated two types of random sequence sets to simulate the size and GC content of the *C. elegans* and human genomes (100 million and 3 billion bases of random sequence, respectively, as described in Methods). The number of false positives found with each al-

Complete Genome	Size (Mbp)	tRNAscan 1.3 (CPU hours)	EufindtRNA (CPU hours)	tRNA CM (CPU hours)	tRNAscan-SE (CPU hours)
P. anserina mito	0.1	0.14	< 0.001	2.8	0.019
H. influenzae	1.8	2.54	< 0.001	51	0.069
S. cerevisiae	12	16.7	0.02	333	0.33
C. elegans	100	139	0.15	2,780	1.8
Human	3,000	>4170	7.1	83,300	36.6

Table 2.5: **Analysis time in hours required for various complete genomes and tRNA search algorithms.**
Actual genome scan times are given for tRNAscan-SE and EufindtRNA (genome simulation times used for human). Estimated scan times are given for tRNAscan 1.3 (400 bp/s) and tRNA covariance model analysis (tRNA CM; 20 bp/s).

gorithm appear in Table 2.4 along with false positive rates from actual genomic sequence (discussed below). Analysis of the simulated genomes gave consistent false positive rates between the various trials, at approximately 0.40 false positives per million bases for tRNAscan 1.3, a little more than half that for EufindtRNA, and zero for both tRNAscan-SE and covariance model analysis. In ten independent *C. elegans* genome simulations, an average of 42.5 tRNAs were identified by tRNAscan 1.4. The sequences for the false positive tRNAs were saved and analyzed with the original tRNAscan 1.3 program to confirm that false positives were due to the tRNAscan 1.3 algorithm, not the modifications introduced in tRNAscan 1.4. EufindtRNA misidentified an average of 26 false positives per simulated *C. elegans* genome. Both tRNAscan-SE and the tRNA covariance model searches found zero positives for every trial (only one genome simulation was searched with the tRNA covariance model due to the extreme CPU demands). As seen in Table 2.5, minor differences among analysis times for the various methods for microbial genomes become substantial when analyzing larger eukaryotic genomes. Analysis of the single *C. elegans* genome simulation with covariance models required almost four CPU-months.

For the five human genome simulations, tRNAscan 1.4 produced an average of 1118 false positives per genome (had tRNAscan 1.3 been used, it would have taken almost half a CPU year per trial). EufindtRNA searched the simulated genomes in just over seven hours

per trial, giving an average of 684 falsely predicted tRNAs for each. Had we searched the entire 3 billion nucleotide human genome simulation with tRNA covariance model analysis, it would have taken over nine CPU-years for each trial (Table 2.5). Based on the histogram of covariance model scores against 500 million bases of simulated human sequence data (not shown), we estimate that the tRNA covariance model search of the simulated human genome would have produced zero false positives. tRNAscan-SE required an average of a day and a half to scan each of the three billion nucleotide test sets, and produced no false positives in any of the five trials (the exact same sequences were used as in the trials described above for tRNAscan 1.4 and EufindtRNA).

A concern not addressed by the random sequence genome simulations is the "false positive" rate caused by certain classes of SINEs that are suspected to be derived from tRNA genes (Daniels & Deininger, 1985; Deininger, 1989). These elements have similarity to known tRNA genes and contain well conserved RNA polymerase III internal A and B box promoters. To assess tRNAscan-SE's ability to identify and exclude these types of pseudo-tRNAs, the repeat element database Repbase maintained by Jerzy Jurka (ftp://ncbi.nlm.nih.gov/repository/repbase) was scanned. Of the reference sequences searched, tRNAscan-SE did not produce any false positive tRNA identifications. Covariance model analysis, however, did misidentify 12 of 775 rodent B2 SINE sequences and two ALU-like sequences (bovine ALU-like repetitive element & rat ALU type III-like repetitive element), all with scores between 20 and 28 bits. Rat identifier (ID or R.dre.1) sequences, also known to have high similarity to alanine, proline, and other tRNAs, were searched within Genbank and dbEST (database of expressed sequence tags, (Boguski et al., 1993)). tRNAscan-SE misidentified four rat ID element sequences total, one from Genbank (RATRSIDH) and three from dbEST (R46943, R46943, R82886). The extreme sensitivity of covariance model analysis is also unable to distinguish between these SINEs and true tRNAs, giving bit scores between 24.5 and 33.1 bits. tRNAscan 1.3 requires strong adherence to secondary structure rules, thus does not call any of these pseudogenes as tRNAs. The rest of Repbase, including consensus and database collections of ALU, L1, THE, MIR, MIR2,

THR, and B1 repetitive elements, were also searched with tRNAscan-SE, giving no other false positives.

The selectivity of tRNAscan has already affected genome sequence annotation detrimentally. In 58.4 Mbp of *C. elegans* genomic sequence, tRNAscan 1.3 produced 29 tRNAs which were judged to be false positives (0.50 fp /Mbp) based on searching with the tRNA covariance model, visual inspection of secondary structure, and lack of primary sequence similarity to any other tRNAs within the genome. Since both the Washington University Genome Sequencing Center (St. Louis) and the Sanger Center (Cambridge, UK) used tRNAscan 1.3 in semi-automated sequence annotation until very recently, 16 of these 29 false positives are annotated as tRNAs in finished, submitted Genbank entries. This false positive rate is very close to that seen in the random *C. elegans* genome simulation (0.42 fp/Mbp), giving additional confidence to the estimates based on simulated sequence data.

tRNAscan-SE produced no obvious false positives in the *C. elegans* genomic sequence, but did identify 8 tRNAs that were judged to be possible pseudogenes by manual inspection (Table 2.3). Eleven other tRNAs were automatically identified as pseudogenes via primary or secondary structure scores that fell below minimum values described in the methods. All 19 pseudogenes had strong similarity to other tRNAs within the genome, and contained unusual features such as 3-16 bp truncations of the 5' end of the gene, or other large insertions or deletions within the sequence. One could consider detection of these possible pseudogenes a desirable feature of tRNAscan-SE's sensitivity. Further studies of these unusual tRNAs may help better elucidate aspects of genome dynamics, genetic element mobility, and evolution.

2.4.4 Selenocysteine tRNA Detection

There are not enough selenocysteine tRNA sequences to properly evaluate tRNAscan-SE's selenocysteine detection accuracy. Three selenocysteine tRNAs (one each from *H. influenzae, M. jannaschii*, and *C. elegans*) were detected in recent genome sequence data. The *H. influenzae* tRNA, previously unrecognized in the literature, was detected by the prokaryotic

selenocysteine-specific routines and covariance model. The tRNA from the distantly related *M. jannaschii*, however, was detected by the standard EufindtRNA algorithm and general tRNA covariance model. The failure of the specialized routines may have been due in part to the fact that this is the first and only archaebacterial selenocysteine tRNA available to date. For the remaining non-archaeal selenocysteine tRNAs, use of the specialized models boosts covariance model scores from the 20-40 bit range to 45-72 bits. Since accurate t-RNA secondary structure prediction relies on correct alignment of the tRNA sequence to the covariance model, use of selenocysteine-specific models for these tRNAs improves the accuracy of structure predictions. A search of the non-redundant database (nrdb) maintained at NCBI revealed no new selenocysteine tRNAs from species for which there was no previously noted sequence.

2.4.5 Intron Detection

tRNAscan-SE correctly predicted the introns for the 13 species of intron-containing tRNAs in the *S. cerevisiae* genome (Westaway & Abelson, 1995). tRNAscan 1.3 often gives multiple intron predictions for each tRNA, making correct placement uncertain. EufindtRNA does not attempt to predict intron boundaries at all (Pavesi et al., 1994).

Detection of tRNAs containing long introns, usually group I or group II, is problematic. The default maximum tRNA length for tRNAscan-SE is 192 bp, but this can be increased (option -L <max length>) to allow searches with no practical limit on tRNA length. In the first phase of tRNAscan-SE, EufindtRNA searches for A and B boxes of the specified maximum distance apart, and passes only the 5' and 3' tRNA ends to covariance model analysis for confirmation (removing the bulk of long intervening sequences). Using this option, tRNAscan-SE was able to detect three of the four long tRNAs initially missed by all four methods in the Genbank tRNA subset search (the fourth tRNA was undetectable with EufindtRNA even with the intron removed before analysis). Group I or II introns in tRNAs tend to occur in positions other than the canonical position of protein-spliced introns, so tRNAscan-SE mispredicts the intron bounds and anticodon sequence for these

cases. 5' and 3' tRNA bounds were correct for all three unusual tRNAs.

2.4.6 Performance on Mitochondrial tRNAs

Although tRNAscan-SE was designed with non-organellar tRNA detection in mind, we also tested it on a complete mitochondrial genome, that of Podospora anserina (Genbank ID PANMTPACGA). tRNAscan-SE detected 22 of the 27 annotated tRNAs (81.5 %), tRNAscan 1.3 detected 18 of 27 (66.7%), and covariance model analysis detected all 27 tRNAs (Table 2.3). Since organellar genomes are usually small, the computational demand of covariance model analysis alone (without the use of fast first-pass scanners) is not prohibitive. For this reason, tRNAscan-SE can be run in covariance model analysis-only mode (-C option) for maximum sensitivity, bypassing dependence on tRNAscan 1.4 and EufindtRNA. This mode gives the same results as would be obtained by running the covariance model search program alone, but in addition, produces annotated tRNA output identical in format to that found in the default tRNAscan-SE search mode.

2.5 Discussion

2.5.1 Speed, Sensitivity, and Selectivity

The most sensitive and selective tRNA detection method that we are aware of utilizes probabilistic RNA covariance models (Eddy & Durbin, 1994), which are based on stochastic context-free grammar techniques. However, searching with covariance models has two drawbacks. First, it is extremely CPU-intensive, requiring days to weeks of processor time to scan megabase-size genomic data from higher eukaryotes. Second, the general nature of the approach hampers output of tRNA-specific feature information such as anticodon, isotype, and intron position. Our goal in the development of tRNAscan-SE was to produce a practical (*i.e.,* fast) application of stochastic context-free grammar-based RNA analysis methods with sensitivity and selectivity as close as possible to using native covariance model searches. tRNAscan-SE achieves this goal.

tRNAscan-SE increases tRNA covariance model search speed by 1,000 to 3,000 fold while offering nearly equal sensitivity and slightly improved selectivity. Selenocysteine tRNA detection features are built into tRNAscan-SE, including modifications to EufindtRNA and the use of selenocysteine tRNA covariance models. With these additions, tRNAscan-SE correctly identifies both of the selenocysteine tRNAs in the Sprinzl database not detected by normal covariance model analysis. The Genbank version of one of these two selenocysteine tRNA sequences, CTTRSEL from *C. thermoaceticum*, was also detected within the Genbank tRNA subset (the other selenocysteine tRNA was not in the Genbank subset).

tRNAscan-SE also extends the maximum length of tRNAs detectable to almost any length. In covariance model analysis, search time increases as the square of the maximum tRNA length, so the search window has typically been limited to 150 bp. In tRNAscan-SE, the first-pass scanners define the approximate bounds of a tRNA, and for tRNAs with very long introns, intervening sequences can be cut out based on the first-pass analysis. This allows detection of rare, abnormally long tRNAs without greatly increasing the overall average search time. In the Genbank subset, tRNAscan-SE detected four tRNAs (HALTGW plus three detected with the -L option) whose introns, ranging from 104 to 850 bp, exceeded the normal length limit for covariance model detection.

2.5.2 tRNA False Positives & Pseudogenes

Of the 5,591 total false positives identified by tRNAscan 1.4 in 15 gigabases of simulated human sequence (Table 2.4), in only six instances did it agree with EufindtRNA (relaxed parameters) in falsely identifying a sequence as a tRNA. The majority of false positives found by tRNAscan 1.4 seem to have tRNA-like secondary structure but lack similarity to conserved tRNA primary sequence. EufindtRNA, on the other hand, identifies correctly spaced primary sequence promoter elements, yet tends to err because it does not check for proper tRNA secondary structure.

These observations hold up on examination of false positives from actual genomic sequence from *C. elegans*. Most of the 29 false positives identified by tRNAscan 1.3 were

discarded by covariance model analysis because of the lack of primary sequence similarity to the general tRNA model. EufindtRNA, on the other hand, more commonly identifies pseudogene tRNA fragments, SINE-like repetitive elements, or other tRNA-like sequences containing A and B boxes (Table 2.3). Pseudogenes are recognizable since part of the sequence is very similar to other intact tRNAs, in spite of truncations or large insertions elsewhere in the pseudogene. However, tRNA secondary structure in pseudogenes and SINE-like elements tends to be lost more quickly than primary sequence promoter elements. This may not be surprising in light of the observation that portions of tRNA sequences are thought to help provide mobility for some tRNA-derived repetitive elements (Keeney et al., 1995). Since EufindtRNA (relaxed parameters) only looks for canonical promoter regions, it is prone to finding these instances of pseudogenes and repetitive elements with tRNA promoters in the absence of structural tRNA features.

To some extent, covariance model analysis is also apt to identify truncated tRNAs and other tRNA-derived sequence elements. The minimum cutoff score of 20 bits has been set to include outlying tRNAs with low overall homology to the general tRNA model. However, if a part of a high-scoring tRNA is truncated, the score may be much lower, but still exceed the 20 bit threshold. The most extreme example of this occurs with a tRNA in the *C. elegans* cosmid W03A3. The tRNA has 100% identity with tRNAs on at least four other cosmids, except for a truncation of the first 16 bases that removes the 5' side of the aminoacyl acceptor stem and the first half of the A box promoter sequence (part of the D-loop). tRNAscan 1.3 did not detect this pseudogene because of the lost base pairings in the D-loop and aminoacyl stems, whereas EufindtRNA could not locate the A box promoter sequence. Covariance model analysis similarly identified three other pseudogenes that neither tRNAscan 1.3 nor EufindtRNA found: one appears to have a 13 bp truncation relative to tRNAs in two other cosmids; one has a peculiar 21 bp insertion in the middle of the A box promoter sequence that makes three near-perfect repeats of the 7-mer "GTCGCGA"; and one cosmid has a pseudo tRNA containing a 55 bp insert in the anticodon loop that does not appear to be a true intron. Since none of these were identified

by either tRNAscan 1.3 or EufindtRNA, tRNAscan-SE necessarily does not detect them.

tRNAscan-SE does, however, detect 19 other tRNA-like sequences that are identified by EufindtRNA and "confirmed" by covariance model analysis (scores greater than 20 bits). These may or may not be pseudogenes. Nine of these involve 5' truncations of 3 to 15 nucleotides relative to other tRNAs in the nematode. It is impossible to determine by computational analysis alone if these are functional tRNAs or inactive pseudogenes. In either case, it is important to be aware of these possible tRNA pseudogenes for possible further experimental and/or computational study. Elucidating a common transpositional mechanism for preferential loss of the 5' end of these tRNAs is a question of interest.

2.5.3 Conclusion

tRNAscan-SE has been designed with the demands of human genome analysis in mind, but can be used for any DNA sequence. We estimate that tRNAscan-SE will detect about 99.5 % of the true tRNAs in the human genome, give zero false positives (except for tRNA-derived SINEs and tRNA pseudogenes), and take approximately 36 hours.

tRNAscan-SE demonstrates that general RNA structural profiles, covariance models, can be used as the basis for very sensitive RNA similarity searching. The primary limitation is speed. Although the strategy of using fast first-pass tRNA scanners in combination with second-stage covariance model analysis is effective here, this is not an attractive general strategy for searching for other RNA gene family members. Except for group I introns (Lisacek et al., 1994), there are no fast, specialized algorithms for detection of other RNA gene families, and much effort is required for creating these highly specialized new programs. Further work will focus on algorithmic development of covariance model search methods that will reduce both time and memory requirements, allowing faster searches for larger RNA genes without the need for first-pass screens.

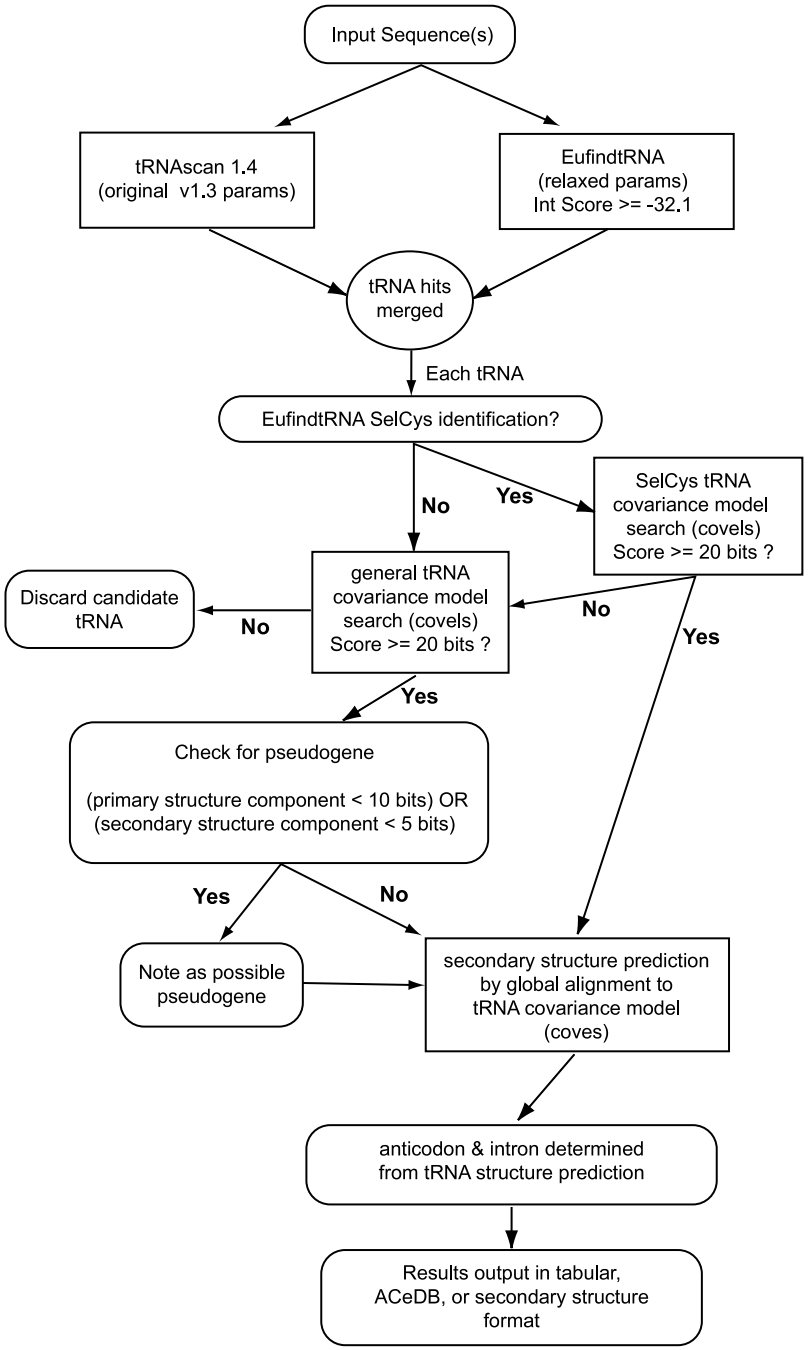

Figure 2.1: **Schematic diagram of tRNAscan-SE algorithm.** Steps carried out by tRNAscan-SE are shown in ovals and rounded-edge boxes. tRNA selection and analysis performed by external programs are shown in rectangles.

Chapter 3

Analysis of the Genomic Complement of *C. elegans* tRNAs

3.1 Introduction

With the completion of the genome sequence of the nematode *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998), we have the opportunity to study the first complete transfer RNA (tRNA) collection from a complex, multicellular eukaryote. Two studies have been published describing the complete collection of tRNAs within the single-celled eukaryote *S. cerevisiae* (Percudani et al., 1997; Hani & Feldmann, 1998). Those analyses confirmed early theoretical predictions on the minimal complement of tRNA genes needed by a eukaryote (Guthrie & Abelson, 1982), as well as giving evidence for the relationship between tRNA species copy number, intracellular tRNA concentration, and protein codon usage.

tRNAs are a critical link in the fidelity of information transfer from messenger RNA (mRNA) to protein sequence. Accurate incorporation of amino acids during translation depends on correct "reading" of the genetic code specified by three-base codons (Crick, 1966). Because there are only 20 amino acids, it is obvious there is excess coding capacity among the 64 possible permutations of the triplet code. Naively, one might expect to find only a subset of the possible codon combinations in use by a particular organism to simplify the cellular machinery needed to translate all possible proteins. In fact, organisms contain all possible codon combinations within (or terminating) their mRNA sequences. Thus, most amino acids are coded for by more than one synonymous codon triplet. If one tRNA were required for each possible codon, this would require the cell to maintain over 60 different tRNA species to be able to translate all possible codons. In fact, many tRNAs specifically recognize more than one codon through non-Watson-Crick base pairings, commonly known as the "wobble hypothesis" (Crick, 1966).

Crick initially proposed the wobble rules based on the observation that codons specifying the same amino acid commonly share the same first two nucleotides. He guessed that the first position of the tRNA anticodon could base pair with more than one possible nucleotide in the third position of mRNA codons. Specifically, a third position anticodon G could

pair with U or C, a U with A or G, and an I (inosine) with U, C, or A (it had been shown that genomically encoded adenosines in the first position of tRNA anticodons are almost universally deaminated to inosine). These simple rules, summarized in Table 3.1, can account for the reduced complement of tRNAs needed for normal translation. In 1982, Guthrie and Abelson (Guthrie & Abelson, 1982) updated and revised the wobble rules, based on observations of characterized yeast tRNAs and their anticodon modifications. They predicted that 46 different tRNA species would be found in yeast (28 were known at the time), and perhaps in all eukaryotes.

Codon	Anticodon	
	Crick (1966)	Guthrie (1982)
U	A, G, or I	G or I
C	G or I	G or I
A	U or I	U*
G	C or U	C

Table 3.1: **Original and Revised Wobble Rules.** "Codon" column indicates third position base in codon. "Anticodon" columns indicate first position base in anticodon. U* indicates modified uridine.

Transfer RNAs are the most extensively modified RNAs studied to date. Some modifications are involved in allowing accurate aminoacylation and/or assumption of the native conformation, but those at position 34 (the first anticodon position) either expand base pairing ability (for example, A to I modification), or restrict pairing ability (various modifications of U). Only one tRNA has been directly RNA sequenced in *C. elegans*, Leu-AAG, which was found to contain an I at position 34 (Tranquilla et al., 1982). Numerous *C. elegans* tRNAs appear in the Sprinzl tRNA database (Steinberg et al., 1993), although they are all derived from DNA sequences, devoid of modification information.

Codon selection has been observed to be non-uniform, depending on organism, genomic location, and transcription level of the coded gene. These are seen as the result of a balance between mutational bias and selection for translationally optimal codons (Sharp

et al., 1993). Highly transcribed genes such as ribosomal proteins and structural proteins tend to have the most non-random, biased codon selection, whereas lowly expressed genes such as regulatory factors tend to have fairly unbiased codon selection. In *Escherichia coli* and *S. cerevisiae*, highly expressed genes bias towards codons decoded by the most abundant tRNA species (Bennetzen & Hall, 1982). Furthermore, a positive correlation has been observed between the cellular abundance of yeast tRNAs and overall codon frequency (Ikemura, 1982). This is likely an instance of co-adaptation in which both codon selection and intracellular tRNA concentration change to reach an optimal balance.

Intracellular tRNA levels are controlled by several possible factors: gene copy number, individual transcription rates, and post-transcriptional regulatory mechanisms. In prokaryotes, pressure for genome compactness appears to severely limit the number of redundant tRNA gene copies (*i.e.*, *Haemophilus influenzae* has 58 tRNAs (Lowe & Eddy, 1997), *E. coli* has 86 (Blattner et al., 1997)). Thus, the latter two factors are the most likely determinants of tRNA concentrations (Dong et al., 1996). In yeast, tRNA copy number varies greatly between 1 and 16 depending on the tRNA species, yielding 274 total genes assorted among 42 unique tRNA classes (Percudani et al., 1997; Hani & Feldmann, 1998). A strong correlation between gene copy number, intracellular tRNA level, and overall codon preference has been observed (Percudani et al., 1997; Hani & Feldmann, 1998). These studies confirmed that tRNA levels are primarily influenced by gene copy number in yeast, and that relative tRNA levels may be predicted based on codon preference within highly expressed genes.

The genome of *C. elegans* is approximately eight times larger than that of *S. cerevisiae*, with 27% versus 72% of the genome coding for exons within worms and yeast, respectively. Thus, *C. elegans* is under less evolutionary pressure to maintain a compact genome and may also use tRNA gene copy as a strategy for regulating intracellular tRNA concentration. In contrast to yeast, however, *C. elegans* is a complex, multicellular eukaryote with many tissue types, no doubt requiring some degree of tissue-specific regulation of tRNA levels. The internal RNA polymerase III promoter sequences for tRNAs, the A and B boxes, do not change between redundant tRNA copies, although upstream and downstream sequences are

not conserved and have been shown to modulate eukaryotic tRNA gene expression (Wilson et al., 1985; Young et al., 1986; Reynolds, 1995). It is unclear to what extent these external enhancer elements are responsible for controlling tRNA concentration. Thus, tRNA copy number may not be predictive of tRNA levels in multicellular eukaryotes like *C. elegans*.

In this study, we analyze the complete complement of tRNAs within *C. elegans* to answer three main questions. First, do the predicted tRNAs fulfill the minimal 46 classes believed to be necessary for translation of all possible codons? Does tRNA copy number correspond to biased codons within the most highly expressed genes, thus implying that gene copy is a major determinant of intracellular tRNA concentration? And finally, we examine many apparent tRNA pseudogenes to find several possible examples of novel SINE elements (Daniels & Deininger, 1985; Deininger, 1989), the first repetitive elements of this class described in *C. elegans*.

3.2 Methods

Transfer RNAs in *C. elegans* were detected using the program tRNAscan-SE with the default (eukaryote-specific) parameters (Lowe & Eddy, 1997). The *C. elegans* genome sequence was retrieved from the Genome Sequencing Center website (http://genome.wustl.edu/-gsc/C_elegans/) on December 11, 1998. Probable pseudogenes were detected automatically by the program, and were visually inspected to classify the type of pseudogene.

Codon frequency counts were derived from Wormpep16, a manually reviewed collection of 16,328 predicted protein coding genes from *C. elegans* (available at ftp://ftp.sanger.ac.uk/-pub/databases/wormpep/wormpep16cdnas.dna). A simple PERL script was written to tally codon usage. Start codons and stop codons were not included in codon usage counts.

3.3 Results and Discussion

tRNAscan-SE predicted 592 tRNAs and 194 pseudogenes within the *C. elegans* genome. After visual inspection, 13 borderline tRNAs (scoring between 20 and 30 bits) were judged to

be likely pseudogenes. Thus, the revised tally is 579 tRNAs and 207 tRNA-like pseudogenes. tRNAscan-SE was also used in the analyses presented in the *C. elegans* genome publication (*C. elegans* Sequencing Consortium, 1998), which gives a different overall tRNA count. We believe that data was incorrectly reported due to simple error or changes in the final version of the sequence.

3.3.1 Intron Occurrence and Genome Distribution

A breakdown of tRNAs by isoacceptor type and anticodon is given in Tables 3.3 and 3.4. Only four *C. elegans* tRNA species contain introns (Ile-UAU, Leu-CAA, Thr-UGU, Tyr-GUA), compared to 10 species in *S. cerevisiae* (Ile-UAU, Leu-CAA, Leu-UAG, Lys-UUU, Phe-GAA, Pro-UGG, Ser-GCU, Ser-CGA, Trp-CCA, Tyr-GUA). The total percentage of intron-containing tRNA genes in *C. elegans* is 5%, whereas 21% of *S. cerevisiae* tRNA genes contain introns. The role of introns within tRNAs has not been widely studied, although for several yeast tRNAs, an intron is necessary for correct addition of anticodon base modifications (Johnson & Abelson, 1983; Strobel & Abelson, 1986). Clearly, the evolutionary pressures to increase or reduce introns within tRNAs is independent of those influencing the occurrence of mRNA introns, which occur at a low frequency in yeast relative to *C. elegans*.

tRNA species copy number in *C. elegans* ranged from 1 (Arg-CCG, Selenocysteine (SeC)-UCA) to 33 (Pro-UGG). tRNAs were somewhat evenly distributed among the chromosomes except for a disproportionate number on the X chromosome, which contained nearly half of all tRNAs in the genome (see Table 3.2).

3.3.2 *C. elegans* Follows Wobble Predictions

Based on the revised wobble rules specific to eukaryotes (Guthrie & Abelson, 1982) (Table 3.1), I anticipated finding 46 distinct tRNA species in the *C. elegans* collection. Indeed, we found precisely 46 families, plus a single selenocysteine-inserting (SeC) tRNA. Interestingly, the 46-family prediction was based on an incomplete set of tRNAs from *S. cerevisiae*, which

Chromosome	Size (Mbp)	tRNA Count	Density (tRNAs/Mbp)
I	14.0	61	4.4
II	14.9	55	3.7
III	13.0	57	4.4
IV	16.8	57	3.4
V	21.2	76	3.6
X	17.4	273	15.7

Table 3.2: **Genome Distribution of tRNAs.**

turned out to have only 42 tRNA species (Percudani et al., 1997; Hani & Feldmann, 1998). The missing four families were accounted for by known differences in anticodon modifications in existing tRNAs which could expand base pairing ability and compensate for the absent species (Percudani et al., 1997; Hani & Feldmann, 1998). Almost no direct modification data for *C. elegans* tRNA anticodons exist, but based on the full complement of tRNAs found, *C. elegans* tRNAs are expected to be very typical of eukaryotic tRNAs.

The most comprehensive collection of tRNA genes, the Sprinzl tRNA database (Steinberg et al., 1993), had formerly identified 42 total *C. elegans* tRNAs falling into 35 families. Thus, 12 new families were identified based on the completed genome sequence.

3.3.3 tRNA Gene Redundancy and Codon Frequency

We found a positive correlation between tRNA gene copy number and overall codon usage. For ten out of twelve amino acid isotypes which are decoded by more than one tRNA, tRNA rank order based on copy number was the same as rank based on frequency of associated codons (see Tables 3.3 and 3.4; when tRNAs were predicted to decode more than one codon, codon frequencies were summed for rankings). Glutamate and lysine, both 2-box isotypes, were the two exceptions.

tRNAs may also be ranked relative to all other tRNAs based on copy number and codon frequency. Figure 3.1 shows a consistent trend of increasing tRNA gene number with increasing codon usage (see Table 3.5 for labelled rankings). The general pattern is

remarkably similar to a plot of the same type for *S. cerevisiae* tRNAs (Percudani et al., 1997). The degree of similarity is somewhat surprising in view of the fact that *S. cerevisiae* is a single-celled eukaryote under very different metabolic stresses relative to a metazoan.

No previous studies report general intracellular tRNA levels in *C. elegans*, so we were not able to correlate tRNA redundancy to cellular tRNA concentrations directly. However, codon usage within highly expressed genes has been shown to correlate with tRNA concentration in *S. cerevisiae* (Ikemura, 1982). In general, highly expressed genes are optimized for efficient translation, thus codon selection favors the most readily available tRNAs. By comparing tRNA gene copy number to the codons that are favored in highly expressed genes, we can infer whether intracellular tRNA concentration is influenced by gene copy number.

Tables 3.3 and 3.4 include a column, "Preferred Codon for Highly Expr. Genes", which indicates codons that are statistically preferred in a set of characterized, highly expressed *C. elegans* genes (Stenico et al., 1994). For the 11 of 12 amino acid isotypes which are coded for by more than one tRNA, the most redundant tRNA species decodes the codons most preferred by highly expressed proteins. Glutamine was the single exception to the rule. We infer from this relationship that tRNA gene copy number is likely to be a major determinant of intracellular tRNA concentration levels.

3.3.4 tRNA Pseudogenes

We grouped the 207 tRNA-like pseudogenes identified by tRNAscan-SE into four classes: I) end-truncated tRNAs, II) insertion-disrupted tRNAs, III) "non-maintained" tRNAs, and IV) non-tRNA, polIII-like elements. Members of classes I and II show almost identical sequence similarity to legitimate tRNAs, with the exception of one contiguous insertion or deletion. About 16 tRNA families contained between one and three pseudogenes in these classes, most likely the result of recent and limited mutational events. Some of these predicted pseudogenes may still be functional – we cannot be certain without experimental characterization.

Class III pseudogenes appear to be derived from tRNAs, but show interspersed point mutations and single nucleotide insertions or deletions when compared to known tRNAs or other class III pseudogenes. Some of the multi-member groups of this type may be tRNA-derived SINE repetitive elements (Daniels & Deininger, 1985; Deininger, 1989). Members of this class could almost be mistaken for legitimate tRNAs, except for the fact that they have reduced tRNA-like secondary structure, and the mechanism used by tRNA families to maintain sequence homogeneity appears to be absent. Lack of such a mechanism has allowed group members to drift by mutation independently, without evidence of structural conservation or covariation.

Figures 3.2 and 3.3 show two example alignments of class III elements. The first group of 12 pseudogenes have TTG in their anticodon positions, although comparison with a true Gln-TTG tRNA shows poor similarity (first sequence in Figure 3.2). Comparison to all other true tRNAs in *C. elegans* does not give any significant hits. It is formally possible this represents a tRNA family, but the lack of tRNA-like secondary structure and frequent point mutations relative to other members argues these are not functional molecules. Figure 3.3 shows another example of this type of element. The sequence in the anticodon position for this group is ATG (His). Interestingly, we found no legitimate tRNAs with this anticodon. Instead, the His-GUG is expected to read CAU codons, as per standard wobble rule (Crick, 1966). As with the TTG pseudogenes, this group possesses weak tRNA secondary structure, as well as frequent base substitutions relative to other members in the group. We found at least three other large groups like these in our analyses.

Class IV pseudogenes contain strong RNA polymerase III internal promoter elements (A and B boxes), but show no recognizable similarity to other tRNAs or tRNA-like pseudogenes. Some members of class IV may in fact be unidentified pol-III transcribed RNA genes. Alternatively, these may be nonfunctional tRNAs that have drifted beyond obvious sequence similarity with their original families.

3.3.5 A tRNA-derived SINE

In our study of *C. elegans* tRNA pseudogenes, we discovered a large family of related elements that we believe are likely retrotransposons, dubbed "tde-1" for tRNA-Derived repetitive Element-1. An alignment of 52 members of this family is shown in Figures 3.4 and 3.5. We estimate there are over 200 copies of tde-1 in the *C. elegans* genome. Tde-1 resembles SINE elements found in other metazoans based on its high copy number, apparent derivation from a non-coding RNA (tRNA in this case), and lack of secondary structure conservation. Like other SINES, this element appears to have included flanking sequence, outside the ancestor RNA gene, as part of the mobile element. The 5' 74 nucleotides of tde-1 contain strong primary sequence similarity to tRNA pol-III promoters, but poor tRNA-like secondary structure. 3' to the tRNA-like sequence is an additional 175 nucleotides present in over 100 family members. Careful examination of the tde-1 alignment shows fairly random substitutions throughout the sequence, leading one to believe there is little or no selection for biological function.

3.4 Conclusions

In conclusion, our study of the complete *C. elegans* tRNA family has produced several new observations. First, the Guthrie revised wobble rules (Guthrie & Abelson, 1982) appear to apply well to *C. elegans*. In the absence of biochemical data on anticodon modifications, we are now able to infer *C. elegans* tRNAs contain modifications that are typical within Eukarya based on family representation. Second, tRNA genome copy number correlates well with codon usage. Based on highly-expressed genes' codon preference for the most redundant tRNA species, we also infer that tRNA copy number is a major determinant of intracellular tRNA concentration. Finally, there appears to be a great diversity of tRNA-like pseudogenes within the *C. elegans* genome. We identify and partially classify over 200 of these elements, in contrast to the single tRNA pseudogene found in the *S. cerevisiae* genome (Lowe & Eddy, 1997). We also present what we believe is the first example of

a retrotransposon SINE repetitive element in *C. elegans*. The study of pseudogenes and repetitive elements in metazoans will no doubt be a rich area of genome research in the future, as they are molecular fossils that may give new clues regarding genome dynamics and evolution. The opportunity to study complete gene families, including pseudogenes, is one of the unique benefits yielded by the current "genome rush".

Isotype	Codon	tRNA	tDNA Anticodon	tRNA Anticodon	Genomic Copies	Codon Frequency	Pref by ⇑ Expr. Genes	Notes
Ala	GCU	Ala-1	AGC	IGC	22	22.4	•	
	GCC					11.9	•	
	GCA	Ala-2	TGC	UGC	8	20.1		
	GCG	Ala-3	CGC	CGC	4	7.8		
Gly	GGA	Gly-1	TCC	UCC	31 + 1p	31.4	•	
	GGC	Gly-2	GCC	GCC	13	6.4		
	GGU					11.0		
	GGG	Gly-3	CCC	CCC	3	4.4		
Pro	CCA	Pro-1	TGA	UGA	32 + 3p	25.9	•	
	CCU	Pro-2	AGG	IGG	6	9.1		
	CCC					4.4		
	CCG	Pro-3	CGG	CGG	4	9.0		
Thr	ACU	Thr-1	AGT	IGU	17	19.5		
	ACC					10.3	•	
	ACA	Thr-2	TGT	UGU	12	20.3		3 w/intr
	ACG	Thr-3	CGT	CGU	7 + 1p	8.5		
Val	GUU	Val-1	AAC	IAC	18	24.8		
	GUC					13.2	•	
	GUA	Val-2	TAC	UAC	5	10.3		
	GUG	Val-3	CAC	CAC	5	14.1		

Table 3.3: **Four-box tRNA Families in *C. elegans*.** tRNAs were named and numbered based on predicted isotype and frequency rank within genome. "Codon" entries are grouped with the major tRNA-decoding species based on standard "wobble" rules (Crick, 1966). "tDNA Anticodon" inferred from tRNAscan-SE (Lowe & Eddy, 1997) analysis. Experimental tRNA anticodon modification data is available only for Leu-1 (Tranquilla et al., 1982); the only modifications assumed for "tRNA Anticodon"s are the common first-position adenosine to inosine (I) conversions. Pseudogenes recognizably derived from "legitimate" tRNA species are included in "Genomic Copies" as "p" counts. "Codon frequency" is the number of codons per thousand total codons.

Isotype	Codon	tRNA	tDNA Anticodon	tRNA Anticodon	Genomic Copies	Codon Frequency	Pref by ⇑ Expr. Genes	Notes
Arg	CGU	Arg-1	ACG	ICG	19 + 2p	11.0	•	
	CGC					4.9	•	
	CGA	Arg-2	TCG	UCG	10	11.6		
	CGG	Arg-3	CCG	CCG	1	4.4		
	AGA	Arg-4	TCT	UCU	7 + 1p	15.6		
	AGG	Arg-5	CCT	CCU	4	3.8		
Ser	AGC	Ser-4	GCT	GCU	8 + 1p	8.1		
	AGU					12.3		
	UCU	Ser-1	AGA	IAC	15	17.3	•	
	UCC					10.5	•	
	UCA	Ser-2	TGA	UGA	7	20.7		
	UCG	Ser-3	CGA	CGA	6	11.5		
Leu	CUU	Leu-1	AAG	IAG	19 + 2p	21.6	•	
	CUC					14.5	•	
	CUG	Leu-2	CAG	CAG	6	11.8		
	CUA	Leu-3	TAG	UAG	3	8.1		
	UUG	Leu-4	CAA	CAA	7	20.4		7 w/intr
	UUA	Leu-5	TAA	UAA	4	10.5		
Phe	UUC	Phe-1	GAA	GAA	13 + 1p	24.4	•	
	UUU					25.3		
Asp	GAC	Asp-1	GTC	GUC	27 + 2p	16.7	•	
	GAU					35.6		
Glu	GAG	Glu-1	CTC	CUC	23	23.2	•	
	GAA	Glu-2	TTC	UUC	17 + 3p	40.9		
His	CAC	His-1	GTG	GUG	18 + 10p	9.0	•	
	CAU					14.2		
Gln	CAA	Gln-1	TTG	UUG	18 + 13p	27.2		
	CAG	Gln-2	CTG	CUG	6 + 1p	13.6	•	
Asn	AAC	Asn-1	GTT	GUU	20 + 1p	18.7	•	
	AAU					31.0		
Lys	AAG	Lys-1	CTT	CUU	31	25.5	•	
	AAA	Lys-2	TTT	UUU	15 + 1p	38.9		
Met	AUG	Met-i	CAT	CAU_i	8 + 1p	–		
	AUG	Met-1	CAT	CAU	9	23.9		
Ile	AUU	Ile-1	AAT	AAU	21 + 1p	33.4		
	AUC					18.9	•	
	AUA	Ile-2	TAT	UAU	7 + 1p	10.1		3 w/intr
Cys	UGC	Cys-1	GCA	GCA	13	9.1	•	
	UGU					11.6		
Trp	UGG	Trp-1	CCA	CCA	10	11.1		
SeC	UGA	SeC	TCA	UCA	1	0.0		
Tyr	UAC	Tyr-1	GTA	GUA	19	14.0	•	19 w/intr
	UAU					18.2		
Sup	UAG	None			0	0.0		
	UAA	None			0	0.0		

Table 3.4: **Non four-box tRNA Families in *C. elegans*.** See Table 3.3 for column headings.

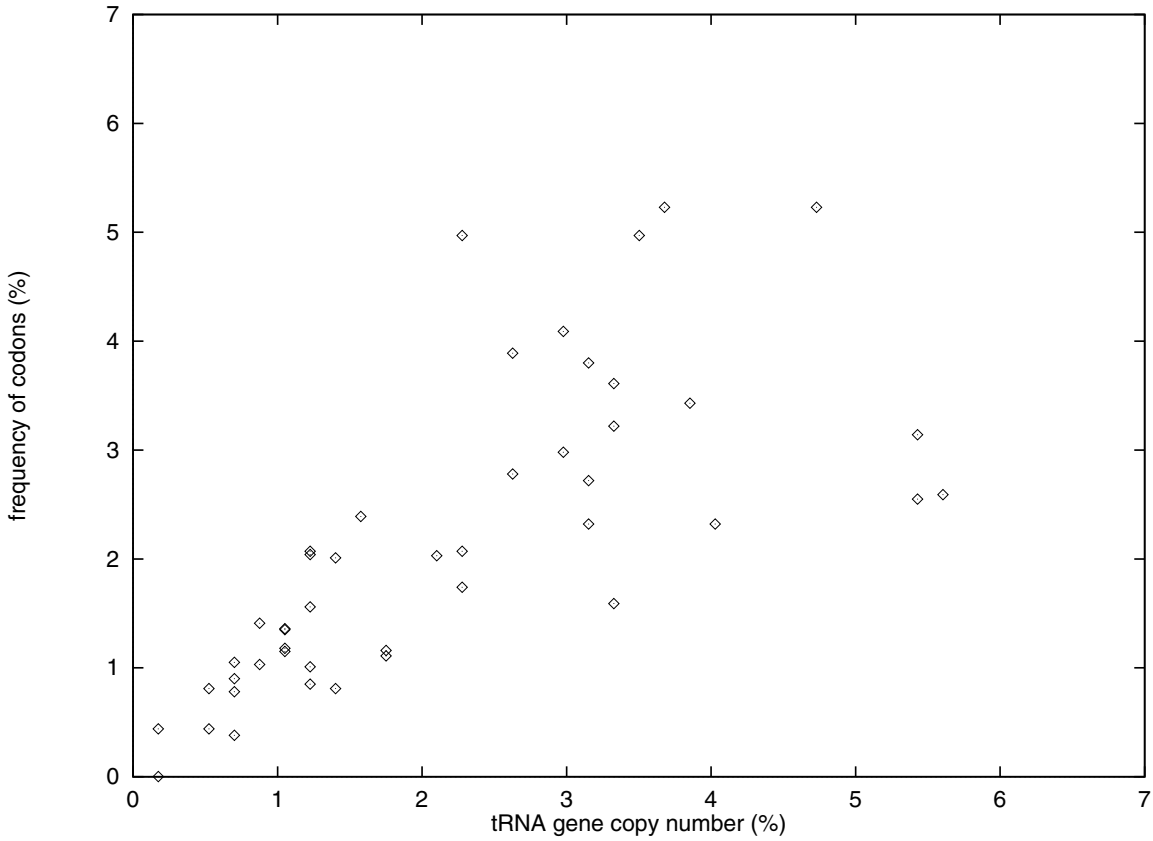

Figure 3.1: **tRNA gene copy number versus codon frequency.**
tRNA copy number is the count of each tRNA divided by the 571 total tRNAs in the genome (pseudogenes not included). Frequency of codons is the frequency of all codons expected to be decoded by a given tRNA. Initiator methionine and start codons are not included in counts. See Table 3.5 for anticodon labels to data points.

Anticodon	tRNA Gene Copy Number (%)	Frequency of Codons (%)
UGA	5.604	2.590
UCC	5.429	3.140
CUU		2.550
GUC	4.729	5.230
CUC	4.028	2.320
IGC	3.853	3.430
AAU	3.678	5.230
GUU	3.503	4.970
ICG	3.327	1.590
IAG		3.610
GUA		3.220
UUG	3.152	2.720
IAC		3.800
GUG		2.320
UUC	2.977	4.090
IGU		2.980
UUU	2.627	3.890
IAC		2.780
GCC	2.277	1.740
GCA		2.070
GAA		4.970
UGU	2.102	2.030
UCG	1.751	1.160
CCA		1.110
CAU	1.576	2.390
UGC	1.401	2.010
GCU		0.810
UGA	1.226	2.070
UCU		1.560
UAU		1.010
CGU		0.850
CAA		2.040
IGG	1.051	1.350
CUG		1.360
CGA		1.150
CAG		1.180
UAC	0.876	1.030
CAC		1.410
UAA	0.701	1.050
CGG		0.900
CGC		0.780
CCU		0.380
UAG	0.525	0.810
CCC		0.440
CCG	0.175	0.440
UCA		0.001

Table 3.5: **tRNA Gene Copy Number Versus Decoded Codon Frequencies.**

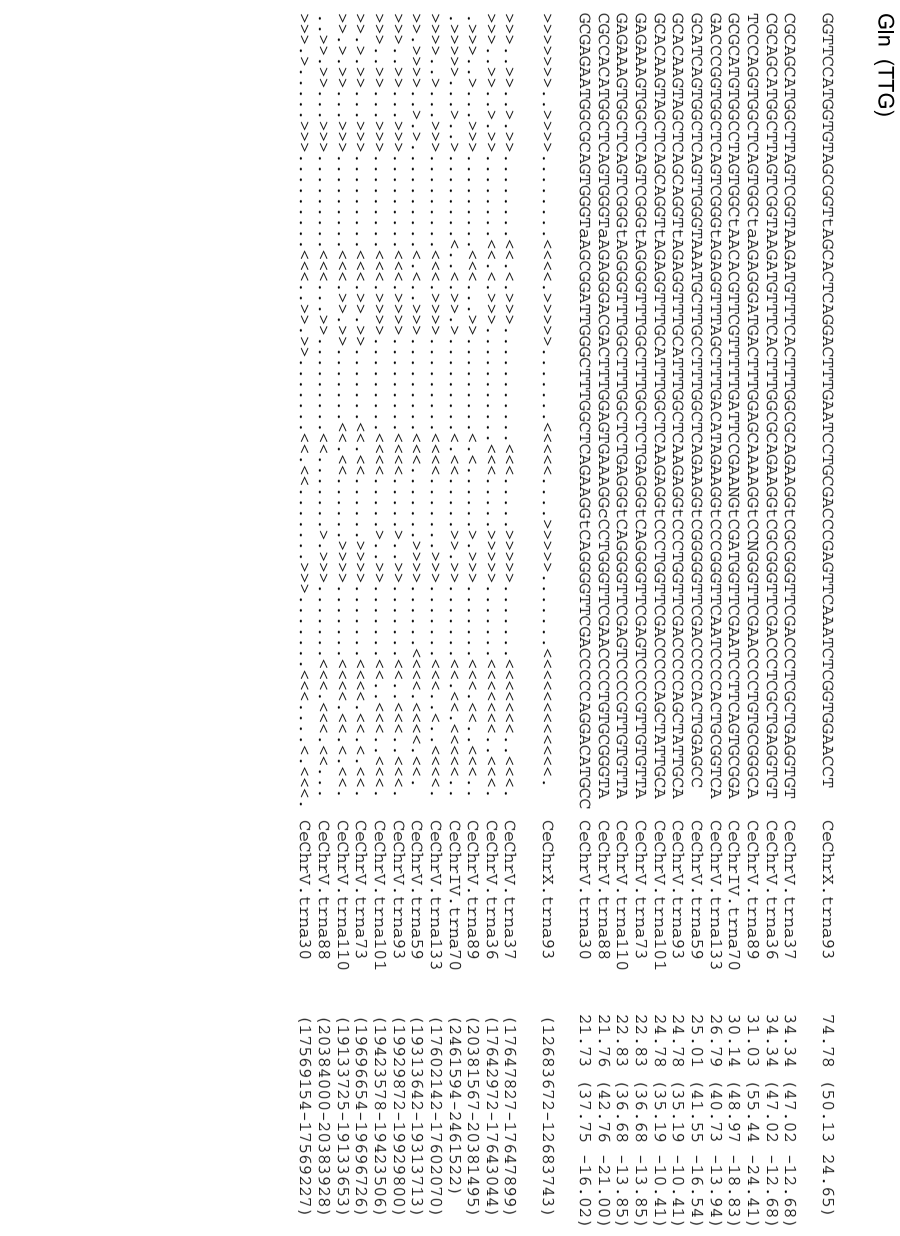

Figure 3.2: **tRNA-like pseudogenes with TTG (Gln) "anticodons".**
The first sequence is a "legitimate" Gln-TTG tRNA, followed by an alignment of 12 tRNA pseudogenes with TTG in their anticodon positions. Below sequence alignments are corresponding secondary structure predictions and bounds from tRNAscan-SE. Nested ">" and "<" denote base pairings. Right three columns of scores indicate: a) overall tRNA score, b) primary sequence score, c) secondary structure score (in bits). Note loss of pairing potential in pseudogenes, and low secondary structure scores relative to true Gln-TTG tRNA.

His Pseudo (ATG)

```
GAACTTTTGGCTCAACTGGTaAGAGGGATACCAATGGAGAAATAGGtCATGGGTTCGATTCCCAACGAGGCCA  CeChrII.trna6     20.50 (22.61 -2.11)
GGCGTCGTGGCTCAGTCGGTAAGCGgTTGGGCTATGAAGCAAAAGGtCCCGGGTTCGGATCCCCGGAAGCTT   CeChrII.trna36    22.64 (36.80 -14.16)
GCGCGCGTGGCGCAGTCGGTAAGATGTCAGACTATGACCCAGAAGGtCACAGGTTCGATTCTTtGTCGAAGGAT CeChrIII.trna6    21.18 (44.33 -23.15)
GTCGTAGTGGCTCAGCTGGGtAGAGGTTGGCCTATGGCTCTGAAGGtCCGGGGTTCAATCCCCACAGCGAAA   CeChrIV.trna77    23.70 (34.25 -10.55)
GTCTTCGTGGCTCAGTTGGGaAGAGGTTTGACTATGGTTCAGAAGGtCAGGGGTTCGACCCCATCGGTGGCT   CeChrV.trna32     31.86 (41.03 -9.17)
GACCCGCTGGCTCAACTGGGtAGAGGTTTGGCTATGGCCCAGAGGGtCAAGGGTTCGACCCCCGCCGAGGCA   CeChrV.trna34     24.73 (27.72 -2.99)
GTTCCGGTGGCTCAGCTGGGtAGAGAGATGGCTATGGGGCAAAT GtCCGGGGTTCGATTCCCCGCTAGAGTT  CeChrV.trna39     23.71 (22.98 0.73)
GTCGAAGTgGCTCAGGTGGGAAGAGGGATAGCTATGGGGAATAGGGtCCTGGGTTCGAGTCCCCGCTTGGGCA  CeChrV.trna40     20.38 (25.94 -5.56)
GCACCGGTGGCTCAGCTGGGtAGAGGGACGGCTATGGCTCAGGAGGtCAGGGGTTCGACTCCCTGCTGTGGCT  CeChrV.trna52     29.70 (41.49 -11.79)
GCGCTGGTGGTTCAGCTGGGtAGAGGGGCGGCTATGGCTCAGTAGGtCTGGGGTTCGAACCCCAAGCGTGGCC  CeChrV.trna53     22.42 (32.97 -10.55)
  GCCAGTAGCTCAGTTGGGtAGAGGCTTGGCTATGACTCTGGAGGtCAAGGGTTCGAGCCCCGGCGTGATA   CeChrV.trna54     22.39 (28.96 -6.57)
GCACCTGTGGCTCAGTCGGGtAGAGGAATGTCTATGGCTCCGGAGGtCAGGGGTTCGAACCCCAACCGTGGCC  CeChrV.trna55     21.60 (38.43 -16.83)
GCGCCgATGGCTCAGGTGGGtAGAGGTTTGGCTATGACGCAGAAGGtCACGGGTTCGACCCCCGGTGGGGTAA  CeChrV.trna57     35.22 (37.06 -1.84)
GTCGCTGTGGCTCAGTTGGGtAGAGGTTTGGCTATGGCGCAGAAGGtC GGGGTTCGATTCCCACTGGGGTA   CeChrV.trna58     31.93 (38.61 -6.68)
GCACAAGTAGCTCAGTTGGGtAGAGGCTTGGCTATGGCTCTGGAGGtCGAAGGTTCAATCCCCGCCGTGGTA  CeChrV.trna60     24.15 (37.08 -12.93)
GTCCCGGTGGCTCAGCTGGGtAGAGGCTTGGCAATGGTCCAGAAGGtCAAGGGTTCGAGCCCCAGTGGTGGCA  CeChrV.trna62     20.30 (39.19 -18.89)
GTCGCAGTGGCTCAGTAGGGtAGAGGTTTGGCTATGGCGCAGAAGGtCGAGGGTTCGAACCCCCACTGTGGTA  CeChrV.trna63     35.77 (39.55 -3.78)
GCACCTGTGGCTCAGTCGGGtAGAGGAATGTCTATGGCTCCGGAGGtCNNGGGTTCGAACCCCAACCGTGGCC  CeChrV.trna72     22.60 (42.85 -20.25)
GCCTCGGTGGCTCAGTTGGGtANAGGTTTGGCTATGGNCNAGAAGGtCAAGAGTTCGAACCCTGCAGAGGCC  CeChrV.trna74     41.74 (37.82 3.92)
GCCNGGGTGGCTTAGTGGCtaAGAGTGATGCCGATGGAGTGGTAGGtCAGGGGTTCGAGCCCCCGAATGGGTG  CeChrV.trna90     22.21 (33.23 -11.02)
GCCCGGGTGGCTTTGTGGCgaAGAGTGATGCCTATGGAGTGGTAGGtCAGGGGTTCGAGCCCCCGCATGGGTG  CeChrV.trna91     27.44 (32.08 -4.64)
GCGCCGATGGCTCAGGTGGGtAGAGGTTTGGCTATGACGCAGAAGNtCACGGGTTCGNNNCCCGGTGGGGTA  CeChrV.trna94     26.36 (31.38 -5.02)
GNCGCTGTGGCTCAGTTGGGtAGAGGTTTGGCTATGGCGCAGAAGGtC GGGGTTCGATTNCCACTGGGGTA  CeChrV.trna95     29.73 (39.41 -9.68)
GCACAAGTAGCTCAGTTGGGtAGAGGCTTGGCTATGGCTCTGGAGGtCGAAGGTTCAATCCCCGCCNTGGTA  CeChrV.trna97     24.15 (37.08 -12.93)
GCACCGNTGGCTCAGCTGGGtAGAGGGACGGCTATGGCTCANGAGGtCAGGGGTTCGACTNCCCTGCTGTGGC  CeChrV.trna99     24.34 (25.24 -0.90)
GACCAAGTGGCGCAGTCGGTAACAGGGATGGCTATGACCCAGAAGGCCAcGGGTTCGACCCCCcGCTAGaGTCA CeChrV.trna102    22.90 (36.85 -13.95)
GCCTCGGTGGCTCAGTTGGGtAGAGGTTTGGCTATGGCCCAGAAGGtCAAGAGTTCGAACCCTGCAGAGGCC  CeChrV.trna104    40.32 (40.06 0.26)
  GCCAGTGGCTCAGTCGGGtAGGGCTTTGGCTATGGCTCTGAGGGtCAGGAGTTCGAGTCCCCGTTGTGGCA  CeChrV.trna109    23.67 (33.68 -10.01)
GCTGCGGTAGTTCAGTTGGGtAGAAGTTTGACTATGGCTCCAAGGGtCACAAGTTCAATCCCCGCTTGACCCA  CeChrV.trna129    20.15 (41.44 -21.29)
TCAGTAGTGGCGCAGTGGCtAACAGATCTGGCTATGACGCAGAAGGCCAgAGGTTCGATCCCcGCCACTGTT   CeChrV.trna136    23.10 (31.01 -7.91)


>..>>>....>>>.........<<<..>>.>.......<..<<......>>>>.......<<<<..<<<..<.  CeChrII.trna6     (1434127-1434199)
>>>.>>...>>>.........<<<..>>.>.........<.<<......>.>>>.......<<<.<<<.<<<.   CeChrII.trna36    (12273825-12273896)
.>...>>..>.............<.>>.>.........<.<<.....>>>>>.......<<<.<<<<...<<..  CeChrIII.trna6    (2314474-2314547)
.>>>>.>...>>>.........<<<.>>>>.........<<<<......>>>>......<<<<.<.<<<<..   CeChrIV.trna77    (1105319-1105248)
>>>.>>>...>>>.........<<<.>>>>>.......<<<<<.....>.>>>......<<<.<<<<.<<<.   CeChrV.trna32     (17594652-17594723)
>.>.>>....>>>.........<<<.>>>>>.......<<<<<.....>.>>>.....<<<..<<.<.<.    CeChrV.trna34     (17598779-17598850)
>>>>.>>...>>>.........<<<...>>.........<<.....>>>>>.......<<<<<<<.<<<<.   CeChrV.trna39     (18196002-18196073)
>>>.>>>...>>>.........<<<..>>..........<<......>.>>>.......<<<.<<<<.<<<.   CeChrV.trna40     (18200092-18200164)
>>..>>>...>>>.........<<<...>>>........<<.<.....>>>>.......<<<<.<<<..<<.   CeChrV.trna52     (19167955-19168027)
>>..>>....>>>.........<<<..>>>>.......<<.<<.....>.>>>.....<<<.<.<<..<<.   CeChrV.trna53     (19168249-19168321)
  >.>>>...>>>.........<<<.>>.>>........<<.<<.....>>>.....<<<..<.<<.<.    CeChrV.trna54     (19170166-19170235)
>>..>.>...>>>.........<<<...>>.........<.<.....>>>.....<<<..<.<..<<.   CeChrV.trna55     (19208239-19208311)
>>>.>>.>...>>>.........<<<.>>>>.........<<<<.....>>>>.....<<<<.<<<.<<<.  CeChrV.trna57     (19312558-19312631)
 >>>.>.>...>>>.........<<<.>>>>.........<<<<.....>>>.....<<<.<.<.<<<.   CeChrV.trna58     (19313343-19313413)
>>.>>.>...>>>.........<<<.>>.>>........<<.<<.....>..>>.....<<..<<.<<<<.   CeChrV.trna60     (19314106-19314177)
```

Figure 3.3: **tRNA-like pseudogenes with ATG (His) "anticodons".** Alignment of 30 tRNA pseudogenes with ATG in their anticodon positions. Below sequence alignments are corresponding secondary structure predictions for 15 pseudogenes from tRNAscan-SE. Nested ">" and "<" denote base pairings. No "legitimate" tRNAs with ATG anticodons were found in the *C. elegans* genome.

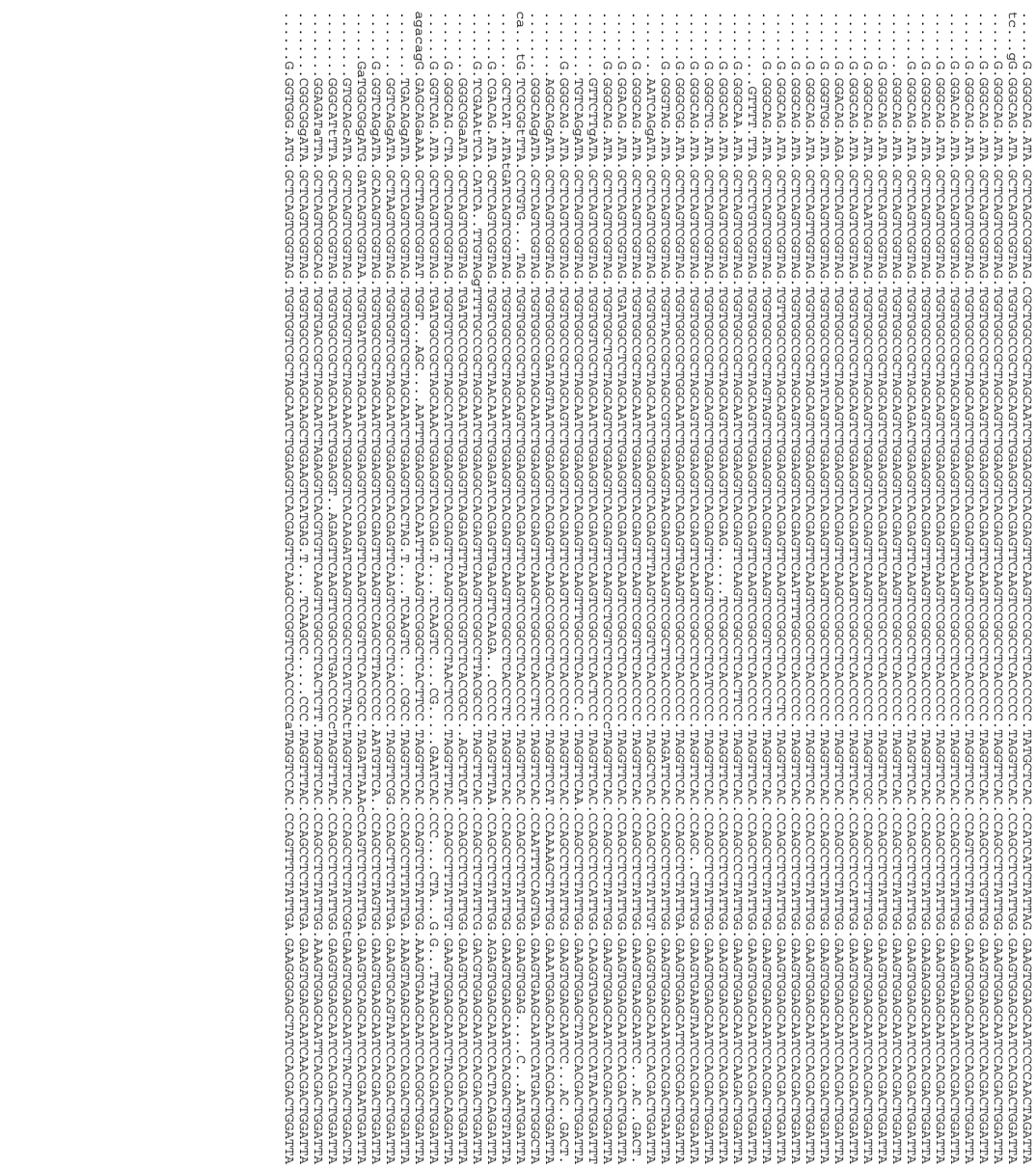

Figure 3.4: ***C. elegans* tRNA-derived SINE element alignment (5' half).** Alignment of 52 out of >200 genomic copies of tRNA-derived SINE-like element. The first 74 nucleotides of these elements were detected by tRNAscan-SE as tRNA-like with strong pol-III promoters, but poor tRNA secondary structure.

```
C02D4.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCTCAGTGG.GAGCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTACATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
C18B2.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
C01G5.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
K02B9.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTC.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTCAATCC
F17A2.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
C24A3.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
R11F4.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAAACCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
T05C1.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGATTCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CT.T.....
C53D6.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
F56F10.t1       TCGGCCACAGTCCCCGGCTAGGACGTGACTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
T21C12.t1       TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTAAAGAACGGATCGTC....CTTTAATCC
F39D8.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
M106.t1         TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCTTC....CTTTAATCC
D1053.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCATTTGAAGAACGGATCGTC....CTTTAATCC
F38B2.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAAT.C
C05E7.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAATGGATCGTC....CTTTAA...
C05D9.t1        ACGGCCACAGCCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCTCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
F46G10.t1       TCGGCCACAGTTCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACTACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCATC....CTTCAATCC
F45E6.t1        TCGGCCACAGTCCCCGGCTAGGTCGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGCC....CTTTAATCC
T01E8.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
C09G1.t1        TCGGCCACAGTCCCCGGCTCGGACGTGGCTT.AAATTA.CAACCCAGTGG.GATCACCACCAGGCAGTGTACCTGAGTCCCAGATCCGCAGTGCATAGCACTTGGAGAACGGATCGTC....CTTTGACCC
T24D8.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
E03A3.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTTaAAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
F19H6.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAGCGGATCGTC....CTTTAATCT
F55G7.t1        TCGGCCACAGTCCCCGGCTAGAACGTGGCTT.AAATTA.TAGCCCAGTGG.GATCACCACCAGGCAGTGTACCAGAATCCCAGATCCGCAGTGCATAGCACTTGAAGATCAGATCGTC....CTTCAAT..
C52B9.t1        TCGGCCAGAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CCTTAATCG
F46F2.t1        TCGGCCACAGTCTCCGGCTAGGCCGTGGTTT.AAATTA.TAGCCCAGTGG.AATCACCACCAGGCAGTGTACCTGAATTCCAGATCCGCAGTGCATAGCGTTTGAAGAACGGATCGTC....ATTTAATCC
T24D11.t1       .GGACCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.TAGCCCAGTGG.GAGCTTCACCGGGCAGCGTACCTGACTCCCAGATCCGCAGTGCATAGCACTTGGAGAACGGATCGTC....CTTTAATCC
T14B1.t1        TCGGCCACAGTCCCCGGCTAGGATGTTGCTT.AAATTA.TAACCCAGTGG.GATCACCACCAGGCAGTGTACCTGACTCCCAGATCCG.TGTGCATAGCACTTGAAGAACGAATCGTC....CTTTAATTC
C04E7.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGGtGATCACCACCAGGCAGTGTACCTGAACCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCT
F31F6.t1        TCCGCCACAGTCTCCGGCTAGGACGTGGCTT.AAATTA.TAGCCCAGTAG.GAGCACCATCAGGCAGTGTACCTGGCTCCTAGATCCGCAGTGCATCGCACTTGAAGAAGGGATCGTC....TTTTAATCC
C49F8.t1        TCGGCCACAGTCCCCGACTAGGACGTGTCTT.AAATTA.TATCCCAGTAG.GATCACCACCAGGCAGTGTACCTGACTCCCGGATCCGCGGTGCATAGCACTTGAAAAACGGATCGTC....CTTTAATCT
M03C11.t1       .CGGCCACAGTCCCCGGATAGGACGTGGCTT.AAATTAaAAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCC
C11G10.t1       TGGACCACAGTCCCTGGCTAGGACGTGGCTT.AAACTA.AAGCCCAGCGG.GAGCACCACCAGGCAGGGTACCTGACTCCCAGATCCGCAGTGCAGAGCGCTTGTAGAAGGGATCGTC....CTTAAATCC
F33C8.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.TAGCCCAGTGG.GATCACCATCAGGCAGCGTACCTGAATTCCAGATCAGCAGTGAATATCACTCGTAAAACGGATCGTC....CTGTAATCC
R11G10.t1       TCGGCCACA.TCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCA.TGCATAGCACTTGAAGAACGGATCGTC....CTTTAATCA
T24B8.t1        TCGGCCACAGTCCCCGGCTAGGACGTGGCTT.AAATTA.CAGCCCAGAGG.GATCACCACCAGGCAGTGTACCTGAATCCCAGATCCGCAGTGCATA......GA...TCG.....TC....CTTTAATCC
T04C10.t1       TCGGCCACAGTCCCAGGCTAGGATGTGGCTT.AAATTA.TAGCCCAGTGA.GATCACCACCAGTCAGTTTACCTGAATCCCAGTTCCGCAGAGCATAGCACTTGAAGAACGGATCGTC....TTTTAATCC
T02C5.t1        TTGTCCGCAGTCCCCGGCTAGGACGTGGCTT.AAATTA.TAGCCCGGTGG.GAACACCACCAGGCAGTGCACCTAACTCCCAGATCCAAAGTGCATAGCGCTTGTAGATCGGATTGTC....CTTTAATCA
T24C2.t1        TCGGCCACAGTCCCCGGCTTGGATGTGGCTT.AAATTA.TAGCCCAGTTG.GAGGGTCACCAGGCAGTGGGCCTGACTCTCAGATCTGCAGTGCATAGCGCTTGAAGAACGGATCGTC....CTTTAATCC
C27C12.t1       TCGGCCACAGTCCCCGGCCCGGACGTGGCCT.TATCTA.TATCCCAGTGG.GATCACTACCAAGCAGTGTATCTGACTCCCAGATCCGTAGTGCATAGCGCTTGA.AAACGGATCGTC....CTTTAATCC
T14C1.t1        TCGGCCACCGTCCCCGGCTAGGACGTGGCTT.AAATTA.TATCGCAGTGG.GAGCACTACCAGGCAATGCACCTGACTCCCAGATCCGCAATGCACAGCGCTTGAAGATCGGATCGTC....TTGTAATCT
C05G5.t1        TCGGCCACAGTCTCCGGCTAGGACGTGGCTT.AAATTA.TAGCCCAGTGG.GAGCACAACCAGGCAGTGTACCTAACTCCCAGATCCGCAGTGCGTTGCGCTTGAAGAACGAATCGCC....CTTTAATCT
T01B4.t1        TCGGTCACAGTCCCCGGCTAGGACGTGGCTA.AAAGTA.TTGCCCAGTGG.GAGCACCACTAGGCGGTGTACCTGACTCCCAGAGCCGCAGTGCATAGCGTTTGAGGAACAGATCGTT....CTTTAATCC
T14F9.t1        TCGGCCACGGTTCCTGGCTAGGACGTGGCTC.AAATTA.TAGCCCAGTGG.GAGCACCACCAGGCAGTGAACCTGACTCCCAAATCCGCAAGTCTTG....GTGAGATTTAGATTTACctcgCTGTAATAC
F54B11.t1       TCGGCTACATTCCCCGGCTAGGACATGGCTT.TAACCA.TAGCACAGTGG.GAGCACCGCCAGGCAGTGTACCTGACTCC..........GTGCACAGCACTTGAAGAACGGATCGTC....CTTTAATCC
F53B1.t1        TCGGCCACAGTCACCGGCTGGAACGTAGCTT.AAATTA.TAGCCCAGTGG.GAGCGCCACCATGCAGTGTGGCTAACTCCCAGTCCCGCAGTGCATAGCGCTTGAAGAACGAATTGTC....TTTTAATCC
C18B12A.t1      TCGGCCACAGTCCCCGGCTAGAACGTAGCTT.AAATTA.TAGGCCAGTGG.GATCACCTCCAGGCAGTGTTTTTGAATCCAAGATCCGCAGTGCAAAGCGCTTGAAGAACGGATCGTC....ATTTAATCT
F40E10.t1       TAGGCCACAGTCCCTGGCTAGGACGTGGCTT.AAAAAA.TAGACCAGTGG.GATCACCACCAGGCAGAATCC.....TCCCAG.CCCGCAGTACATAGCACTTGAAGAACGGATCGTC....CTATAATTA
F40E10.t2       TTGGCCATAGTTCCA.GCCAGGACTTGGCTT.AAATTA.TAGCTCAGTGG.GAGCCCCACCAGACAGCGTACCTGGCTCCCAGATTCGCAGTGCATAGCGGTTAAAAAACAGATTGTT....CTTAAGTTT
C06G1.t1        TCGGTCACGGTCCCCGGCTTGAACGTGGCTC.AAATAA.TAGCCCGGTGG.GAGCCCCACCAGACAGTGCAACTGAATTCCAGATCCGCAGTGCATAACACCTGAAGTACGGATCGTC....CTTTAATCC
F09C8.t1        TTGGTCACAGTCCCC.GCTTGGACGTGGCTT.AAATCA.TGGTTCAGTAG.GAGCACAACCATACAGTGTACCTGAATCACTGCTCCCGCCGTGCATAGCGCTTGAAGAACAGTTCAAC....CTTTGATCC
```

Figure 3.5: ***C. elegans*** **tRNA-derived SINE element alignment (3' half).**

Chapter 4

A Computational Screen for Methylation Guide snoRNAs in Yeast[1]

[1]This chapter was co-written with Sean Eddy, and appears in a shortened version in Lowe & Eddy, *Science* **283**: 1168-1171, 1999.

4.1 Abstract

Numerous small nucleolar RNAs (snoRNAs) appear to be required for 2'-O-ribose methylations of eukaryotic ribosomal RNA. Most of the genes for this snoRNA family had been unidentified, despite the availability of a complete *Saccharomyces cerevisiae* genome sequence. Using new probabilistic modeling methods akin to methods used in speech recognition and computational linguistics, we computationally screen the yeast genome and identify 22 new methylation guide snoRNAs, snR50-snR71. Gene disruptions and other experimental characterization of these and other previously proposed guide snoRNAs confirm their methylation guide function. In total, we assign 51 of the 55 ribose methylated sites in yeast rRNA to 41 different guide snoRNAs.

4.2 Introduction

The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced, and is thought to contain about 6000 protein coding genes (Goffeau et al., 1996). However, this is not the total number of genes in yeast. Some of the largest eukaryotic gene families produce functional RNAs rather than protein products. Yeast contains approximately 140 tandemly repeated copies of ribosomal RNA genes (Goffeau et al., 1996) and 274 dispersed transfer RNA genes (Lowe & Eddy, 1997). The number of different identified functional RNAs is growing. In particular, a series of recent papers on small nucleolar RNAs (snoRNAs) has suggested the presence of large snoRNA gene families in eukaryotic genomes (Smith & Steitz, 1997; Tollervey & Kiss, 1997; Bachellerie & Cavaille, 1997).

snoRNAs appear to be involved at various stages of eukaryotic ribosome biogenesis, a complex process taking place in the nucleolus (Hadjiolov, 1985). Ribosomal RNA (rRNA) undergoes cleavages and modifications before assembly with ribosomal proteins into the mature ribosome (Woolford, 1991). Co-localized ribonucleoprotein particle complexes (RNPs) have been found to be essential for rRNA modifications (Tollervey et al., 1991; Mattaj et al., 1993). The three most common rRNA modifications are ribose methylation, pseudouridyla-

tion, and base methylation (Maden, 1990). The RNA component of these nucleolar RNPs, the small nucleolar RNAs, make up a diverse family of molecules that appear to fall into two major classes based on conserved sequence features: box H/ACA snoRNAs and box C/D snoRNAs (Balakin et al., 1996). Some H/ACA snoRNAs are required for specific pseudouridylations (Gannot et al., 1997; Ni et al., 1997). C/D box snoRNAs appear to have multiple roles in the nucleolus, one of which, rRNA ribose methylation, is the focus of this study.

Most C/D box snoRNAs contain one or more long 10-21 bp stretches of exact complementarity to ribosomal RNA (Bachellerie et al., 1995; Maxwell & Fournier, 1995). Many of these complementary regions within rRNA contain 2'-O-methyl modifications, which initially suggested that these snoRNAs might be involved in specifying the location of these modifications. Genetic disruption of U24 snoRNA in *S. cerevisiae* causes loss of the predicted target methyl groups (Kiss-Laszlo et al., 1996). The same study showed that alteration of the rRNA complementary region was sufficient to cause addition of a predictable ectopic methyl at a new position on rRNA. snoRNA depletion experiments in *Xenopus* oocytes have showed that methylation guide snoRNAs are necessary for specific methylation in vertebrates as well (Tycowski et al., 1996; Dunbar & Baserga, 1998).

The function of these ribose methylations remains unknown. The modifications are well conserved throughout eukaryotes, with more than 75% of 2'-O-methyl modified nucleotides in yeast aligning with homologous modified nucleotides in human ribosomal RNA (Maden, 1990). The modifications are located non-randomly in the most phylogenetically conserved regions of rRNA (Raue et al., 1988). Although their phylogenetic conservation suggests selective pressure, removal of two ribose methyls via genetic deletion of U24 snoRNA had no obvious effect on normal cell growth in yeast (Kiss-Laszlo et al., 1996).

The total number of rRNA ribose methyls in *Saccharomyces carlsbergensis*, a close relative of *S. cerevisiae*, has been estimated at 55 (Klootwijk & Planta, 1973). Forty-two of these methyls have been placed to specific nucleotide positions in the rRNA (Veldman et al., 1981; Raue et al., 1988; Maden, 1990). In *S. cerevisiae*, 11 previously isolated C/D box

snoRNAs have been predicted to be responsible for methylations at 12 sites (Kiss-Laszlo et al., 1996; Smith & Steitz, 1997), fewer than one fourth of the total ribose methylations. Experimental evidence supporting these predictions is available only for U24 (Kiss-Laszlo et al., 1996). If the hypothesis is correct that snoRNAs guide most or all ribose methylation in eukaryotes, most members of this gene family remain unidentified in *S. cerevisiae*.

Because the *S. cerevisiae*genome is completely sequenced (Goffeau et al., 1996), it is reasonable to consider identifying methylation guide snoRNAs computationally. However, sequence similarity of snoRNAs across phyla and within the gene family is generally weak, thus methods such as BLAST (Altschul et al., 1990) and FASTA (Pearson & Lipman, 1988) fail to identify new genes by similarity to known snoRNAs. Attempts have been made to identify snoRNAs by pattern searches based on the rRNA complementary guide sequence and other conserved features, but feature consensus is poor. If searches are limited to snoRNAs that occur within introns and that target known methylation sites (so the complementary region in rRNA is known), this strategy has been somewhat effective (Nicoloso et al., 1994; Nicoloso et al., 1996) since the false positive rate is minimized. However, in *S. cerevisiae*, most snoRNAs do not occur in introns, and a quarter of the rRNA methylations have not been precisely mapped.

Formal probabilistic models, based in part on methods used in speech recognition and computational linguistics, have been introduced for searching for complicated consensus features in biological sequence (reviewed in (Durbin et al., 1998)). Hidden Markov models (reviewed in (Eddy, 1996)) are probably the best known of these approaches. Another class of model called stochastic context-free grammars (SCFGs) has been used to construct probabilistic profiles of RNA genes that allow sensitive searching for RNA secondary structure (Eddy & Durbin, 1994; Sakakibara et al., 1994a). Using these probabilistic modeling techniques, we can produce an integrated model of snoRNAs that takes into account the rRNA complementary region, the consensus C, D and D' boxes, terminal stem base pairings, as well as the relative position of these features within the snoRNAs. Once defined, the snoRNA gene model can be trained on previously identified, "trusted" members of the

gene family, and updated as new snoRNAs are found and verified.

The combination of probabilistic modeling approaches and the availability of the complete genome for *S. cerevisiae* has made it feasible for us to execute a "computational genetic screen" for the missing members of the methylation guide snoRNA family. In this study, we have identified 22 new guide snoRNAs, and experimentally verified guide function for all but one which appears to be genetically redundant. Combined with verification of new methyl target sites for other known snoRNAs, we can now assign a guide snoRNA to all but 4 of the 55 ribose methyl sites in *S. cerevisiae* rRNA.

4.3 Experimental Procedures

4.3.1 snoRNA Search Algorithm and Model Scoring

The snoRNA search algorithm is diagrammed in Figure 4.1. The program sequentially searches for snoRNA features in the query sequence in the following order. A box D sequence matching the pattern "(A/C)UGA" is identified. The highest scoring box C sequence (7 bp pattern scored by log odds weight matrix) is located 35-200 bp upstream from box D. The intervening sequence is checked for an rRNA complementary sequence of 9 bp or greater, allowing a maximum of three mismatches and any number of G-U pairings. The highest scoring box D' sequence (four bp pattern scored by log odds weight matrix) is identified just 3' to the rRNA complementarity if the rRNA match is not immediately adjacent to the D box. Finally, the rRNA methylation site guided by the candidate snoRNA is predicted by counting five bp upstream of box D or D'.

Each candidate snoRNA alignment was then scored against our probabilistic model (Table 4.1). SnoRNAs were ranked based on a final log odds score (Barrett et al., 1997) that incorporated information from each of the snoRNA features. The initial model was trained on 35 human C/D box snoRNAs proposed to function as methylation guides (Kiss-Laszlo et al., 1996). Nine previously isolated yeast snoRNAs were shown to match to this snoRNA gene model with significant scores (25.91 - 43.55 bits). In a search of randomly

generated sequence[2] equivalent in size to four complete yeast genomes, the maximum score for a false positive (29.65 bits) exceeded the score for only one of the nine known snoRNAs. Thus we believed we had sufficient training data to search for unidentified snoRNAs in the yeast genome.

4.3.2 snoRNA Gene Disruptions

snoRNA disruptions were generated by homologous gene replacement in S. cerevisiae haploid strain yM4585 (Mat a his3Δ200 lys2-801 leu2-3,2-112 trp1-901 tyr1-501 URA3+ ADE2+ CAN^S) and diploid strain yM4587 (Mat a/Mat α his3Δ200/his3Δ200 lys2-801/ lys2-801 leu2-3,2-112/ leu2-3,2-112 trp1-901/ trp1-901 tyr1-501/ tyr1-501 URA3+/ URA3+ ADE2+/ ADE2+ CAN^S/ can^r) kindly provided by M. Johnston. The disruption scheme is as described by Baudin *et al.* (1993), using a protocol provided by L. Riles. Disruption constructs were generated by the polymerase chain reaction (PCR) using forward and reverse primers with 5' ends containing 41 bp of genomic sequence flanking the predicted snoRNA, plus 19 bp matching a bacterial vector pBM2815 (from M. Johnston) containing the HIS3 marker gene. The resulting constructs contained the HIS3 gene bordered by 41 bp of genomic sequence found just upstream and downstream of the target snoRNA gene. Both haploid and diploid strains were transformed with disruption constructs using a standard LiOAc transformation protocol (Schiestl et al., 1993). Transformants growing on YPD His- plates were picked and assayed by PCR for correct integration of the HIS3 marker gene replacing the target snoRNA. In several cases, diploid transformants were sporulated to obtain haploid disruption mutants.

4.3.3 Mapping of rRNA Ribose Methylations by Primer Extension

Reverse transcriptase primer extensions were carried out on total RNA with 22-26 nt mapping primers complementary to ribosomal RNA. The sequences of all mapping primers are

[2]Random sequences were generated by a fifth order Markov chain based on 6mer frequencies within the yeast genome.

available by WWW (Lowe & Eddy, 1998). Total yeast RNA (0.4 μg/μl) was annealed with end-labeled mapping primers (0.15 pmol/μl) at 60 $^\circ$C for 4 min. Primer extensions were carried out in 5 μl reactions containing 0.8 μg RNA and 0.3 pmol ^{32}P end-labeled primer in the presence of 50 mM Tris-Cl pH 8.6, 60 mM NaCl, 9 mM $MgCl_2$, 10 mM dithiothreitol, 1 mM each dNTP, and 0.2 U/μl AMV reverse transcriptase for 30 minutes at 37 $^\circ$C. Low dNTP concentration reactions were carried out the same except using 0.004 mM of each dNTP and 5 mM $MgCl_2$. Each reaction was analyzed by electrophoresis next to an RNA sequencing ladder on an 8 % polyacrylamide gel. RNA sequencing ladders were generated by AMV reverse transcription in a similar manner as the above primer extensions. Total yeast RNA (0.2 μg/μl) was annealed with end-labeled mapping primers (0.15 pmol/μl) at 60 $^\circ$C for 4 min. Sequencing primer extensions were carried out in 5 μl reactions containing 0.4 μg RNA and 0.3 pmol end-labeled primer in the presence of 50 mM Tris-Cl pH 8.6, 60 mM NaCl, 10 mM $MgCl_2$, 10 mM dithiothreitol, 0.33 mM each dNTP, 0.2 mM one of four (A,C,G,T) dideoxy NTPs, and 0.2 U/μl AMV reverse transcriptase for 30 minutes at 37 $^\circ$C.

4.3.4 Verification of snoRNA Transcription and 5'-ends

snoRNA transcription was assessed by reverse transcription of total RNA with oligonucleotide primers complementary to internal snoRNA sequence. Sequences of internal snoRNA primers are available by WWW (Lowe & Eddy, 1998). Primer extension reactions were carried out using the same conditions as used for 1 mM [dNTP] rRNA primer extensions. RNA sequencing ladders were run adjacent to snoRNA primer extensions on 8% polyacrylamide gels to assess fragment lengths. Primer extensions on RNA from snoRNA-disrupted yeast strains were run next to primer extensions with wild-type strain RNA to verify loss of the snoRNA band of expected size.

4.4 Results

4.4.1 Computer Search Algorithm and Probabilistic snoRNA Model

We implemented a greedy search algorithm to identify 2'-O-methylation guide snoRNAs in genomic sequence. The program sequentially identifies six components characteristic of these genes (see Figure 4.1): box D, box C, a region of sequence complementary to ribosomal RNA, box D' if the rRNA complementary region is not directly adjacent to box D, the predicted methylation site within the rRNA based on the complementary region, and the terminal stem base pairings, if present. The program also notes the relative distance between identified features within the snoRNA, information we found critical to reducing the false positive identification rate.

Each candidate snoRNA alignment is scored against a probabilistic model (Figure 4.2) trained on experimentally verified yeast or human snoRNAs. snoRNAs are ranked based on the final log odds score (Barrett et al., 1997) incorporating information from each of the snoRNA features. Although a dynamic programming algorithm incorporating the probabilistic model at the initial search phase could have been used, we opted for a greedy search followed by probabilistic scoring in the interest of speed. A final report is generated for each snoRNA, including component features and scores plus the target rRNA methylation site. Initial profiles of snoRNA features were provided by Kiss-Laszlo *et al.* (1996) as a consensus structure for methylation guide snoRNAs that was based on 21 novel and 14 previously identified human snoRNAs. Nine previously isolated yeast snoRNAs were shown to conform to this snoRNA gene model, thus we believed we had sufficient training data to search for unidentified snoRNAs in the yeast genome. As new snoRNAs were identified and verified, they were added to the model training set.

4.4.2 snoRNAs Assigned to 39 of 42 Known Ribose Methyl Sites

We began our search for new snoRNAs by identifying family members that target known 2'-O-ribose methyl sites. Because there is very little information on ribose methyl modifications

for *S. cerevisiae* rRNA, we inferred the position of 42 ribose methyls based on mapping data from *S. carlsbergensis* (Maden, 1990). We applied the snoRNA search program to the sequence of *S. cerevisiae*, and extracted from the output only candidate snoRNAs that could target one of the inferred methyl sites. These candidates were divided based on target methyl site and sorted by score, producing 42 different lists of best-to-worst snoRNA predictions, one for each methyl site. Depending on search parameter cutoffs and the specific target methylation site, the program found dozens to over a hundred predictions for each methylation site. Candidates overlapping predicted protein coding regions were noted and disfavored relative to other strong, non-overlapping candidates. Seven previously published snoRNAs have been predicted to guide methylation at eight of the 42 sites (U14, U18, U24, snR39, snR39b, snR40, snR41; see Table 4.2). Our searches did not show improved snoRNA predictions over the previously identified snoRNAs, so we did not pursue new assignments for these eight sites.

We tested the top scoring snoRNA gene predictions corresponding to the remaining sites by gene disruption (Baudin et al., 1993). Each snoRNA-disrupted strain was tested for the ability to methylate at the predicted rRNA site by a dNTP concentration-dependent primer extension assay (Maden et al., 1995; Kiss-Laszlo et al., 1996). Out of 30 gene disruptions, 24 loci were verified as methylation guide snoRNAs. Seven of these had been previously identified as C/D box snoRNAs (Table 4.2: snR13, snR47, snR48, snR74, snR76, snR77, snR79). Seventeen snoRNAs were new (Table 4.2: snR50-snR57, snR60-snR63, snR66, snR68-snR71). Two sample primer extension gels demonstrating typical loss of rRNA methylation sites for snoRNA disruption mutants appear in Figures 4.3 and 4.4. Primer extension assays for two of the snoRNA disruption mutants, snR55 and snR70, showed a noticeable but minor change in the primer extension pattern at the expected sites (Figures 4.5 and 4.6, thus we qualify these assignments as "inconclusive". Twenty-three additional primer extension gels for the other verified snoRNAs can be found at (Lowe & Eddy, 1998). None of these snoRNA gene disruptions was lethal, nor did we observe impaired growth on rich media.

Given the proposed rule of one snoRNA per methylation site, 24 verified guide snoRNAs implies assignments to 24 of the 34 known but unassigned methyl sites. However, we found that some of these snoRNAs guide modification at more than one methylation site, as previously seen for U24 (Kiss-Laszlo et al., 1996). The search program predicted and we experimentally verified one additional methylation target site for snR47, snR48, and snR51 (Table 4.2). We also found an additional target site for snR41, a snoRNA previously predicted to guide at a different methyl site (Kiss-Laszlo et al., 1996). We verified snR41 methylation guide function for both the previously predicted site (SSU-Gm1123) as well as the newly predicted site (SSU-Am541). With these additional site assignments, 28 of 34 known but previously unassigned sites can be attributed to guide snoRNAs.

We have made a tentative methylation assignment to one additional new C/D box snoRNA, snR59 (Table 4.2), whose expression we have verified (see below). We predict snR59 guides methylation at the same site already assigned to snR39, LSU-Am805. Neither snR59 nor snR39 has been checked for functional redundancy. Both snoRNAs are intronic, thus we did not attempt to generate null mutants. The homologous knockout method that we chose uses a large marker gene to replace the target snoRNA, and we have observed that such insertion-based disruptions appear to interfere with host protein intron splicing.

Our searches gave no strong snoRNA candidates for four of the remaining six methylation sites: LSU-Cm648, LSU-Gm1448, LSU-Am2279, and LSU-Gm2919. The one common factor among these sites is that they are all one nucleotide adjacent to methyl sites for which snoRNA assignments or strong predictions have been made. This led us to believe that an unusual snoRNA interaction may be responsible for methylation in these cases. Our disruption of snR13 showed loss of the predicted target LSU-Am2278 (described above), as well as the ribose methyl one nucleotide adjacent, LSU-Am2279 (see Figure 4.4). U24 disruption previously showed loss of LSU-Am1447, as expected, plus the adjacent methyl at LSU-Gm1448 (Kiss-Laszlo et al., 1996). Taking these data into consideration, we predicted that U18, in addition to modifying LSU-Am647 (Kiss-Laszlo et al., 1996), may also direct methylation at LSU-Cm648. We disrupted the intron-encoded U18 to test our prediction,

but could not assay a mutant haploid clone since our deletion was lethal. Because U18 is nonessential (Balakin et al., 1996), we believe we inadvertently disrupted function of the essential host gene, elongation factor-1β. We also have a strong candidate snoRNA targeting the site adjacent to Gm2919, snR52. Our disruption of snR52 did not result in loss of either Um2918 or Gm2919, possibly due to functional redundancy (see discussion). We present a model in the discussion that may account for the two observed (snR13, U24) and one hypothesized (U18) methylation assignments.

Out of the 42 previously mapped ribose methyl sites, this leaves three sites for which we could not assign a C/D box snoRNA: SSU-Am436 for which we have no prediction, plus LSU-Um2918 and LSU-Gm2919, for which we could not confirm our snR52 prediction.

4.4.3 snoRNAs Assigned to 12 of 13 Previously Unmapped Ribose Methyl Sites

Initial estimates of the total number of ribose methyl sites derive from early studies of ribosomal RNA modifications in *Saccharomyces carlsbergensis* (Klootwijk & Planta, 1973). These experiments implied 55 distinct 2'-O-ribose methyl groups based on two dimensional gel analysis of ^{14}C-methyl labeled, ribonuclease T_1-digested rRNA. In these experiments, 2D gel spots, each containing a T_1 fragment, were excised and analyzed for the methylated nucleotide and the surrounding sequence. Using the sequence context information, the locations of 42 ribose methyls were mapped unambiguously to the rRNA sequence. 13 ribose methyls could not be placed precisely due to insufficient sequence information within the T_1 fragment (Veldman et al., 1981; Raue et al., 1988).

From the results we described in the previous section, we knew that the *S. carlsbergensis* methyl site data agreed well with observed concentration-dependent stops in *S. cerevisiae* rRNA primer extensions. Although one report has described difficulty detecting some ribose methylation sites in vertebrate rRNA by primer extension methods (Yu et al., 1997), our slightly optimized protocol allowed clear detection of all but one of the 42 mapped *S. carlsbergensis* ribose methyl sites in *S. cerevisiae*. For SSU-Cm1637, we could only detect

a weak, non-concentration dependent band, likely due to two strong stops adjacent to the putative ribose methyl (Figure 4.6). Based on our ability to visualize the known ribose methyl sites by primer extension assay, we believed we would be able visualize unmapped sites as well.

We used three lines of evidence to predict, then experimentally verify the position as well as the snoRNA assignment for each of the unmapped sites. First, we knew between one and five nucleotides of sequence context for each of these sites based on known sequences for T_1 ribonuclease digest fragments of rRNA that contain the unplaced ribose methyl groups (Klootwijk & Planta, 1973; Veldman et al., 1981). Second, we checked the existing collection of C/D box snoRNAs for previously unrecognized rRNA complementary regions that could target sites not included in the list of known ribose methyls. Third, we went back to the *S. cerevisiae* genome search results from our program and extracted all high scoring snoRNAs that could target new rRNA methyl sites.

Using these methods, we identified and verified 12 of the 13 unmapped methyl sites by primer extension. Six new sites were assigned to known C/D box snoRNAs, and the other six were assigned to newly identified snoRNAs (Table 4.2, target sites in boldface). Each of the 12 new methyl sites can be correlated with a T_1 digest fragment for one of the 13 unmapped ribose methyls. We could not identify the location of the single unmapped methyl site in small subunit rRNA (T_1 fragment GmU). snR190 has also been predicted to target a potential methylation site at LSU-Gm2393 (Kiss-Laszlo et al., 1996). In our primer extension assay, this site does not give a visible band, nor does its sequence context correspond to an unassigned T_1 fragment.

For each of the 12 newly mapped sites, we disrupted the corresponding guide snoRNA (except the intronic snR38 gene), and confirmed loss of the expected methylation site (see Table 4.2, snoRNAs assigned to methyl sites in boldface). None of the verified guide snoRNAs was found to be essential, nor did gene disruption cause noticeably impaired growth.

As in the previous section, several of these snoRNAs guide methylation at more than one site. The snR40 disruption showed loss of the newly mapped LSU-Um896 in addition to the

previously predicted SSU-Gm1267 (Kiss-Laszlo et al., 1996). For snR60, gene disruption showed loss of newly mapped LSU-Gm906, as well as previously mapped LSU-Am815 (see previous section). snR67 disruption showed loss at two newly mapped sites (Table 4.2).

We took advantage of the tandem arrangement of seven snoRNA genes snR72 through snR78 and constructed a septuple deletion mutant of these genes. The septuple snoRNA deletion mutant was still viable with no obvious change in growth rate on rich media. While we did construct a single locus deletion mutant for snR78, we tested the rest of the snoRNAs in the tandem array (snR72-snR77) via the septuple mutant. We therefore cannot prove a 1:1 mapping of the six methyl site losses to snR72-snR77, although the strong rRNA complementarities for each snoRNA support this conclusion.

For each snoRNA disruption mutant, we checked the status of the predicted target methyl site(s) as well as at least two other sites in the neighboring rRNA region. We observed specific methyl site loss only at the predicted target site in all instances but one. In the case of the tandemly grouped snoRNAs, we observed a polarity effect in that snoRNAs downstream from those being disrupted showed either partial or complete loss of methylation at their target sites (unpublished).

4.4.4 Expression and 5' ends of New Yeast snoRNAs Verified

As an additional line of evidence to confirm our 22 newly identified snoRNA loci, we verified gene expression by primer extension, and mapped 5' snoRNA ends to nucleotide resolution. Most previously isolated human and yeast guide snoRNAs end four to five nucleotides upstream of the C box, and two to four bases downstream of the D box ((Kiss-Laszlo et al., 1996); Genbank C/D box snoRNA entries). Our program assumes a 5' end that is four nucleotides upstream of the C box and two nucleotides 3' of the D box. All 5' ends we mapped for new snoRNAs occurred 4-5 nucleotides from the true C box (Figure 4.7). We also assayed the 5' end of snR13 since the Genbank entry showed it extends 21 nucleotides 5' of the C box, the only guide snoRNA exception to the 4-5 bp rule. Our primer extension agreed with the unusually long 5' end given in the snR13 Genbank entry (SCU16692).

5' end mapping allowed us to check that the program had chosen the correct C boxes of new snoRNAs. C' boxes have been observed to occur in the interior of snoRNAs (Kiss-Laszlo et al., 1996; Kiss-Laszlo et al., 1998), and could be mistaken for legitimate C boxes. For 21 of 22 new snoRNAs tested, our program picked a C box that matched the experimentally determined 5' end. One snoRNA, snR63, surprised us by giving a primer extension product over 200 bp in length, far longer than any other methylation guide snoRNA identified. Assuming that the 3' end D box prediction is correct, snR63 appears to be 255 bp in length.

Transcription of the only two apparently redundant snoRNAs, snR59 and snR39, was also verified. Since these snoRNAs are intronic, mapping of the predicted mature 5' ends showed that both appear to be processed from their host gene mRNA transcripts.

Finally, we also tested each snoRNA disruption strain to verify that that we had eliminated snoRNA production completely. Each disruption strain showed complete loss of the snoRNA extension product of the expected size (Figure 4.7).

4.4.5 snR70 Identified by Comparative Genomics

We also used the program to scan genome sequence from other eukaryotes including human, *C. elegans*, and *Schizosaccharomyces pombe* (manuscript in preparation). We identified *S. cerevisiae* snR70 by comparison with our results from the *S. pombe* snoRNA search. One tandem group of three *S. pombe* snoRNAs appears to be syntenic with three tandem *S. cerevisiae* snoRNAs (snR41, snR70, snR51). The 5' most snoRNAs in each array both target SSU-Gm1123, the middle snoRNAs both target SSU-Cm1637, and the 3' most snoRNAs both target LSU-Um2726. Our snoRNA search program had not identified snR70 as one of the top snoRNA candidates due to what appears to be a single nucleotide bulge within the snoRNA/rRNA paired region. Given the strong prediction for the middle *S. pombe* snoRNA, we then re-examined the sequence between *S. cerevisiae* snR41 and snR51, and found snR70. As described above, we verified expression of snR70, and observed loss of the weak stop at Cm1637 for the disruption mutant. Considering the evolutionary distance

between these organisms, genuine synteny would imply that these snoRNAs are at least one billion years old.

4.5 Discussion

Using a computational genetic approach, we have identified 22 novel methylation guide snoRNAs in Saccharomyces cerevisiae, all of which are expressed, and all but one of which has been functionally verified by gene disruption. We have also identified and verified 12 of the remaining 13 unmapped ribose methyl sites in *S. cerevisiae* ribosomal RNA based on strong snoRNA predictions. Four of these new sites were recently inferred from existing C/D box snoRNAs and confirmed independently by another research group (Cavaille & Bachellerie, 1998).

4.5.1 snoRNAs Assigned to 51 of 55 Total Ribose Methyl Sites

By snoRNA gene disruption, we have verified snoRNA-directed modification for 41 of the 55 total ribose methylation sites in ribosomal RNA. Three additional sites for U24 have previously been demonstrated (Kiss-Laszlo et al., 1996). Five other sites are strongly predicted to be guided by experimentally isolated C/D box snoRNAs but have not been confirmed yet because the deletion is lethal (U14), or the snoRNA is in an intron (U18 x 2 sites, snR38, snR39/snR59). Two more sites have been tentatively assigned to newly identified, expressed C/D box snoRNAs (snR55, snR70). This leaves four sites for which we could not assign a prediction (SSU-Am436), locate the methyl site (SSU-Gm?), or experimentally verify a prediction (LSU-Um2918, LSU-Gm2919). We believe there are several explanations for the inconclusive or missing snoRNA assignments.

Technical difficulties with the primer extension assay are the most likely cause for the two inconclusive methyl site assignments. The occurrence of other modifications or secondary structures on or near the target rRNA nucleotide can make interpretation of primer extensions difficult. For example, there is a very strong stop on wild-type rRNA at SSU-

G1266 that prevents almost all read-through by the reverse transcriptase (Figure 4.5, lanes 5 and 6). This primer extension stop has been previously observed, and hypothesized to be some type of base modification (Bakin & Ofengand, 1995). For the snR55 disruption mutant, a noticeable but far from complete loss of the stop at G1266 is visible (Figure 4.5, lane 10), as well as improved read-through of larger products. We believe that the decreased intensity of the stop indicates loss of the Um1265 ribose modification, although we cannot be certain by this assay. We also believe at least one unknown base modification or secondary structure element at SSU-C1640 and/or SSU-G1641 is responsible for difficulty in visualizing the small, faint ribose methyl stop for SSU-Cm1637 (Figure 4.6, lanes 5 and 6). The uncharacterized, strong stops just downstream of Cm1637 make loss of the weak stop in the snR70 disruption mutant (lanes 7 and 8) somewhat less convincing. Use of an alkaline hydrolysis primer extension assay (Kiss-Laszlo et al., 1996), or another apparently more sensitive methyl assay (Yu et al., 1997) may give clearer results in these cases.

Incomplete or undetected methyl site loss could also be due to functional redundancy of snoRNAs. The snR52 disruption mutant showed no change at one of its two predicted methylation sites, LSU-Um2918. However, we believe that snR52 may still be involved in modification at Um2918 based on a perfect 11 bp rRNA complementarity, a high snoRNA score of 34.46 bits, and the fact that snR52 has already been confirmed to guide methylation at another methyl site. In this case, we believe an unidentified, functionally redundant snoRNA may exist.

A more obvious example of functional redundancy appears to exist between snR39, and an apparent homologue, snR59. Both snoRNAs are intron-encoded, one within the ribosomal protein gene YL8A on chromosome VII, and one within YL8B on chromosome XVI. The snoRNAs appear to have been duplicated relatively recently with their host genes (Mizuta et al., 1995). An alignment of the snoRNAs shows that only a single nucleotide differs between the box C, box D, and rRNA complementary regions. We did not detect any other pairs of close homologues, although less similar but functionally redundant snoRNAs could exist. Multiple disruptions of potentially redundant snoRNAs may be necessary to

test for methylation guide function.

It is possible that some sites may be modified by a different mechanism. Prokaryotes contain a fraction of the rRNA modifications found in eukaryotes and do not appear to contain snoRNAs. Thus, a handful of site-specific enzymes may accomplish these modifications without snoRNAs. Conservation of such enzymes in yeast would obviate the need for several specific guide snoRNAs. In yeast, three 2'-O-methyl groups (SSU-Cm1637, LSU-Gm2616, LSU-Um2918) are conserved with homologous modified positions in prokaryotes (*E. coli* SSU-Cm1402, LSU-Gm2251, and LSU-Um2552, respectively (Raue et al., 1988)). For one of these sites, we have found and confirmed a yeast guide snoRNA (snR67). However, at the other two conserved sites, we have either failed to verify a guide snoRNA (Um2918), or failed to obtain definitive evidence of methyl site loss (Cm1637). Protein methyltransferases targeting these specific sites may account for our difficulty in finding and/or verifying guide snoRNAs in these cases.

4.5.2 A Nearly Complete Set of Methylation Guide snoRNAs in Yeast

In summing all expressed snoRNAs that have either been verified or strongly predicted to guide ribose methylation, we count 41 genes that can be assigned to 51 rRNA methylation sites (Table 4.2). We estimate that up to two methylation guide snoRNAs remain to be identified for the two unassigned methylation sites (SSU-Am436, SSU-Gm?), and two to four snoRNAs may be identified as being redundant with known snoRNAs for SSU-Um1265, SSU-Cm1637, LSU-Um2918, and LSU-Gm2919.

In addition to the 22 new guide snoRNAs we identified and verified (Table 4.2, snoRNAs in boldface), we were able to verify that a number previously identified C/D box snoRNAs guide methylation at additional methyl sites. The search program pointed out that snR40, snR41, snR47, and snR48 each contain additional methylation target sites, each of which we were able to verify experimentally. The program was also able to classify snR13, a previously known C/D box snoRNA for which the function was undetermined (Smith et al., 1997).

Eight previously identified C/D box snoRNAs that we predicted as methylation guide

snoRNAs and verified experimentally appear in Genbank as the "Z" snoRNAs (SCSNORZ2 - SCSNORZ8, SCZ9SNOR; Zhou, H. and Qu, L.H., unpublished). It was recently suggested but not experimentally shown that these are legitimate guide snoRNAs (Cavaille & Bachellerie, 1998). With demonstration of function, we suggest assigning the standard "snR" names snR72 through snR79 to Z2 through Z9. While this manuscript was in preparation, seven additional Z snoRNAs were deposited in Genbank as Z10-Z16 (Zhou, H. and Qu, L.H.), again with no reference to function or scientific publication. All seven correspond to snoRNAs independently identified, assigned to methylation sites, and verified as novel snoRNAs in this work (Table 4.2), and we propose snR names for them as well.

Some snoRNAs may direct other modifications not detected by the dNTP concentration-dependent primer extension assay, or may have other functions in assembly of the ribosome. Currently, only three identified C/D box snoRNAs have no demonstrated function: snR4, and snR45, and snR190. All three have been shown to be nonessential (Zagorski et al., 1988; Balakin et al., 1996). Our search program did not detect any rRNA complementarities in snR4 or snR45, thus we cannot predict their function. For snR190, our program agrees with a possible target modification site previously suggested by Kiss-Laszlo *et al.* (1996). However, based on experimental evidence, we believe such a ribose methyl site does not exist or is too weakly methylated to be detected by our rRNA methyl mapping experiments.

4.5.3 Multiple Methylations Guided by Single snoRNAs

Although U24 was shown to have two methyl targets, it has been proposed that one snoRNA generally modifies one site. In this work, we have observed three distinct ways in which a single snoRNA appears to direct multiple ribose methyl modifications. The first and most obvious involves "double guide" snoRNAs which contain two different guide sequences, one located at the 5' end of the snoRNA, the other at the 3' end. A double guide snoRNA has been previously observed for human and yeast U24 (Kiss-Laszlo et al., 1996). To this, we add six new double-guide snoRNAs (snR40, snR41, snR47, snR51, snR60, snR67), three of which are at new loci. snR60 is particularly interesting given that the two nucleotides it

modifies are on opposing sides at the base of a 12 bp helix formed in the rRNA secondary structure (Gutell, 1994). It could be imagined that snR60 acts as a chaperone to bring the two ends of the helix together, expediting rRNA folding. However, this is the only example which shows obvious target site proximity. U24, snR41, and snR67 also guide modification of nucleotides within the same ribosomal subunit, although a spatial relationship as in the case of snR60 is not apparent from rRNA secondary structure predictions (Gutell, 1993; Gutell et al., 1993). snR40, snR47, and snR51 modify nucleotides on different ribosomal subunits, thus proximity is difficult to estimate.

A second method appears to utilize two different D' boxes with the same complementary region to guide methylation at two different sites. Our disruption of snR48 resulted in loss of methylation at Gm2790 and Gm2788, implying that the first methylation site is measured from the D' box "AUGU" and the second by D' box "GUUA" (the "GU" overlaps between D' boxes). These are the two most atypical D' boxes among all confirmed yeast snoRNAs to date. Even so, none of the canonical traits of methylation guide snoRNAs (Kiss-Laszlo et al., 1996) are violated. A second example of this is predicted to occur in a newly identified *S. pombe* snoRNA but at a different pair of methylation sites (manuscript in preparation).

For U24, snR13, and U18, an additional adjacent ribose methyl modification may be due to a bulge within the snoRNA-rRNA duplex. Previous disruption of U24 has been observed to result in loss of methylation at Am1447 and unexpectedly at Gm1448 (Kiss-Laszlo et al., 1996). Our disruption of snR13 results in loss of Am2278 and at Am2279. In both cases, we think that a single nucleotide bulge within the snoRNA could "slide" the rRNA target one base pair closer to the reference D' box without disrupting the necessary base pairings (see Figure 4.8). The one nucleotide slide places the adjacent site the canonical modification distance (5 bp) away from the D' box. In addition to the sites modified by snR13 and U24, two other pairs of adjacent sites in the rRNA may be modified in a similar manner, LSU-Am647/Cm648 and LSU-Um2918/Gm2919. U18 is predicted to modify LSU-Am647, but would allow a nucleotide bulge in the snoRNA-rRNA duplex to guide at LSU-Cm648 as well (Figure 4.8). We were not able to assay a U18 disruption mutant so this interaction

is still hypothetical. Although we could not verify snR52 assignment to the remaining pair of adjacent methylation sites at LSU-Um2918 and LSU-Gm2919, this snoRNA fits the bulge model as well (Figure 4.8). Kiss-Laszlo *et al.* (1996) proposed an alternative mechanism, that loss of the adjacent methyl site (Gm1448) for the U24 disruption could be due to involvement of an independent methyltransferase that requires a ribose methyl site for sequential addition of an adjacent methyl site.

4.5.4 Methylation Guide snoRNA Consensus Structure

With a large number of the 2'-O-methylation guide snoRNAs identified and confirmed in *S. cerevisiae*, we now have an improved consensus structure for this gene family. The structure is similar to that presented by Kiss-Laszlo *et al.* (1996), now with more clearly defined lengths between features. Some aspects of the expanded sample of snoRNA features are worth noting. Two snoRNAs break with the canonical "CUGA" D box sequence. There is a fairly tight length distribution between features, which we used to our advantage to eliminate false positives. For the D' box guided sites, the complementary sequence always occurs within 25 bp of the C box, and the gap between the D' and D boxes (3' end) is always two or more times longer than the gap between the C box and complementary sequence (5' end). Thus, the complementary sequence never occurs directly in the middle of snoRNAs. We also noted a strong bias towards uridine at the nucleotide immediately 5' to the guide D or D' box (U 61%, A 26%, G 9%, C 4%). This may be a valid extension of the D/D' box motif. Examination of the snoRNA-rRNA duplex matches/mismatches (Table 4.2) shows that 17 duplexes contain at least one G-U pairing. Almost as many duplexes contain non-Watson-Crick, non-G-U pairs as well, violating the commonly described "perfect" stretches of snoRNA/rRNA pairing. Also, no clear patterns emerged for snoRNAs containing terminal stem loops. All intronic snoRNAs had at least weak terminal stems, although roughly half of the extragenic and tandem array snoRNAs also contained stems. Double guide snoRNAs also do not show a consistent pattern for stem formation.

4.5.5 Genomic Organization of the snoRNA Gene Family

With nearly all methylation guide snoRNAs identified, we can assess the general genomic organization of the gene family (Table 4.2). All chromosomes except VI contain at least one methylation guide snoRNA. Most are dispersed as independent singlets or within five small clusters of 2-7 tandemly arrayed guide snoRNAs. A total of 19 singlets occur outside of known protein coding genes, presumably as independent transcription units. All tandemly arrayed snoRNAs within the same cluster are oriented on the same strand, and are separated by between 80 to 148 bp. Recent results indicate these genes are polycistronic (Petfalski et al., 1998; Chanfreau et al., 1998a; Chanfreau et al., 1998b; Qu et al., 1999). Six yeast snoRNAs occur within the introns of host protein genes, all on the pre-mRNA coding strand. The mixture of snoRNAs in yeast occurring within introns, tandem arrays, and as singlets is in contrast to vertebrates, where all currently known guide snoRNAs are within host gene introns. Polycistronic arrays of snoRNAs have also been reported in plants (Leader et al., 1997; Shaw et al., 1998). Some plant polycistrons contain a mix of snoRNAs from both major families of guide snoRNAs (C/D box and H/ACA box snoRNAs), whereas none of the yeast tandem arrays contain members outside of the C/D box family.

It has long been noted that C/D box snoRNAs often occur in the introns of ribosomal proteins in vertebrates (Maxwell & Fournier, 1995). Only two C/D box snoRNAs, snR39 and snR59, occur in ribosomal proteins of yeast, but an unexpectedly large number, ten, occur immediately adjacent to ribosomal proteins. These snoRNAs occur more often on the opposite strand than the same strand as the ribosomal protein genes.

4.5.6 Implications for Genome Sequence Analysis

One of the goals of genome sequencing is to identify all the genes in an organism. Computational methods for protein coding gene identification are reasonably well developed, especially for compact genomes with few or no introns. Protein coding genes have open reading frames, codon bias, and other telltale statistical signals that can be recognized. On

the basis of such algorithms and other genetic characterization, the yeast genome is said to contain 6000 genes and to have a coding density of about 75% (Goffeau et al., 1996).

These genefinding algorithms do not attempt to search for noncoding functional RNA genes. Examples of noncoding functional RNAs have been known for decades, but their diversity and numbers seem small. New discoveries of enigmatic noncoding RNA genes, such as the mammalian tumor suppressor *H19* (Brannan et al., 1990) or the mammalian X-dosage compensation gene *Xist* (Brockdorff et al., 1992; Brown et al., 1992), are interesting but perhaps exceptional. However, it seems possible that, in fact, a large number of noncoding RNAs remain to be discovered; not only computational screens but experimental screens tend to be biased against RNAs. Many functional RNAs are not polyadenylated, so are not well represented in oligo-dT primed cDNA libraries or in EST sequencing projects. Many RNAs are small genes that occur in redundant copies, and RNAs are of course not affected by stop codons or frameshifts, so they are probably somewhat refractory to genetic screens. To date, most functional RNAs have probably been identified by biochemical means.

Here, we have extended the known gene family of methylation guide C/D box snoRNAs to 41 loci in yeast. Pseudouridylation guide snoRNAs are probably encoded by another large dispersed gene family (Ni et al., 1997; Gannot et al., 1997). Yeast genome sequence analysts probably would not have guessed that careful computational analyses had missed the presence of two large gene families and almost 100 new genes. By themselves, the snoRNAs do not substantially alter the estimate of 6000 genes in yeast, nor the 75% coding fraction. However, given that one or two large gene families of functional RNAs escaped detection, how many others are there? How much "extragenic" DNA is actually encoding functional RNAs? How many of the systematic gene knockouts being generated in yeast will also knock out an unsuspected RNA gene (especially intronic ones), and thus superpose two genetic phenotypes on the resulting disruption? Using probabilistic models, we are beginning to gather the tools necessary to computationally screen genome sequences and answer some of these questions.

4.6 Acknowledgments

We are grateful to Linda Lutfiyya and Linda Riles from the M. Johnston lab for protocols and guidance in all aspects of yeast handling, gene disruptions, colony PCR and RNA preparations. We would also like to thank Skip Fournier, Jingwei Ni, Dmitry Samarsky, and Ted Maden for helpful discussions and sharing of unpublished observations. Thanks to Steve Johnson and Linda Lutfiyya for careful reading of the manuscript. This work was supported by NIH grant number R-01-HG01363 and by a gift from Eli Lilly.

4.7 Data availability

All computer code, snoRNA search results, oligonucleotide primers, rRNA primer extensions gel, and other referenced data can be found on line (Lowe & Eddy, 1998). All new snoRNAs (snR48, snR50-snR71) have been submitted to the Saccharomyces cerevisiae Genome Database (SGD; http://genome-www.stanford.edu/Saccharomyces/), and can be accessed directly by searching for SNR locus names (e.g., "SNR50", or "SNR*"). Sequences are available in Genbank by accessions AF06461-AF064283 for snR48, snR50-snR71, respectively. Other yeast snoRNA Genbank accession numbers are as follows: snR190 and U14 (X96815), U18 (U12981), U24 (Z48760), snR13 (U16692), snR38 (U26012), snR39 (U26011), snR39b (X94605), snR40 (U26015), snR41 (U26016), snR47 (U56648), Z2-Z8 (Z69294-Z69300), Z9 (Z70300).

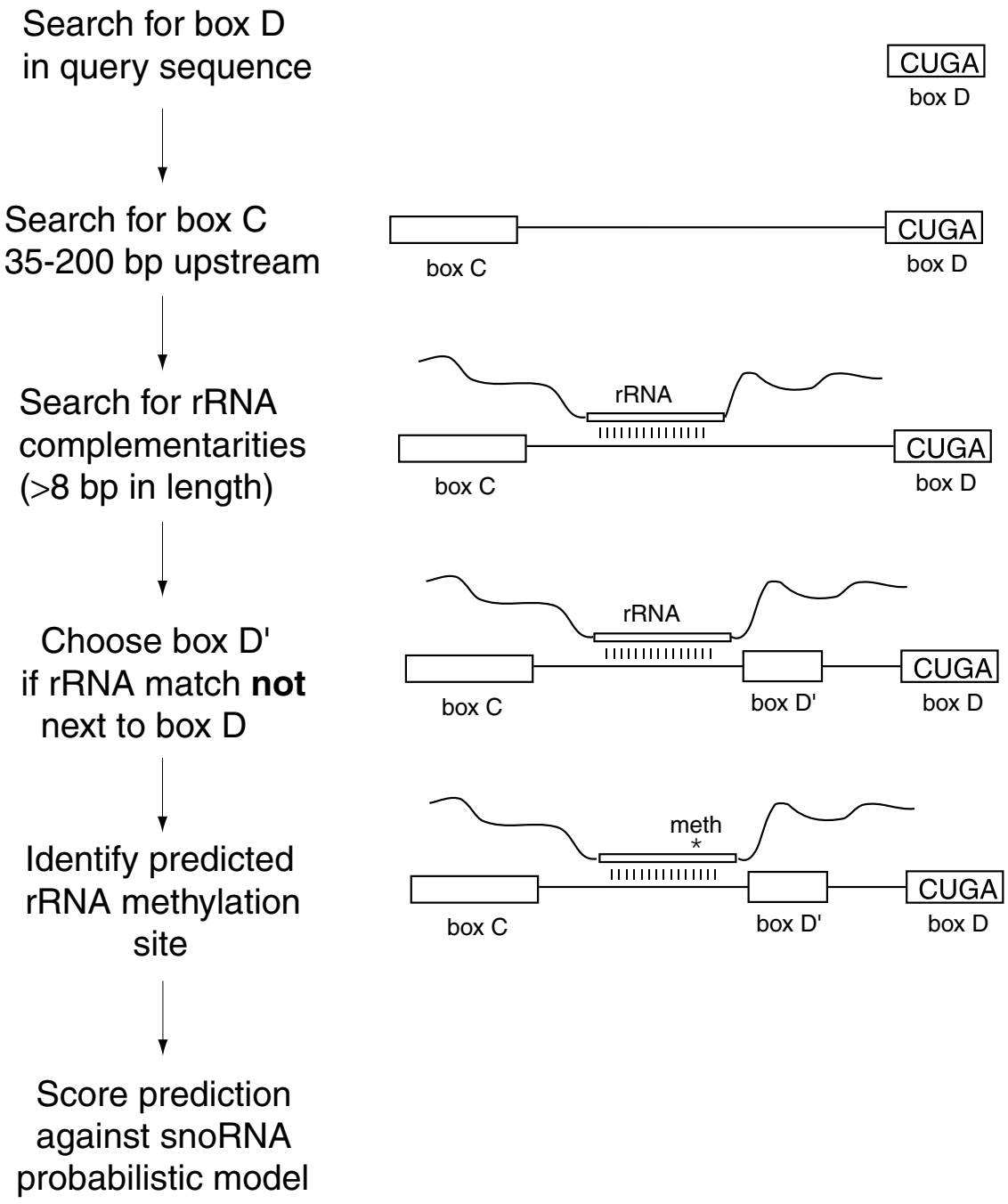

Figure 4.1: **Schematic diagram of snoRNA search algorithm.**

State number	Feature	Model	Consensus	Feature Score (bits)		
				Best	Average	Worst
1	Terminal Stem	SCFG, 4-8 bp	6 bp (when present)	7.60	3.09	0.35
2	Box C	7 bp ungapped HMM	AUGAUGA	12.73	11.63	5.84
3	Gap	Duration model	Length 6-10 bp	-1.59	-2.09	-4.76
4	Guide Sequence	HMM	12 bp duplex	15.67	11.11	2.54
5	Box D'	4 bp ungapped HMM	CUGA	7.34	4.85	-3.74
6	Gap	Duration model	Length 36-45 bp	-1.59	-2.43	-5.36
7	Box D	4 bp ungapped HMM	CUGA	8.05	7.92	5.43
8	Gap	Duration model	Length 56-75 bp	-1.50	-2.10	-4.17
9	Guide Sequence	HMM	14 bp duplex	18.96	13.98	9.95

Table 4.1: **Summary of states within snoRNA probabilistic model.**
State numbers correspond to Figure 4.2. "Ungapped HMM" states represent fixed-length conserved sequence motifs. The state for the terminal stem is analogous, but models base pairs rather than single positions (*e.g.,* a stochastic context-free grammar, SCFG (Durbin et al., 1998) instead of a hidden Markov model, HMM). Duration models for gaps are estimated from binned length distributions (e.g., the probability that a gap will be 11-20 nt, 21-30 nt, etc.). The guide state is a hidden Markov model dependent on the rRNA target sequence; it includes terms for the probability of starting the complementarity at a given position relative to rRNA (this probability is high near known methylation positions), the length of the complementarity, and the probability of mismatches and noncanonical base pairs in the complementarity. For each state, the most common feature ("consensus") is shown to indicate the overall pattern we search for. The best, average, and worst feature scores are given for 41 methylation guide snoRNAs as an indication of the relative contribution of each state to the overall information in the model. For more detail, see the program source code (Lowe & Eddy, 1998).

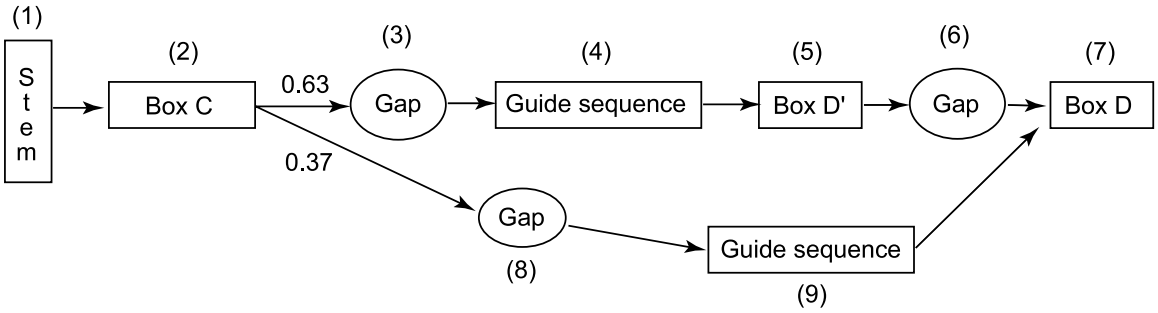

Figure 4.2: **Schematic of the probabilistic snoRNA model.**
States (boxes and ovals) are connected by transitions (arrows). Each numbered state is a probabilistic model of a sequence feature (Table 4.1). Transition probabilities are 1.0, except those shown for transitions 2→3 and 2→8, which account for the proportion of snoRNAs with a guide sequence adjacent to box D' and those with a guide sequence adjacent to box D, respectively.

Legend for Table 4.2. (next page)

Newly identified snoRNAs or methylation sites are in boldface. Previously identified snoRNAs newly determined as methylation guides are in italics. "Match/Mismatch" column refers to the number of base pairings (G-U included) and mismatches found within the snoRNA complementary region/rRNA duplex. "Len" refers to the known or predicted mature snoRNA length in nucleotides. "Position" and "Chr" refer to the 5' end and chromosomal genomic locus according to current version of the yeast genome available at the Saccharomyces cerevisiae Database (http://genome-www.stanford.edu/Saccharomyces/). Strand designations: (W) = Watson, upper/forward strand; (C) = Crick, lower/complement strand. rRNA positions are numbered as in (Maden, 1990). • = New data presented in this work; ⊗ = tentative assignment due to inconclusive assay for methyl loss; ⊕ = previously identified: (1) (Kiss-Laszlo et al., 1996), (2) (Nicoloso et al., 1996), (3) (Ni, 1998); ND = Not determined.

snoRNA	Target Methyl Site	Predicted Targ Site	Verified Targ Site	Match/ Mism	Genomic Placement	Len	Chr	Position
U14	SSU-Cm414	⊕ (1)	ND	13/1	cluster 5	123	10	139388 (C)
U18	LSU-Am647	⊕ (1)	⊕ (3)	15/0	intronic	102	1	142357 (W)
	LSU-Cm648		⊕ (3)					
U24	LSU-Cm1435	⊕ (1)	⊕ (1)	14/0	intronic	87	13	500071 (C)
	LSU-Am1447	⊕ (1,2)	⊕ (1)	12/0				
	LSU-Gm1448							
snR13	LSU-Am2278	•	•	10/0	extragenic	107	4	1393187 (W)
	LSU-Am2279		•					
snR38	**LSU-Gm2812**	⊕ (1)	ND	13/0	intronic	95	11	282830 (W)
snR39	LSU-Am805	⊕ (1,2)	⊕ (3)	13/0	intronic	89	7	365249 (C)
snR39b	LSU-Gm803	⊕ (2)	•	14/0	extragenic	95	7	366466 (C)
snR40	SSU-Gm1267	⊕ (1,2)	•	11/1	extragenic	97	14	89208 (W)
	LSU-Um896	•	•	12/0				
snR41	SSU-Am541	•	•	11/1	cluster 3	105	16	719237 (C)
	SSU-Gm1123	⊕ (1,2)	•	20/1				
snR47	SSU-Am619	•	•	11/0	extragenic	99	4	541738 (C)
	LSU-Am2218	⊕ (3)	•	12/0				
snR48	LSU-Gm2788	•	•	17/1	extragenic	112	7	609578 (W)
	LSU-Gm2790	•	•	15/0				
snR50 (Z14)	LSU-Gm865	•	•	12/0	extragenic	89	15	259489 (W)
snR51	SSU-Am100	•	•	16/1	cluster 3	107	16	718803 (C)
	LSU-Um2726	•	•	14/1				
snR52 (Z13)	SSU-Am420	•	•	13/0	extragenic	92	5	431217 (C)
	LSU-Um2918	•	No	11/1				
	LSU-Gm2919	•	No					
snR53	SSU-Am796	•	•	11/0	cluster 4	91	5	61699 (W)
snR54	SSU-Am973	•	•	13/0	intronic	86	13	163620 (C)
snR55 (Z10)	SSU-Um1265	•	⊗	12/0	cluster 2	98	12	794793 (C)
snR56	SSU-Gm1425	•	•	11/0	extragenic	86	2	88181 (W)
snR57	SSU-Gm1570	•	•	15/0	cluster 2	88	12	795023 (C)
snR58 (Z12)	**LSU-Cm661**	•	•	13/0	extragenic	96	15	136182 (C)
snR59	LSU-Am805	•	⊕ (3)	14/0	intronic	78	16	173826 (W)
snR60 (Z15)	LSU-Am815	•	•	10/0	extragenic	104	10	348929 (C)
	LSU-Gm906	•	•	19/0				
snR61 (Z11)	LSU-Am1131	•	•	11/0	cluster 2	90	12	794574 (C)
snR62	LSU-Um1886	•	•	14/0	extragenic	100	15	409863 (C)
snR63	LSU-Am2254	•	•	12/0	extragenic	255	4	323470 (C)
snR64	**LSU-Cm2335**	•	•	11/0	extragenic	101	11	38812 (W)
snR65	**LSU-Um2345**	•	•	11/0	extragenic	100	3	175909 (W)
snR66 (Z16)	LSU-Um2415	•	•	12/0	extragenic	86	14	586088 (W)
snR67	**LSU-Gm2616**	•	•	11/2	cluster 4	82	5	61352 (W)
	LSU-Um2721	•	•	11/0				
snR68	LSU-Am2637	•	•	12/0	extragenic	136	9	97111 (W)
snR69	LSU-Cm2945	•	•	18/3	extragenic	101	11	364418 (W)
snR70	SSU-Cm1637	•	⊗	9/1	cluster 3	164	16	719047 (C)
snR71	LSU-Am2943	•	•	9/1	extragenic	89	8	411228 (W)
snR72 (Z2)	**LSU-Am874**	•	•	14/0	cluster 1	91	13	298554 (W)
snR73 (Z3)	**LSU-Cm2956**	•	•	12/1	cluster 1	103	13	298306 (W)
snR74 (Z4)	SSU-Am28	•	•	13/0	cluster 1	80	13	298138 (W)
snR75 (Z5)	**LSU-Gm2286**	•	•	11/0	cluster 1	85	13	297915 (W)
snR76 (Z6)	LSU-Cm2195	•	•	13/1	cluster 1	101	13	297727 (W)
snR77 (Z7)	SSU-Um578	•	•	14/0	cluster 1	84	13	297506 (W)
snR78 (Z8)	**LSU-Um2419**	•	•	12/0	cluster 1	82	13	297277 (W)
snR79 (Z9)	SSU-Cm1006	•	•	16/1	extragenic	85	12	348511 (C)

Table 4.2: **C/D box snoRNAs in *S. cerevisiae* that function as methylation guides.** (see legend previous page)

Legend for Figures 4.3 to 4.6, showing experimental confirmation of methylation guide function for yeast snoRNAs.

AMV reverse transcriptase (RT) primer extensions were performed on total RNA from wildtype (wt) and snoRNA-disrupted (ΔsnR) strains to verify loss of target 2'-O-methyl groups in ribosomal RNA. RNA sequencing ladders of ribosomal RNA regions being assayed appear in lanes 1-4. Lanes 5 and greater contain pairs of RT primer extensions on the same RNA sample in which odd lanes use high dNTP concentration (1.0 mM) reactions and even lanes contain low dNTP concentration (0.004 mM) reactions. 2'-O-methyl modified nucleotides are characterized by appearance of termination bands in low but not high dNTP concentration reactions. Bands due to known 2'-O-methyl groups are labeled to the right of primer extension gels, and generally occur one nucleotide 3' to the nucleotide containing the methyl group. The gels in Figures 4.3 and 4.4 depict typical results for functional confirmations, and those in 4.5 and 4.6 represent the two most difficult primer extensions to interpret for loss of 2'-O-methyls, presumably due to other types of neighboring nucleotide modifications.

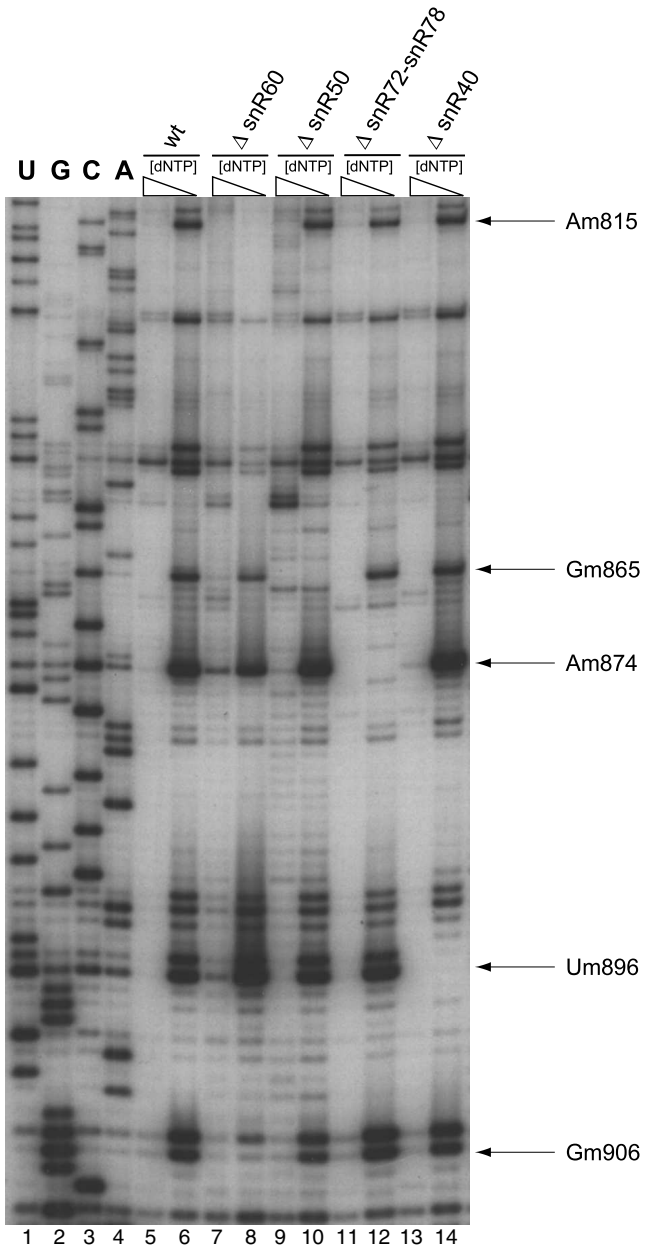

Figure 4.3: **Experimental confirmation of methylation guide function for snoRNAs snR60, snR50, snR72, and snR40.** Loss of 2'-O-methyl bands in low dNTP-concentration reactions for mutant strains (even lanes 8 and greater) relative to the wildtype strain (lane 6) indicates loss of the methylation site and thus functional confirmation. Polyacrylamide gel electrophoresis of primer extensions using ^{32}P end-labled primers annealing to 25S rRNA from position 914-939.

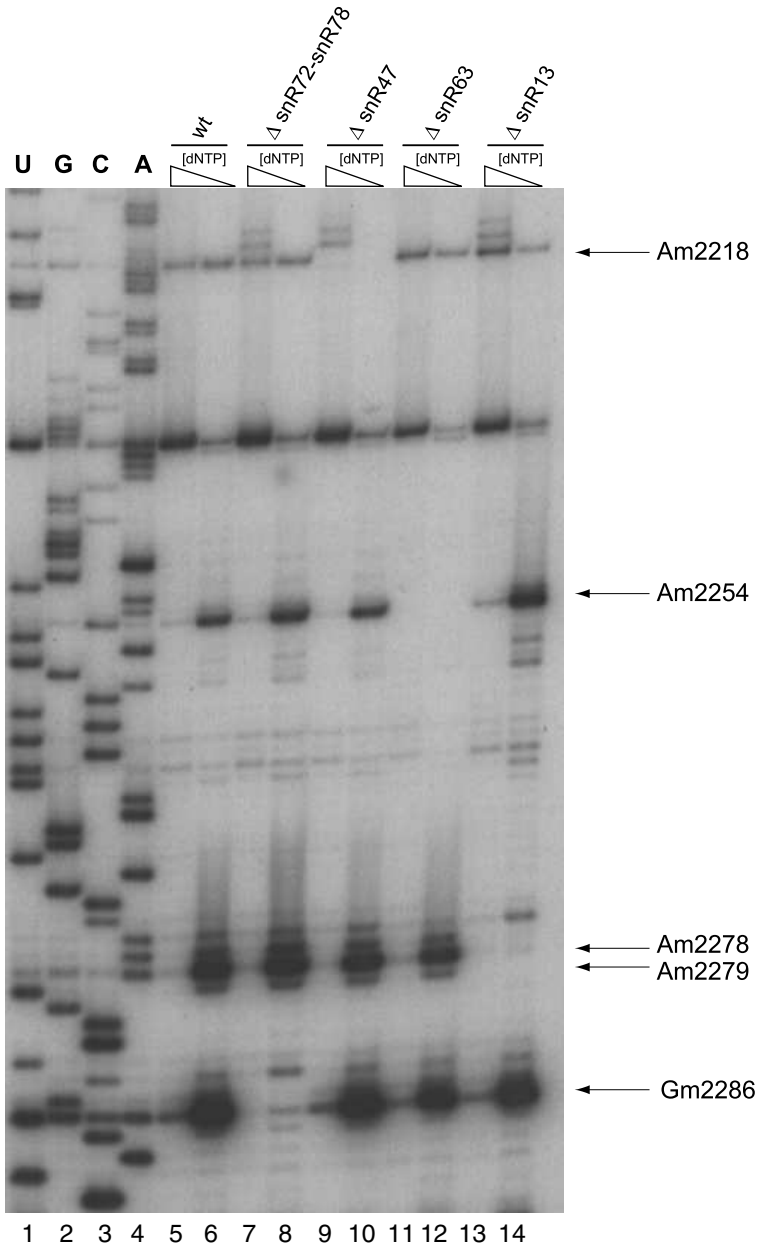

Figure 4.4: **Experimental confirmation of methylation guide function for snoRNAs snR75, snR47, snR63, and snR13.** Loss of 2'-O-methyl bands in low dNTP-concentration reactions for mutant strains (even lanes 8 and greater) relative to the wildtype strain (lane 6) indicates loss of the methylation site and thus functional confirmation. Polyacrylamide gel electrophoresis of primer extensions using ^{32}P end-labled primers annealing to 25S rRNA from position 2305-2328.

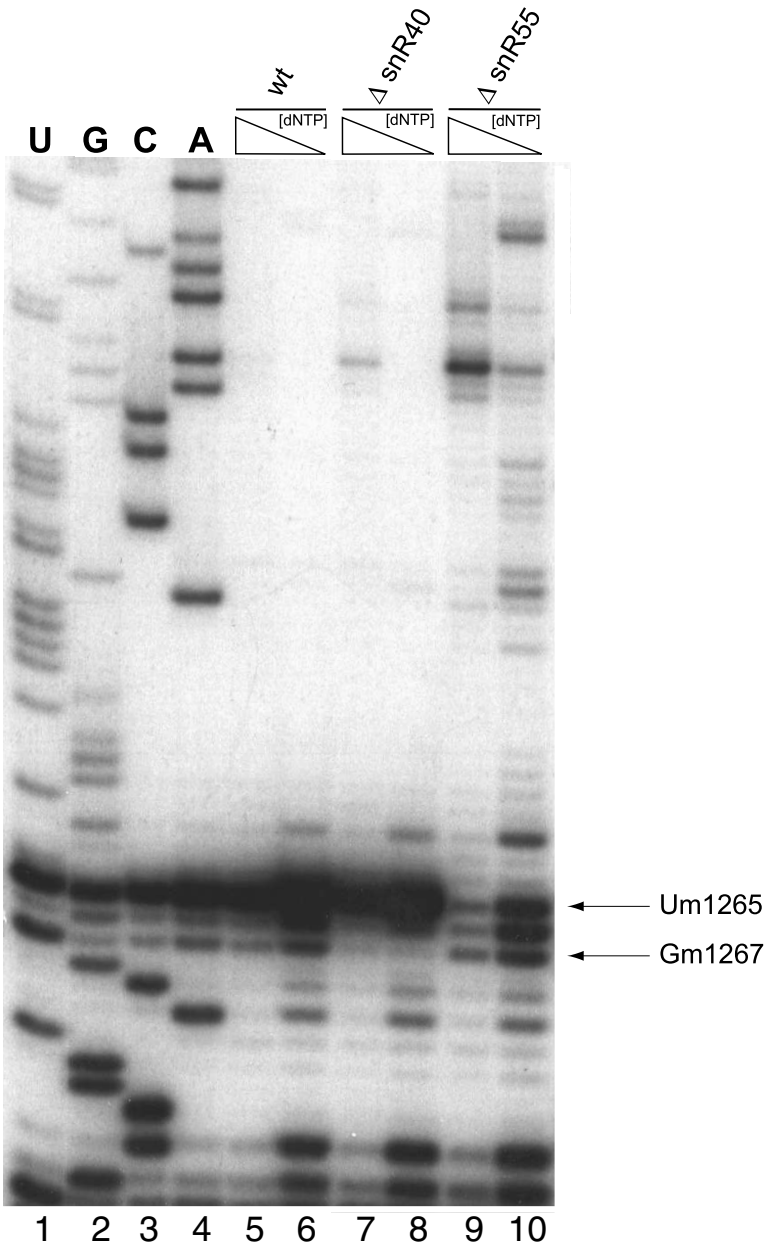

Figure 4.5: **Experimental confirmation of methylation guide function for snoRNAs snR40 and snR55** Loss of 2'-O-methyl band in low dNTP-concentration reaction for mutant strain (lanes 8 & 10) relative to the wildtype strain (lane 6) indicates loss of the methylation site and thus functional confirmation. Polyacrylamide gel electrophoresis of primer extensions using ^{32}P end-labled primers annealing to 18S rRNA from position 1291-1315.

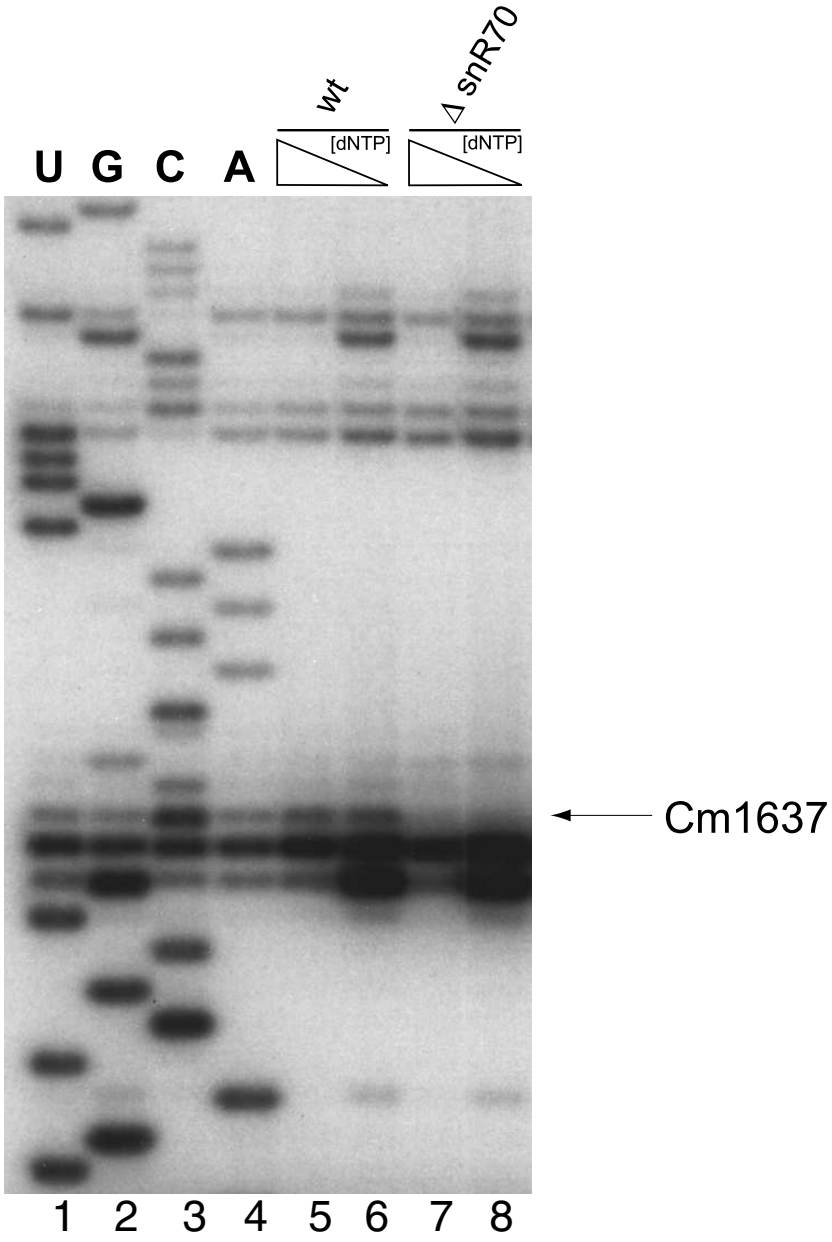

Figure 4.6: **Experimental confirmation of methylation guide function for snoRNA snR70.** Loss of 2'-O-methyl bands in low dNTP-concentration reactions for mutant strains (lane 8) relative to the wildtype strain (lane 6) indicates loss of the methylation site and thus functional confirmation. Polyacrylamide gel electrophoresis of primer extensions using ^{32}P end-labled primers annealing to 18S rRNA from position 1652-1675.

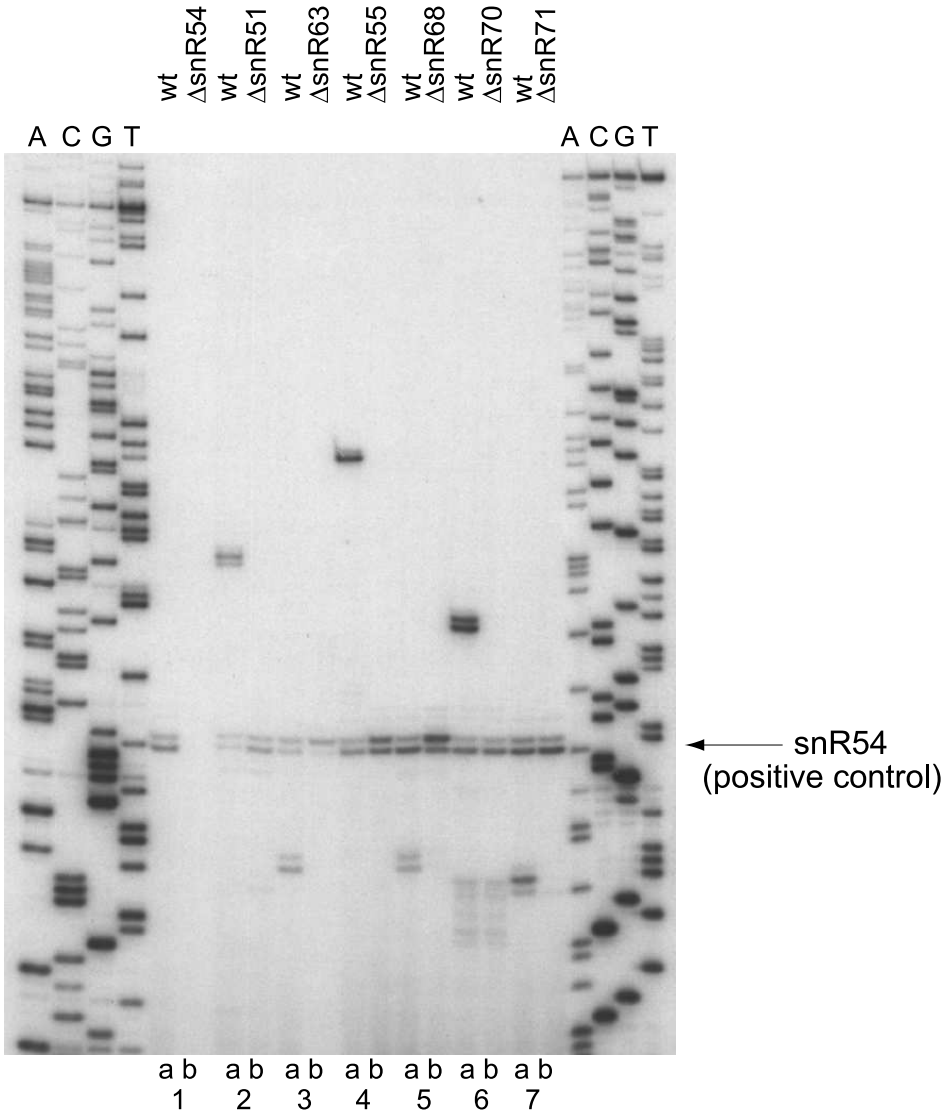

Figure 4.7: **snoRNA primer extensions demonstrating expression of newly identified methylation guide snoRNAs.**
Reverse transcriptase primer extensions on total RNA from wildtype and snoRNA-disrupted strains. ^{32}P end-labeled primers complementary to internal snoRNA sequence for snR54 (lanes 1a,b), snR51 (lanes 2a,b), snR63 (lanes 3a,b), snR55 (lanes 4a,b), snR68 (lanes 5a,b), snR70 (lanes 6a,b), snR71 (lanes 7a,b) were used. snoRNA expression in wildtype RNA reactions (lanes 1a, 2a,...7a) was confirmed, as was loss of expression in snoRNA-deleted strains (lanes 1b, 2b,...7b). The snR54 internal snoRNA primer was included in all reactions as a positive control of intact RNA and active primer extension. RNA sequencing ladders of unrelated sequence are included on either side of snoRNA primer extensions for fragment size reference.

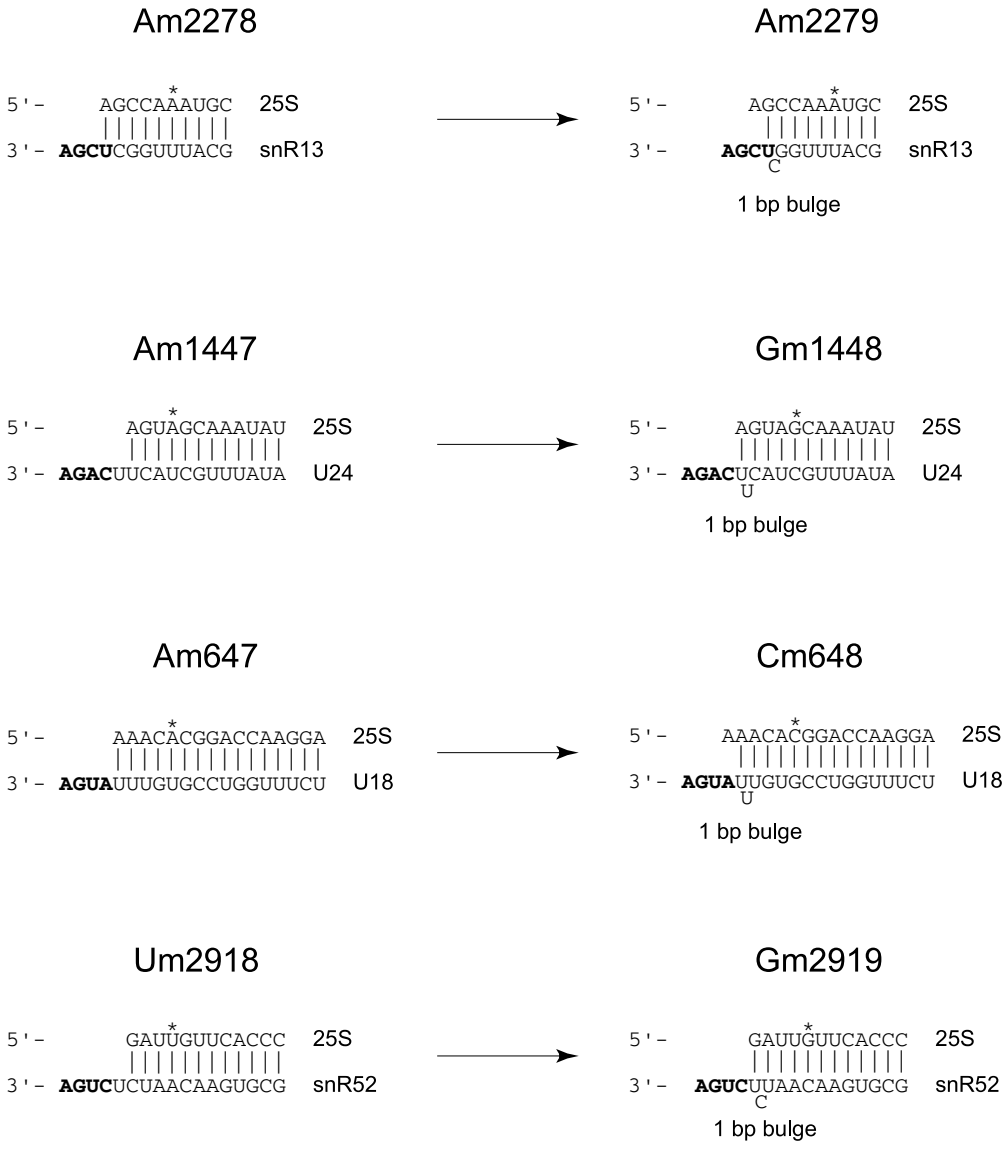

Figure 4.8: **Model for Addition of Adjacent 2'-O-methyls via the same snoRNA.** Listed are the four instances in yeast rRNA in which 2'-O-methyl groups occur just one nucleotide from other 2'-O-methyls. On the left hand side, base pairings between yeast rRNA and functionally confirmed (snR13, U24) or predicted (U18, snR52) methyl guide snoRNAs are depicted. Spacing between the D or D' box and rRNA sequence in each case presumably determines the location of 2'-O-methyl modification, invariably 5 bp from the end of the D/D' box (Kiss-Laszlo et al., 1996)). If the D' box is allowed to slide one nucleotide closer in via a single nucleotide bulge, the new placement of the D' box could conceivably guide addition of a second 2'-O-methyl group at the adjacent position. In each case, the single nucleotide bulge in the snoRNA would not necessarily disrupt required base pairings within the snoRNA/rRNA duplex.

Chapter 5

Small Nucleolar RNAs in Archaeal Genomes[1]

[1]This chapter was co-written with Patrick Dennis, and will be submitted for publication as a collaboration between the Eddy and Dennis labs.

5.1 Abstract

Eukaryotic ribosomal RNA (rRNA) contains dozens of post-transcriptionally modified nucleotides. The most numerous type of modification, 2'-O-ribose methylation, requires a family of small nucleolar RNA genes (snoRNAs) which specify the position of methylation by direct base pairing interactions. SnoRNAs have not been reported in Bacteria or Archaea. Using biochemical and computational methods, we have identified archaeal snoRNA genes in species covering both major branches of the Archaea. Eighteen small sno-like RNAs (sRNAs) were cloned from the archaeon *Sulfolobus acidocaldarius* by co-immunoprecipitation with aFIB and aNOP56, the archaeal homologs of eukaryotic snoRNA-associated proteins. From the properties of these sRNAs, we trained a probabilistic model to search for archaeal sRNAs in archaeal genomic sequences. Over 200 additional sRNAs were found across five divergent archaeal genera from seven genomes. Many of these are confirmed experimentally or supported by comparative sequence analysis.

5.2 Introduction

Ribosome biogenesis in eukarya occurs in the nucleolar compartment of the nucleus. Several proteins, including fibrillarin, Nop56 and Nop58, and dozens of snoRNAs are involved in this process (Maxwell & Fournier, 1995; Balakin et al., 1996; Tollervey et al., 1991; Gautier et al., 1997). The snoRNAs can be divided into two major classes: C/D box and H/ACA box RNAs. The C/D box snoRNAs are efficiently precipitated with antibodies against fibrillarin. Most C/D box snoRNAs are involved in targeting ribose methylation within rRNA, whereas most H/ACA box RNAs are involved in targeting the conversion of uridine to pseudouridine within rRNA (Balakin et al., 1996; Cavaille et al., 1996; Kiss-Laszlo et al., 1996; Ni et al., 1997; Gannot et al., 1997). The small number of snoRNAs not involved in nucleotide modification are required for proper endonucleolytic processing of the pre-rRNA (reviewed in (Maden & Hughes, 1997)).

The general mechanism whereby C/D box snoRNAs target ribose methylation is well

established. Each snoRNA contains a unique 9 to 20 nucleotide (nt) sequence located 5' to the D or D' box motif that is complementary to a sequence within small subunit (SSU) or large subunit (LSU) rRNA (see Figure 1.1). During ribosome biogenesis, a snoRNA:rRNA helix is formed and methylation is directed to the rRNA nucleotide that participates in the base pair 5 nt upstream from the start of the D or D' box.

In eukaryotes, it is likely that most, if not all, rRNA ribose methyl modifications are guided by snoRNAs. In the yeast *S. cerevisiae*, methylation guide snoRNAs have been assigned to all but four of the 55 rRNA ribose methylation sites (Lowe & Eddy, 1999). Although no single methylation site and no individual C/D box snoRNA involved only in methylation appear to be essential, global rRNA methylation in yeast is apparently essential. Inhibition of methylation is believed to severely compromise the ability of the rRNA to fold into or maintain the active higher order structure (Tollervey et al., 1993; Maden & Hughes, 1997).

SnoRNAs have only been found in eukaryotic species. Ribose methyl modification levels in bacterial rRNA are much lower. *Escherichia coli* rRNA contains only four ribose methyls and these are anticipated to be modified by individual protein enzymes (Lafontaine & Tollervey, 1998). In contrast, the rRNA of the archaeon *Sulfolobus solfataricus* has been shown to contain 67 ribose methylation sites (Noon et al., 1998), a number similar to that found in eukaryotes. Even though Archaea are unicellular prokaryotic organisms that lack a nucleolus (Woese et al., 1990), their genomes encode homologs to the essential eukaryotic nucleolar proteins, fibrillarin and NOP56/58 (Amiri, 1994; Lafontaine & Tollervey, 1998). Based on these observations, we decided to examine Archaea for the presence of sno-like RNAs using both experimental and computational methods.

5.3 *S. acidocaldarius* has aFIB, aNOP and C/D box sRNAs

To isolate biochemically sno-like RNAs from the archaeon *Sulfolobus acidocaldarius*, members of the Dennis lab[2] first cloned the archaeal homologs to the eukaryotic fibrillarin (aFIB) and NOP56 (aNOP) proteins using sequence information from a related species, *Sulfolobus solfataricus*[3]. The cloned genes were expressed in *E. coli* and the recombinant proteins were purified and used to raise polyclonal antibodies in rabbits. The two antibody preparations were each highly specific and recognize single polypeptides of the predicted size in total *S. acidocaldarius* cell extracts (data not shown). Immunoprecipitates formed with crude cell lysate contained a large amount of non-specific RNA. To eliminate this contamination, an ammonium sulfate-glycerol gradient fractionation procedure was introduced. Following the sedimentation step, the antibodies were used first to monitor the size distribution of aFIB and aNOP56 in the gradient fractions by Western blotting (Chamberlain et al., 1998) (Figure 5.1a). Both aFIB and aNOP56 sedimented as a large heterogeneous complex ranging from about 4S to greater than 50S in size; the larger complexes appeared to be enriched for aNOP56 relative to aFIB.

To detect RNAs that associate with aFIB- and aNOP-containing complexes, aliquots from gradient fractions were immunoprecipitated with either anti-aFIB or anti-aNOP56 antibodies. Following immunoprecipitation, total RNA was extracted with phenol from the supernatants and the pellets, and a portion from each was end-labeled with pCp and

[2]The Dennis lab at University of British Columbia, Department of Biochemistry and Molecular Biology, includes Patrick Dennis, Arina Omer, Anthony Russell, and Holger Ebhart. As collaborators on this project, the Dennis lab performed all *S. acidocaldarius* protein biochemistry, immunoprecipitation, and sRNA cloning.

[3]A clone containing the aFIB and aNOP56 genes from *S. solfataricus* was provided by M.A. Ragan and C.W. Sensen. The Dennis lab used Southern hybridization to identify and clone the corresponding genes from *S. acidocaldarius*. The 16S rRNA sequences from *S. acidocaldarius* and *S. solfataricus* are about 90% identical. The aFIB and aNOP proteins from the two organisms are respectively 76 and 66 percent identical. The accession numbers for the *S. acidocaldarius* aFIB and aNOP genes will be obtained upon submission of the manuscript describing this work

displayed by denaturing PAGE (Figures 5.1b and c). The most abundant RNAs that were co-immunoprecipitated appear as a family of discrete bands ranging in length from about 50-70 nt. This size class of RNAs, which is substantially shorter than eukaryotic C/D box snoRNAs, was invisible when total cellular RNA was labeled with pCp. To obtain cDNA clones, the RNA precipitated from fraction 5 with anti-aFIB and from fractions 6-8 and 10-13 with anti-aNOP56 were gel purified, ligated to the oligonucleotide oAO30, and used as template for RT-PCR[4] (Wu et al., 1998). The PCR products were cloned between the PstI and XhoI sites of pSP72 plasmid.

A total of about 50 clones from each of the two immunoprecipitated RNA pools were sequenced by the Dennis lab. About half had inserts containing random fragments of 16S and 23S rRNA sequence; the other half gave one or more representatives of 18 different sequences which exhibited features characteristic of eukaryotic C/D box snoRNAs (Table 5.4) (Balakin et al., 1996; Kiss-Laszlo et al., 1996; Kiss-Laszlo et al., 1998). Three of the clones, Sac sR5, sR14, and sR18, were independently recovered from the two separate immunoprecipitations. This was expected since anti-aFIB coprecipitates aNOP and anti-aNOP coprecipitates aFIB from crude cell extracts (data not shown). All 18 clones contained well-defined and highly conserved C(AUGAUGA) and D(CUGA) box motifs located respectively near their 5' and 3' ends. Moreover, all contained recognizable internal C' and D' box motifs, giving the RNAs a dyad repeat structure characteristic of eukaryotic

[4]The primer AO30 (5' CTCGAGATCTGGATCCGGG 3') was 5' end-labeled with T4 polynucleotide kinase and $\gamma-^{32}$P-ATP and blocked at the 3' end using terminal deoxynucleotidyl transferase (Gibco BRL) and dCTP. The modified oligo was then ligated to gel purified sRNA for 16 hrs at 4 $^{\circ}$C. The ligation products were reverse transcribed with Thermoscript RT (Gibco BRL) at 55 $^{\circ}$C for 30 min, using AO31 (5' CCCGGATCCAGATCTCGAG 3') as primer. The RNA template was hydrolyzed with RNase H and the cDNA strand was extended with dATP using terminal deoxynucleotidyl transferase. The extended cDNA strand was used as template for PCR (95 $^{\circ}$C denaturation, 65 $^{\circ}$C hybridization, 72 $^{\circ}$C extension, 30 cycles) using AO30 and AO32 [5' GCGAATTCTGCAG$(T)_{30}$ 3'] as primers. The DNA products were cleaved with PstI and XhoI, ligated between the PstI and XhoI sites of plasmid pSP72 and transformed into *E. coli*. Plasmids were isolated and their sequences was determined.

methylation guide snoRNAs (Kiss-Laszlo et al., 1998).

Both the Dennis lab and I used primer extension analysis to confirm the presence of the sRNAs in total RNA extracted from *S. acidocaldarius* (Figures 5.2a and b; gels pictured are from Dennis lab). I designed the *S. acidocaldarius* sRNA primers to overlap the common D box motif and extend through the unique guide region and into the C' box motif of sR1 to sR17. I obtained extension products for 15 of 17 sRNA-specific primers. The lengths of the products were within one or two nucleotides of the cloned 5' ends for all but two sRNAs; sR6 and sR8's products were both 4 nt longer than the cloned ends. In separate experiments in the Dennis lab, Southern hybridizations have confirmed the existence of sR1, sR2, sR5, and sR13 encoding single-copy sequences within *S. acidocaldarius* genomic DNA; the four encoding sequences are not closely linked (data not shown). These data demonstrate that *Sulfolobus acidocaldarius* contains snoRNA-like C/D box sRNAs.

5.4 Methylation sites in Ribosomal RNA

To determine if these sRNAs might function as guides for ribose methylation in ribosomal RNA in a manner similar to eukaryotic C/D box snoRNAs, the sRNAs were examined for potential guide sequences. Regions complementary to rRNA and adjacent to the D or D' boxes were identified for more than half of the sRNAs. Using the D/D' box plus 5 nt rule, we predicted the locations of potential ribose methyl modifications in rRNA and experimentally tested for these using the dNTP concentration-dependent primer extension assay (Maden et al., 1995; Kiss-Laszlo et al., 1996). In this assay, characteristic ribose 2'-O-methyl pauses are displayed in the reverse transcriptase reactions at low- but not at high-dNTP concentrations. Using both *S. acidocaldarius* and *S. solfataricus* total RNAs as template, we were able to identify pause sites at eight predicted methylation sites (Table 5.1). Examples of four such pause sites are shown in Figure 5.3. Gene disruption systems for *S. acidocaldarius* and most other archaea are currently not available; consequently we were not able to demonstrate the loss of predicted rRNA methylation sites upon disruption

of sRNA genes.

5.5 Identification of an *S. solfataricus*sRNA Homolog

To find additional homologs of our cloned *S. acidocaldarius* sRNA genes, we initially ran BLAST (Altschul et al., 1997) on each cDNA clone against the non-redundant nucleotide database, which included four completely sequenced archaeal genomes (*Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidis* and *Pyrococcus horikoshii*). From our previous experience with eukaryotic snoRNAs, we knew generalized similarity search methods like BLAST were usually only effective at finding homologs in closely related species. Although we did not receive any hits to the distantly related complete archaeal genomes, we did recover two weak hits against sequences in other *Sulfolobus* species: Sac sR03 had a hit near the *Sulfolobus shibatae* top6B topoisomerase II gene (opposite strand of 3' UTR, BLAST score 40.1 bits, expect value 0.038), and Sac sR01 had a hit near the *Sulfolobus solfataricus* aspartate aminotransferase gene (3' UTR; BLAST score 38.2 bits, expect value 0.15). Normally we would not have considered these weak hits to be legitimate (both against only the 3' half of the respective sRNAs); however, manual examination of the upstream sequences revealed proper C boxes. Intriguingly, the hit against Sac sR01, when extended upstream, overlapped by 21 nucleotides into the C-terminal end of the aminotransferase coding region. Because the aspartate aminotransferase protein has clear homologs in several other archaeal genomes (with similarity extending up to but not including sRNA overlap), we believe the protein is real with an accurately predicted C-terminal end. Overlap of a snoRNA gene and a protein coding region is unprecedented, and because no other hits were found for the other sRNA clones, it was unclear whether these were true sRNA homologs.

Because of the uncertainty, I carried out primer extension analysis to test for the presence of the Sac sR1 homolog, dubbed Sso sR1, using total *S. solfataricus* RNA as template. A product with a length similar to that of Sac sR01 was detected (Figure 5.2c; gel pic-

tured is from a reproduced experiment from the Dennis lab). Moreover, using the dNTP concentration-dependent reaction, both the Dennis lab and I were able to verify that the predicted U52 position in *S. solfataricus* small subunit (SSU) rRNA is likely to carry a methyl modification (Figure 5.3a). The presence of functionally related sRNAs in two distinct species that apparently guide methylation to the same U52 position in 16S rRNA provides additional support for the existence of eukaryotic-like C/D snoRNAs in the Archaea.

5.6 A Computational Screen for Additional Archaeal sRNAs

An alternate method was needed to find sno-like RNAs in the other sequenced archaeal genomes. Because I have had previous success with a specialized snoRNA gene finding program (Lowe & Eddy, 1999), I decided to tailor a search for archaeal snoRNAs. The original program used a probabilistic model trained on known human and yeast snoRNAs. With the new set of verified *S. acidocaldarius* sRNA genes cloned by the Dennis lab, I retrained the program for archaeal sRNAs. The search algorithm and general model remained as originally described (Lowe & Eddy, 1999). Alignments of the box features (C, D, C', D') of the *S. acidocaldarius* sRNAs were used to create log odds weight matrices reflecting the frequency of each nucleotide at each position in each box feature. The lengths of the rRNA complementary region and the gaps between box features were scored with binned length distributions. Overall, training data for the nucleotide content of the box features did not change significantly, but the distribution of lengths between features did vary; archaeal sRNAs appear to be much more compact than those in eukaryotes, and the rRNA complementary regions are shorter (commonly 8-11 nt long, compared to 12-14 in *S. cerevisiae*).

I started the sRNA genome searches in *Sulfolobus solfataricus*, for which approximately half the genome sequence was available. The program identified many dozens of sRNA candidates, each of which had the potential to target a modification to a particular position in the ribosomal RNA of *S. solfataricus*. Because I had very little *a priori* knowledge of

verified ribose methylation sites in *S. solfataricus* rRNA, I sorted all candidates by overall score, regardless of the target rRNA methylation site. I designed primers against the top twenty sRNA candidates, and performed primer extensions on total *S. solfataricus* RNA to identify sRNA transcripts of the correct length. Based on cloned sRNAs, I assumed that new sRNAs should have a 5' end 2-6 nt upstream from the predicted C box. Ten of the top 13 scoring candidates produced primer extension products of the approximate size (data not shown). An alignment of the 10 verified *S. solfataricus* sRNAs and 3 other predictions is shown in Figure 5.4 (below *S. acidocaldarius* clones for comparison). For seven sRNA candidates, we also attempted to verify a predicted target ribose methylation site, again using the dNTP concentration-dependent primer extension assay. Sites for four sRNAs were verified (see Table 5.2). Because we have evidence for rRNA methylation sites corresponding to a number of verified or predicted sRNAs, we believe that as in eukaryotes, C/D box sRNAs function as a guides for methylation.

5.7 sRNAs in Both Main Branches of the Archaea

With the establishment of sRNAs in two *Sulfolobus* species, we next asked how ubiquitous this class of RNAs might be among the Archaea. Fortunately, genome sequence is available from species covering a wide range of phyla, including members from both main divisions of the Archaea, the Crenarchaea (*Sulfolobus solfataricus*, *Aeropyrum pernix*), and the Euryarchaea (*Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidis*, *Pyrococcus horikoshii*, *Pyrococcus abyssi*, *Pyrococcus furiosus*). Evidence of methylation guide sRNAs in any of the Euryarchaeal species would imply that this feature originated before the split between Archaea and Eukarya. In searching these genomes for guide sRNAs, I found strong candidates in all but one of these seven species (see sRNA sequence alignments in Figures 5.5, 5.6, 5.7, & 5.8). No strong candidates were found in the genome of *M. thermoautotrophicum*.

The search of the *M. jannaschii* genome gave eight strong hits. Strikingly, all eight

candidates contain precise canonical box features for the C (ATGATGA), D' (CTGA), C' (TGATGA), and D boxes (CTGA) (Figure 5.5). All contain 5-9 bp long terminal stems, and each has one or two rRNA complementary guide sequences of 9-13 nt. Again, as seen for the *Sulfolobus* guide RNAs, these RNAs are extremely compact, with very small gaps (1-7 nt) between box features and guide regions. I plan to perform primer extensions with sRNA-specific primers and *M. jannaschii* total RNA (kindly provided by James Brown) to confirm these predictions. I also plan to assay 4-6 sites of predicted methylation by the dNTP concentration-dependent primer extension assay. However, based on the strong feature conservation observed in the candidates, we believe the *M. jannaschii* C/D box sRNAs are also involved in guiding methylation of ribosomal RNA.

Based on high overall score and conserved sequence characteristics, I identified with high confidence four sRNA candidates from *A. fulgidis* (Figure 5.5). Two predictions had 4-5 bp terminal stems. Numerous other candidates that had one or more imperfect features were found, but in the absence of ribose methylation information for *A. fulgidis* rRNA, their authenticity remains suspect.

Aeropyrum pernix, which was the only Crenarchaea for which I had a complete genomic sequence to search, produced 29 candidate sRNAs (25 of these are shown in Figure 5.6). The *A. pernix* sRNA candidates show relatively relaxed sequence feature conservation and spacing, more similar to the crenarchaeal *Sulfolobus* sRNAs identified (Figure 5.4) than the euryarchaeal sRNAs (Figures 5.5, 5.7, & 5.8). This may be a general trend that is better supported as sRNAs from other archaeal species are identified.

5.8 *Pyrococcal* sRNA families

The genomes of three closely related *Pyrococcal* species (*P. horikoshii*, *P. furiosus*, and *P. abyssi*) have been sequenced. This allows powerful comparative analysis of genes that are recognizably homologous because they are recently diverged. However, the evolutionary distance between these particular species is great enough that syntenic DNA without selective

pressure is not conserved due to mutational drift. For our purposes, alignments between highly similar sequences that do not code for proteins generally indicate RNA genes, regulatory elements, or other biologically important sequences. We used this information to support our *Pyrococcus* sRNA predictions without experimentation.

The searches of the *P. horikoshii*, *P. furiosus*, and *P. abyssi* genomes identified 50, 52 and 56 putative C/D box sRNAs, respectively. The complete set of *P. horikoshii* sequences are presented in Figure 5.7. Most of the D and D' guide sequences in the 50 different *P. horikoshii* sRNAs exhibit at least one extended complementarity to a known stable RNA. Because of the close relationship between the three species, it was easy to associate the sRNAs by sequence similarity into 57 homologous groups. Alignments of the first ten homolog groups appear in Figure 5.8. Members of the same group ranged from 80 to 98 percent identity in end-to-end sequence alignments. Forty-six groups were found in all three species, 11 were found in only two species, and two were unique to a single species. Of the 50 *P. horikoshii* sRNAs, only Pho sR33 is unique; the other 49 have at least one homolog in one of the other *Pyrococcal* species. For each *Pyrococcal* homolog group, sequence similarity extends over both the D' and D box guides. Except for a few notable instances, the respective D' and D box guides appear to target homologous RNA methylation sites across the three species. In the cases of Pfu sR5 and Pfu sR7, a single base insertion upstream of the D' box appears to have slightly shifted the target methylation site by one nucleotide, relative to the sRNAs for the other two species (Figure 5.8). In three cases, complementary regions have changed by three or more nucleotides relative to the other two homologous sRNAs, likely changing the region or molecule targeted for methylation. Observations of this type support the view that methylation target selection is an ongoing, dynamic process.

Although we have identified regions of sequence complementarity between most of the *Pyrococcal* sRNA guides and stable RNAs, we cannot be sure which of the matches are functionally significant and which are fortuitous. We suspect that most may be functional since in almost all cases, the guide target complementarities are conserved within the *Pyrococcal*

sRNA families (and in some instances with other genera of archaea).

5.9 *Pyrococcus* sRNA Genome Distribution

With the large collection of sRNA genes identified within the *Pyrococcal* genomes, we were able to make a reasonable assessment of the their locations in the three species relative to protein coding genes and to each other. The sRNA genes are in general dispersed throughout the genome, and in only four cases are two genes located near each other. Three of the pairs are found near each other in all three genomes: sR14-sR22, sR2-sR9, sR12-sR34. All are on opposite strands and oriented away from each other. The distances between pairs ranged from a single nucleotide (Pho sR12-sR34) to 130 nt (Pfu sR2-sR9). The fourth pair, sR50-sR54, found only in *P. furiosus*, is oriented on the same strand and separated by 34 nucleotides. This is the only possible candidate for a polycistronic sRNA transcript.

Examination of the positions of sRNA loci relative to protein coding regions resulted in a unique finding: some (20-35%) sRNA genes appear to overlap partially with either the 5' or the 3' ends of open reading frames (ORFs) on the coding strand. Of the 17 overlaps in the *P. horikoshii* genome, eight occur at the 5' ends of protein ORFs. Based on BLAST results, all of these are likely to be artifacts resulting from incorrect assignment of translation initiation codons. In the nine cases where the overlap occurs at the 3' end of the protein ORFs, the overlaps appear to be valid. In most of these cases, the translation stop codons are provided by either the C or the D' box of the overlapping sRNA. A few sRNAs appear to partially overlap coding regions on the opposite strand, but we found no cases of sRNAs completely within predicted protein ORFs. Almost all sRNAs that do not overlap protein coding regions were located very near ORF boundaries (5-20 nt) and are probably too near to have their own promoters. Thus, they may be co-transcribed with upstream protein encoding genes and processed out of polycistronic transcripts.

We observed only one case where an sRNA was encoded completely within another gene: the *Pyrococcal* sR40 family resides as an intron in the anticodon loop of the gene encoding the

tRNA-Trp. This intron, which exhibits all of the hallmark features of an Archaeal sRNA, has been independently identified by Daniels and coworkers (personal communication). They present evidence that the D' and D box guides target methylation to positions C42 and C37 within the intron-containing precursor tRNA. We recovered this sRNA in our search because the respective guides appear to be capable of targeting methylation to C1252 in 16S rRNA and C1171 in 23S rRNA. Neither the tRNA nor rRNA target predictions have been experimentally verified.

5.10 Conserved Sites of Methylation in 16S and 23S rRNA

The identification of C/D box sRNAs from a wide spectrum of Archaeal genera allowed us to search for guide sequences capable of directing methylation to homologous sites within the respective 16S and 23S rRNAs. Based on CLUSTAL (Higgins et al., 1992) alignments of the rRNA sequences, a total of 19 sites of conserved methylation were identified; of these, 14 were shared between two genera and five were shared between three genera. In nearly all of the 19 cases, the sequence similarity between sRNAs that direct methylation to a homologous site is limited to only the guide region that targets the methylation. Moreover, the directing guides can be either both in the same position (*i.e.,* both D' or both D box associated) or in different positions (*i.e.,* one D' and the other D box associated). In only one instance have we detected strong end-to-end sequence similarity between two sRNAs from different Archaeal genera: Pho sR39 and Mja sR06.

Based on these and other data we cannot tell if guides that direct methylation to homologous sites in rRNA are related to each other by common ancestry (*i.e.,* homology) or by sequence convergence. If the relationship is by homology, it implies that guide and target sequences can co-evolve over long periods of evolutionary time if the selection for the methylation is sufficiently important (*i.e.,* for the folding and stabilization of rRNA structure). Alternatively, the rapid accumulation of substitutions in guide sequences (relative to substitutions in rRNA targets) as seen in numerous *Pyrococcal* families, coupled

with weak selection for methylation at a particular position, might suggest that many of the conserved sites for methylation between genera arose through convergence. Once a particular guide sequence lost its target specificity through the accumulation of nucleotide substitutions, it would become free to "explore sequence space" in order to identify more favorable interactions with rRNA or other types of RNA. We have also noted eight examples where sRNAs from different genera appear to direct methylation to nearby but not precisely identical rRNA sites. In these cases, it is possible that base pairing between the sRNA and rRNA during ribosome biogenesis (*i.e.,* chaperone function) may be more important than the precise positioning of a methyl group within the mature rRNA.

5.11 Evolutionary Origin and Divergence of C/D box sRNAs

The nucleolar compartment of the nucleus of eukaryotic cells contains the abundant protein fibrillarin and a large collection of fibrillarin associated C/D box snoRNAs. Most all of these RNAs function by employing a common helical ruler mechanism to direct 2'-O-ribose methylation to specific sites in 18S or 28S rRNA. The methylations may be important determinants in the folding and the stabilization of the structure of the rRNAs and their assembly into mature ribosomal subunits. Direct cloning of RNAs associated with the archaeal homolog of fibrillarin and computer searches of archaeal genome sequences have revealed the presence of C/D box sno-like RNAs in at least six widely divergent genera from both the crenarchaeote and euryarchaeote branches of the archaeal domain.

Analysis of archaeal C/D box sRNAs relative to eukaryotic snoRNAs reveals interesting structure-function features. The archaeal sRNAs are small, generally 50-60 nt in length, whereas human and yeast methylation guide snoRNAs average roughly 75 and 100 nt, respectively. A much larger proportion of archaeal sRNAs appear to have the ability to guide methylation from both D and D' boxes as "double guides", particularly the *Pyrococcal* sRNAs. Based on program predictions and comparative sequence analysis among *Pyrococcal* families, we estimate at least 50% of archaeal sRNAs guide methylation using both 5' and

3' guide regions, whereas only 20% of human and yeast snoRNAs have been reported to be double-guides (Kiss-Laszlo et al., 1996; Lowe & Eddy, 1999). Often, double guides target sites that are within the same target RNA (*i.e.,* 16S or 23S rRNA) and often they are closely linked within the target RNA. For example, Sac sR12 appears to direct methylation using D' and D box guides to positions G1114 and A1134 in 23S rRNA. This is in contrast to yeast snoRNA double guides, in which there is no apparent correlation between molecules targeted by the same snoRNA.

5.12 rRNA Methylation and Hyperthermophily

We have noted an interesting correlation between optimal growth temperature of an archaeal species and the number of highly probable sRNAs that were identified by our search program. At one end of the spectrum is *M. thermoautotrophicum* (65 °C growth temperature) where no highly probable sRNAs were identified, and at the other end are the *Pyrococcal* species (100 °C growth temperature) where 50 or more sRNAs were identified. The remaining species with growth temperatures between these extremes produced intermediate numbers of candidate sRNAs. There are at least two explanations for this apparent correlation. First, higher growth temperatures may require a larger number of methylation modifications in order to efficiently fold, process, assemble or stabilize rRNA within ribosomal particles. The observation by Noon and colleagues (Noon et al., 1998) that the amount of rRNA methylation in *S. solfataricus* increases with increasing growth temperatures is consistent with this possibility. Second, it is possible that we systematically identified only a specific subset of highly uniform sRNAs that contain the characteristic features found in the cDNA clones isolated from *S. acidocaldarius* and used to define the parameters of our search program. If this is correct it may mean that the sRNA sequence/structure is more rigidly defined at higher growth temperatures; at lower growth temperatures where the sequence/structure requirements might be less stringent, our program would become less efficient in identifying sRNAs.

Do mesophilic and moderately thermophilic Archaeal species contain C/D box sRNAs? The answer is probably yes. The gene encoding the sRNA associated protein aFIB was first identified in two mesophilic Archaeal species, *Methanococcus voltae* and *Methanococcus vannielii* (Amiri, 1994), and both the aFIB and aNOP56 genes are present in the genome of the moderate thermophile, *M. thermoautotrophicum*. Indeed, our search of the *M. jannaschii* genome revealed other less probable candidate sRNAs; it seems likely that some of these candidates are legitimate sRNAs but independent conformation is required.

5.13 Conclusions

These findings imply that an RNA guide mechanism for directing 2'-O-ribose methylation to specific positions in rRNAs was already well established in the common ancestor of archaea and eukaryotes (Woese et al., 1990). Neither a fibrillarin homolog nor C/D box sno-like RNAs have been described in bacteria. Therefore, it is not clear whether C/D box sRNAs were ancestral to the three surviving lineages and then either lost from or not incorporated into the bacterial lineage. Alternatively, C/D box RNAs may be a derived feature in a common ancestor of Archaea and Eukarya.

sRNA	Ab	sRNA PE ?	Guide Box	Target Site	Match/ Mismatch	Methyl Site Confirmed?
Sac-sR1	F	•	D	16S U52 (W)	11/0	•
Sac-sR2	F	•	D	23S C1914	11/0	•
Sac-sR3	F	•	D	23S G2739	10/0	No
Sac-sR4	N	•	D	23S G1995	10/0	ND
Sac-sR5	F,N	•	D	16S G1056 (W)		•
Sac-sR6	F	•		NF		
Sac-sR7	F	•	D'	23S G2649	9/0	•
			D	23S U2692	10/0	•
Sac-sR8	N	•	D'	23S U2972	9/1	ND
			D	23S G334	12/1	•
Sac-sR9	F	•	D'	16S G926	8/0	ND
Sac-sR10	F	•	D'	tRNA Gly-CCC C50	12/0	ND
			D	23S C2539	9/0	ND
Sac-sR11	F	•	D'	23S A2618	10/0	ND
			D	23S A724	11/2	ND
Sac-sR12	F	•	D'	23S G1114	11/0	No
			D	23S A1134	9/1	•
Sac-sR13	N	No	D'	23S G385	10/1	ND
			D'	23S G2996	11/1	ND
			D	23S C2746	10/0	•
Sac-sR14	F,N	•	D'	16S A468	12/0	No
			D	tRNA Gln-UUG U34	10/0	ND
Sac-sR15	F	•		NF		
Sac-sR16	N	•		NF		
Sac-sR17	F	No		NF		
Sac-sR18	N	ND	D'	23S G140	9/1	ND

Table 5.1: **Predicted Target Ribose Methylation Sites for *Sulfolobus acidocaldarius* sRNAs.**
"Ab" column indicates protein antibody used to co-precipitate sRNA, either α-aFIB "F", or α-aNOP56 "N". "sRNA PE" column indicates sRNAs verified to have primer extension products of the correct approximate length. "Guide Box" indicates the box adjacent to complementarity. "Match/Mismatch" indicates the number of Watson-Crick and G-U pairings versus the number of all other pairs in the guide region/target RNA duplex. "Methyl Site Confirmed" indicates predicted methylation sites confirmed by the primer extension pause assay (Maden et al., 1995; Kiss-Laszlo et al., 1996).

sRNA	sRNA PE?	Guide Box	Target Site	Match/ Mismatch	Methyl Site Confirmed?
Sso-sR1	•	D'	16S U605	10/0	•
		D'	16S U33	8/0	ND
		D	16S U52 (W)	10/0	•
Sso-sR2	•	D	16S G1372	10/0	ND
Sso-sR3	•	D	16S C1490	10/0	ND
Sso-sR4	•	D'	16S G473	8/0	ND
		D'	23S G810	8/1	ND
		D	16S C277	9/1	•
Sso-sR5	•	D'	23S A1183	11/0	ND
		D	23S A685	10/0	ND
Sso-sR6	•	D'	23S G2127	9/0	ND
		D	23S G2094	10/0	•
Sso-sR7	•	D	16S C481	10/0	•
		D'	23S A2425	10/0	ND
Sso-sR8	•	D	16S U477	9/0	No
Sso-sR9	•	D'	23S A682	11/0	ND
		D'	23 A2314	8/0	ND
		D	23S A1461	10/1	No
		D	23S A54	9/1	ND
Sso-sR10	•	D	23S A2082	11/0	ND
Sso-sR11	ND	D'	tRNA Gln-CUG G18	10/0	ND
Sso-sR12	ND	D	16S U1344	12/0	ND
Sso-sR13	ND	D	16S G1018 (W)	11/0	ND

Table 5.2: **Predicted Target Ribose Methylation Sites for *Sulfolobus solfataricus* sRNAs.**
"sRNA PE" column indicates sRNAs verified to have primer extension products of the correct approximate length. "Guide Box" indicates the box adjacent to complementarity. "Match/Mismatch" indicates the number of Watson-Crick and G-U pairings versus the number of all other pairs in the guide region/target RNA duplex. "Methyl Site Confirmed" indicates predicted methylation sites confirmed by the primer extension pause assay (Maden et al., 1995; Kiss-Laszlo et al., 1996).

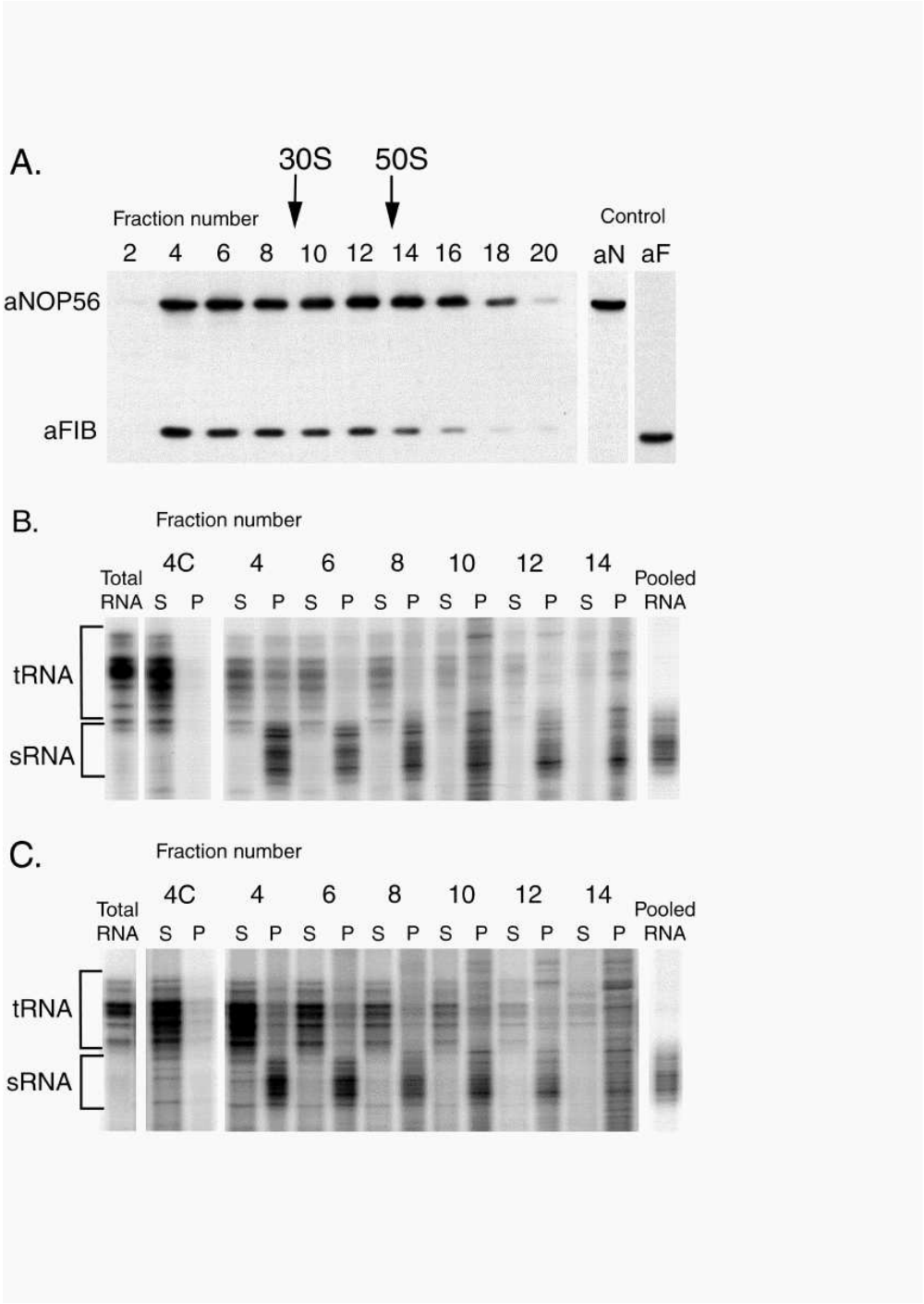

Figure 5.1: **Glycerol gradient sedimentation of aFIB and aNOP56 containing particles present in *S. acidocaldarius* cell free extracts.** (see legend next page)

Legend for Figure 5.1.

A sonicated cell extract was precipitated by addition of 35% ammonium sulfate, redissolved in buffer (50 mM Tris, pH 8), layered onto a 35 ml 5-30% glycerol gradient in the same buffer, and sedimented in an SW27 rotor (10 °C, 17 K, 16 hr). Fractions (1.5 ml) were collected.

(A) Aliquots of every second fraction between 2 and 20 were simultaneously analyzed by Western blotting for the presence of aFIB and aNOP56 using the two antibodies prepared against the recombinant proteins expressed and purified from *Escherichia coli*. The positions of 30S and 50S ribosomal subunits in the gradient are indicated. In the control (right), the antibodies were shown to be highly specific as seen by blotting each separately to *S. acidocaldarius* crude cell extract (right).

(B) Aliquots from every other gradient fraction between 4 and 14 were immunoprecipitated with anti-aFIB as described previously , and RNA was recovered by phenol extraction from the precipitates (P) and the supernatants (S). Only about 0.1% of the RNA in each fraction was coprecipitated with the antibody; the bulk of the RNA was retained in the supernatant. To visualize the precipitated RNAs, aliquots (0.005% and 2.5% of the total RNAs recovered from the supernatant and pellets, respectively) were pCp end-labeled with RNA ligase and displayed on an 8% denaturing polyacrylamide gel. The positions of tRNA and sRNA are indicated on the left. The precipitated RNA recovered from fraction 5 was separated on an 8% denaturing polyacrylamide gel, recovered by electroelution, and used as template for RT-PCR cloning (8). An aliquot of the RNA recovered after electroelution was end-labeled and displayed on an 8% denaturing polyacrylamide gel (right).

(C) Aliquots from every other gradient fraction between 4 and 14 were immunoprecipitated with anti-aNOP56. Other details are as described above except that recovered RNAs from fractions 6-8 and 10-13 were pooled and used for cDNA cloning. An aliquot of the pooled RNA was end-labeled and displayed on an 8% denaturing polyacrylamide gel (right).

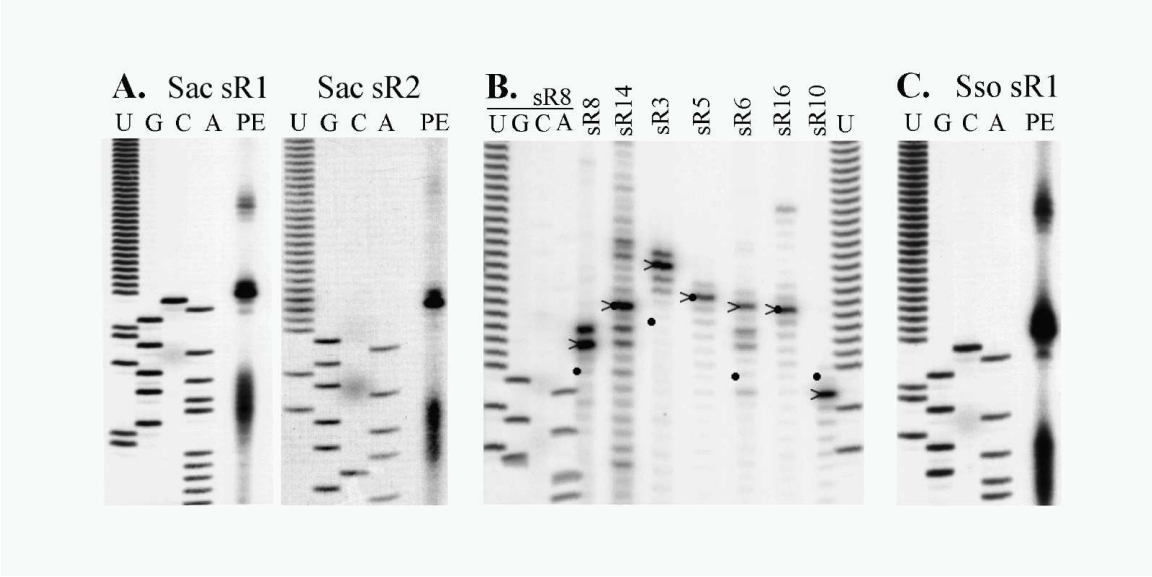

Figure 5.2: **Detection and 5' end mapping of sRNAs from *S. acidocaldarius* and *S. solfataricus*.**
Primers specific for the D box guide region of Sac sR1 to sR17 were 5' end labeled with $\gamma-^{32}$P-ATP and polynucleotide kinase and used in extension reactions with total RNA (20 μg) isolated from *S. acidocaldarius* as template.

(A) The extension products obtained with Sac sR1 and sR2 specific oligonucleotide primers were run alongside a DNA sequence ladder generated with the same primers and Sac sR1 or sR2 cDNAs as template.

(B) The extension products obtained with Sac sRNA-specific primers and run with the DNA sequence ladder generated with Sac sR8 cDNA clone. For each extension reaction, the approximate positions of the 5' terminal nucleotide in the corresponding cDNA is indicated by a (>) beside the lane. [Note: there is an inconsistency in fragment lengths for this experiment between the Dennis lab (gel picture here) and my own primer extensions. This experiment will be repeated and the conflict resolved before submission of this manuscript.]

(C) The primer extension reaction was as in (A) except that the primer was specific to Sso sR1 and total RNA from *S. solfataricus* was used as template. The DNA ladder was generated using Sac sR1 primer and the Sac sR1 cDNA as template. The Sac and Sso primers differ at two internal positions but have the same 5' and 3' end location.

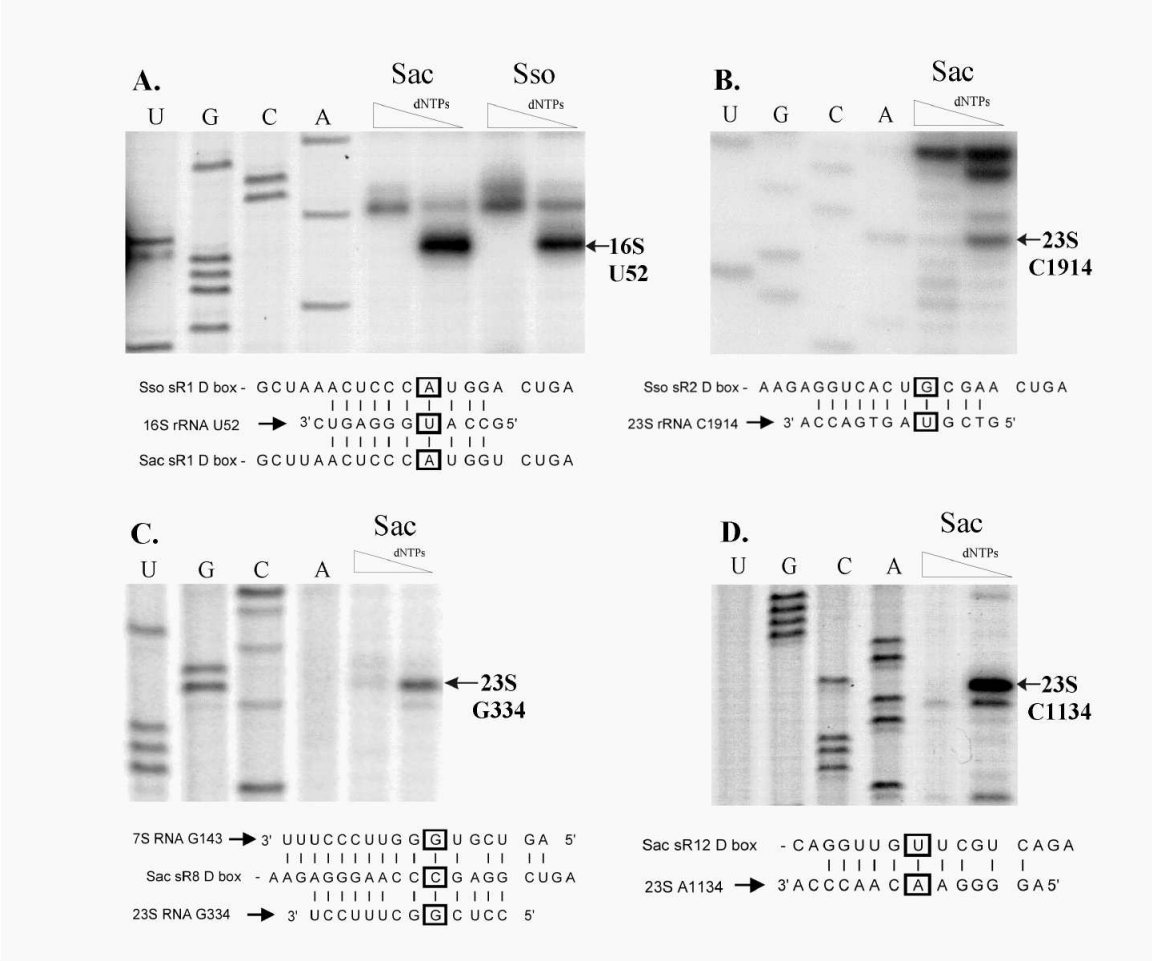

Figure 5.3: **Detection of methylation sites in 16S and 23S RNAs by primer extension pausing.**
Primer extensions were carried out by the dNTP concentration-dependent primer extension assay as previously described (Lowe & Eddy, 1999). RNA sequencing ladders of ribosomal RNA regions being assayed appear in left four lanes. To the right of sequencing ladders, pairs of reverse transcriptase primer extensions on the same RNA sample are shown. Odd lanes use high dNTP concentration (1.0 mM) reactions and even lanes contain low dNTP concentration (0.004 mM) reactions. 2'-O-methyl modified nucleotides are characterized by appearance of termination bands in low but not high dNTP concentration reactions. Proposed rRNA-sRNA guide duplexes appear below primer extension gels. [Note: disregard 7S guide diagram in part C. This will be removed in the final version of the manuscript].

S. acidocaldarius

	C box	Compl R1	D' box	C' box	Compl R2	D box
Sac-sR01	GAG TTGATGA	GAAGTTAAAAAA	GCGA T	GGATGA	GCTTAACTCCCATGGT	CTGA TAAC
Sac-sR02	GA GTGATGA	GACGAGCGCTAA	CAGA GAGAG	TGAAGA	GGTCACTGCGAA	CTGA AGAAA
Sac-sR03	AGG ATGACGA	GACCCAAAATA	TTGA A	CGATGA	TATAACCTGTCTCGG	CTGA TCAGT
Sac-sR04	G TTGATGA	GCACATTCTTTT	CTGA TTTAA	TGAAGA	AAGTGGCCAGGT	CTGA GGTAG
Sac-sR05	GAA ATGATGA	ATGGTCGACGGAA	CGGA CCTA	TGAAGA	ATTGTTGCCGGA	CTGA CAAAC
Sac-sR06	GG ATGATGA	CCAAATAGA	CTGA A	AGATGA	AGAAATGCACCTCAA	CTGA CTAAA
Sac-sR07	G ATGATGA	CAAAGAGCCGAA	TGGA T	TAGTGA	CATCTAATTTTGTGGGCAGCCA	CTGA TAGAG
Sac-sR08	G ATGATGA	AGCCCGCCATCAA	CAGA TAAG	TGAAGA	GGGAACCCGAGG	CTGA GAAT
Sac-sR09	GTTAAAATA ATGATGA	CTAACTCCAATA	CTGA CCAA	TGATGT	CGTAACCCGAAA	CTGA ATAAA
Sac-sR10	GAATG ATGTGGA	ATCCGGGAT	CTGA GAA	TGATGA	CAAAAAGCGCGAGCG	CTGA TTATA
Sac-sR11	GAATGTG ATGATGG	GTCGATGTTA	CTGA TTAGT	TGATGA	GATTATCTCCGG	CTGA GAATCTCGAGTT
Sac-sR12	GA ATGAAGA	ACCCAACCTTAT	CTGA GGTTA	TGATGA	CAGGTTGTTCGT	CAGA TCGATGTGAG
Sac-sR13	AGG ATGATGT	ACTTTCACCCTCA	CTGA AAGG	TGAGGA	TGAGTCCGACTA	CTGA CGCAA
Sac-sR14	GCT GTGAAGA	CGCTAGACTTAGA	CTGA CTCA	TGATGA	AGGGCCAAAGCT	CAGA GCAAAC
Sac-sR15	A GTGATGA	GGAACCAACGAGAG	CTAG TT	TGATGG	CTTCGACGCTCTGCT	CTGA AA
Sac-sR16	GA ATGAAGA	CGTTCCACCCGA	GCGA G	TGATGA	GCGAAACGGTTAATACTG	ATGA TG
Sac-sR17	AGAA ATGAAGA	GTAAAAAACCGG	CTGA GATAAG	TGATGA	CGACGTCTCGCA	CTGA TC
Sac-sR18	AA GTGATGA	CAGAACCCCGGC	TTGA A	AGATGA	TAGAGCCGTGTGAGAA	CTGA TCAAT

S. solfataricus

Sso-sR01	ACAG ATGATGA	ATTCCCGATAGT	ACGA T	TGATGA	GCTAAACTCCCATGGA	CTGA TTAG
Sso-sR02	TGTT GTGATGA	AAGGGAAAGAT	CAGA TCTG	TGAAGA	AGTGGAGCGACA	CTGA GGTG
Sso-sR03	GTCG ATGATGA	GTCAAGAAAA	ATGA TT	TGATGA	TTTTTGAGGTGATCT	CTGA ATAA
Sso-sR04	GGGA ATGATGA	GCCACGCCAGAA	CTGA GCCA	GGATGA	ACGGCTTGGGAG	CTGA CCCC
Sso-sR05	ATTG GTGATGA	CGCCCTTCAGT	TTGA T	TGATGA	TAAGGGCCCGTGTTA	CTGA ACAT
Sso-sR06	GTGT GTGATGA	TAAGGGACCACA	TGGA GTGG	TGAAGA	TAACTTACCCGT	CTGA ATAT
Sso-sR07	ACAT GTGATGA	AGACCTTTGGGA	CTGA TAAAG	TGATTG	AGTGGCGGCTGT	CTGA CATG
Sso-sR08	GGGA ATGATGA	GGGTTCCAGAG	CTGA AGCG	TGATGA	ATGGTTGACACG	CTGA CCCC
Sso-sR09	TAGC GTGATGA	ACCGTGTTTCAG	CTGA TCTG	TGACGA	TAGCCCTTCGG	CTGA GCAA
Sso-sR10	TAGA GTGAAGA	ATTACCCTCGGG	CAGA TTAA	TGATGA	AGACAAGGCATTTGT	CGGA TTTA
Sso-sR11	TTAA ATGATGA	GCTTGACCACTT	ATGA AGCT	AGATGA	TATAATAAAGGTAGCCG	CTGA GTAT
Sso-sR12	TCAA GTGATGT	AGCAGGGACGTA	CTGA AGCTAA	TGTGGA	AAGAGTATAAAGA	CTGA CCTA
Sso-sR13	TGAA ATGATGT	CTGGTGACGGCT	CTGA GCCAG	TGATGT	CGCCAAAAAGT	CTGA TCAA

Figure 5.4: **Alignment of *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus* sRNAs.** Box features and complementary regions (Compl R#) labelled.

M. jannaschii

	C box	Compl R1	D' box	C' box	Compl R2	D box
Mja-sR01	GCAG ATGATGA	CGTTTATCCCCGT	CTGA GTTA	TGATGA	GTAGCAAGCCGG	CTGA TGCC
Mja-sR02	GCGG ATGATGA	ACGGAGTAGCTG	CTGA GCTA	TGATGA	TTGATGGGCGAA	CTGA CGCC
Mja-sR03	GGCA ATGATGA	AAAGAGGGTTAG	CTGA ACTG	TGATGA	TACTTACCCGAA	CTGA GCCA
Mja-sR04	CGCG ATGATGA	GTGCCGACTTCA	CTGA ACTG	TGATGA	AACCTGGGACAG	CTGA GCGT
Mja-sR05	CTCG ATGATGA	GCAATAAAAAG	CTGA CTTAATA	TGATGA	ACCTTTCGGGGTAT	CTGA GAGG
Mja-sR06	GGCG ATGATGA	CAATTTCGCTAT	CTGA TTCTG	TGATGA	CTACTCCCGCAG	CTGA GCCA
Mja-sR07	GGGG ATGATGA	TACATCGATGTG	CTGA ATAT	TGATGA	TGAACGCGCCCTTCT	CTGA CCTT
Mja-sR08	GCCA ATGATGA	CGATTGGCTTTG	CTGA GTCTG	TGATGA	ACCGTATGAGCA	CTGA GGCG

A. fulgidis

Afu-sR01	GCCG ATGATGA	CTGATGGGCGAC	CTGA GAAA	TGATGA	AAGGGAGAGT	CTGA GGCT
Afu-sR02	ATAG GTGATGA	ATCTAGCAGGAT	CTGA GCCGTGGCGA	TGATGA	CGGCTGTACTCT	CTGA TTAC
Afu-sR03	CGGC GTGATGA	TTGACGGGTCTG	CTGA GCGG	TGATGA	CCCGTTGGAGCT	CTGA CCCG
Afu-sR04	AGCG ATGATGA	AAGGGCCCCT	CTGA ACGG	TGATGA	GGGTTTTGAGTT	CTGA GCTC

Figure 5.5: **Alignment of *Methanococcus jannaschii* and *Archaeoglobus fulgidis* sRNA predictions.**

A. pernix

	C box		Compl R1	D' box		C' box		Compl R2	D box	
Ape-sR01	GCCG	ATGATGA	GTTCTAGCCTTA	CGGA	CACG	TGAAGA	CAGTGGCCAGGC		CTGA	GGAT
Ape-sR02	TGGG	ATGATGA	CGCCGTCCCAGC	ATGA	AGGG	TGATGA	CCTCACGACGAT		CTGA	CCAG
Ape-sR03	AGCG	ATGAGGA	CGGTGGAGCGACG	CTGA	CCAAGGTAGGGGA	TGAGGA	GCGACACTCCC		CTGA	GCAC
Ape-sR04	GGCG	ATGAGGA	TAGGCGGGCTTTT	CTGA	GACGTG	TGATGA	GTGCGGTCGTAT		CCGA	GTCC
Ape-sR05	GCTA	GTGATGA	TCCATGCAGGCC	CTGA	GGTCCTGCCCGCGC	CGATGA	AGAGCCTATGGGACTACCCT		CTGA	AGCC
Ape-sR06	GGCT	GTGATGA	CGGCTGGTATTG	CCGA	AGGCAGGGGCCGT	TGAGGA	GGGACACCAGTCT		TTGA	GCCG
Ape-sR07	AAGT	GTGATGA	TTTACTCCCGGA	CCGA	GCAATGCTG	TGAGGA	CTACTTACCCGG		CTGA	CAGT
Ape-sR08	CGGA	GTGATGA	GTAAGGCTGGGATTG	CTGA	AAGCTGTTTTG	AGATGG	AGCGTGATGAGCAGGG		CTGA	GCGC
Ape-sR09	GGGG	ATGGTGA	GGGGTGCGAGCG	CAGA	GCCCCAGGGGCGG	TGAAGA	AGGATTATAACAGCGT		CTGA	CCCT
Ape-sR10	GGGC	GTGATGA	GGGCTTAACGG	CTCA	CGTTTTCT	GGGAGA	GTGTCGCACCT		TTGA	GGGC
Ape-sR11	GCCG	ATGATGT	CTACTCCCGACT	ACGA	CACGCCGTGG	GGATGA	GTTGAAGGCATTTGG		CTGA	GGCA
Ape-sR12	CCCT	ATGAAGA	GGAGTACATGCGGTGG	TGGA	AGCTCCCTTGAAGCGTGTAGGAGACCT	TGAGGA	CCTCGCCGGG		CTGA	GGGT
Ape-sR13	ATGT	ATGATGA	CACGGGGCTTCC	TCGA	CGCCTCCCGCAGACGGGTGGTGTA	TGAAGA	AAGGTAGAGACG		CTGA	CTCA
Ape-sR14	CCCG	ATGAAGA	GTACGCCATCGC	CCGA	CGGACTCTTTCCAGCCGT	GGGGGA	AGCCTGTCCCCC		TTGA	GGGT
Ape-sR17	GGGG	ATGAAGA	GTTCTTTGA	CCGA	CGCTAGTGGTA	TGAGGA	GGCTGAGCTACG		CTGA	CCCC
Ape-sR18	GCCG	ATGAAGA	GTGGTATTACCG	CCTA	GTAGCGGCGATAG	CGATGA	CGAGGCCCCAGTG		CTGA	GGCC
Ape-sR19	TTGA	ATGCTGA	ACTATCAATACAC	TAGA	TAACGTGGGGA	TGACGA	GGCGGGCAGCCT		TTGA	GGAC
Ape-sR20	CGGG	ATGAAGA	CTAGCTAACCGTGG	CTCA	CAGCTATG	GGATGG	CCTAGGGCTCTCGC		CGGA	CCGC
Ape-sR21	GGGA	GTGATGA	GGGTTAAAAGCG	CCGA	CTCTGGGTAGCA	TGGGGA	AGGCCAGGACAG		CTGA	CCAT
Ape-sR22	GGGG	ATGAGGA	CTCCACGGGTT	CCGA	CGCCCCCTCGGGGG	TGGGGA	CACCTGTCTCAA		CTGA	CCCT
Ape-sR23	GCCT	GTGAAGA	CGGAGGTGGGT	GTGG	GCTTA	TGGAGG	CTACGGGGGCTT		CTGA	GGCT
Ape-sR24	GCGG	ATGAAGA	CACTTGCCCCCAC	CTGA	GCCTGTG	TGAAGA	GATTACCGCGGACTC		CAGA	CGCC
Ape-sR27	GGGC	GGGATGA	GGTGTGACTCCAATT	CTGA	GCCTCAAGGGCCGTG	GGGAGA	TCCACCCCTAT		CTGA	CCGC
Ape-sR28	CGGG	ATGAAGA	CTACTCCCCAGA	CCGA	GGCAAGTCTCG	AGATGA	CGACAATTCCTTTAG		CAGA	CCGC
Ape-sR29	CGGT	GTGATGA	ACACCTCGATCT	CCGA	TCGCTGGGA	TGAGGA	TGCGGTGGCTGT		CTGA	CTCT

Figure 5.6: **Alignment of 25 *Aeropyrum pernix* sRNA predictions.**

P. horikoshii

	C box	Compl R1	D' box	C' box	Compl R2	D box
Pho-sR01	AAAGAAGGCG	ATGATGA	AGCCTTCCGCAC	CTGA	ATGA	TGAGGA	GTGGACGGCTTC	CTGA	GCCTACTCCT
Pho-sR02	AAAAAGAGGG	ATGATGA	GTTTTTCCCTCACT	CTGA	GGAG	TGATGA	GGAGCCGATGCA	CTGA	CCTCGATCAT
Pho-sR03	AATTGTGGCG	ATGATGA	ATAGCAAGCCAG	CTGA	AGAG	TGATGA	AGTGAACACCCC	CTGA	GCCTACCTAA
Pho-sR04	CCGAGTTGGG	ATGATGA	GGGAGATTTCGG	CCGA	GTGG	TGAGGA	GACTCGCATGGG	CTGA	CCTTTCTAAG
Pho-sR05	ATAAGGTGTG	ATGATGA	ACGCCATCGATA	CTGA	GATA	TGATGA	CCGGATTCCTGG	CTGA	TTTCTTTTAT
Pho-sR06	TGGAAATGGG	ATGATGA	AGTTTGCTACCC	CTGA	AGAA	TGATGA	ACCCTGCCGTTA	CTGA	CCATTTAGCT
Pho-sR07	GTAATCCGGG	ATGATGA	ACCTCCATCCCAA	CTGA	ATAAA	TGATGA	ATGCACATCAGG	CTGA	CATTACCTTT
Pho-sR08	CGAATAGGCG	ATGATGA	GCTCCATCCCTAC	CTGA	GTTG	TGATGA	ATGTAGCGCGCT	CTGA	GCAACCCTTA
Pho-sR09	CCGACAGGGG	ATGAAGA	GCTTTTGCTTTG	CTGA	GCAGA	TGATGA	CCACGCCCTTCG	CTGA	CCTGCTATTT
Pho-sR10	TTAGCGACTA	ATGATGA	ACTACTCCCGG	CTGA	GGGG	TGATGA	ACCACCTACCGG	CTGA	GGTGAAAGCA
Pho-sR11	CTCTAGCGGG	ATGATGA	CTTTGCCGAGTG	CTGA	GCTGG	TGATGA	GTAAACAGTCGT	CTGA	CTTTCCCTTT
Pho-sR12	ATCGGCTTGG	ATGATGA	GCGTTTACCGGT	CTGA	GCTG	TGATGA	TATCGGCACTGT	CTGA	CTAGTAATCT
Pho-sR13	TCATTGTTGG	GTGATGA	GATGGCGGATTG	CTGA	GAGA	TGATGA	GGACCTTAGGGG	CTGA	TTAAATTTCG
Pho-sR14	CTATTAACCA	ATGATGA	CGGATCAACCGG	CTGA	TCGAA	TGATGA	CGTCCGCATCAC	CTGA	GGGTTTCAAG
Pho-sR15	AGATGAAGAG	ATGATGA	GTAAACCCGTTG	CTGA	GCGG	TGATGA	GAGGATCGACTAG	CTGA	ACAACATCTT
Pho-sR16	ATCAAGTCTG	ATGATGA	ACCTTCCCCTCAC	CTGA	AAGG	TGATGA	GCACACCGGTAGG	CTGA	GGGTGATAAT
Pho-sR17	ACGGTCTGCG	ATGATGA	GAGCGAACTGCA	CTGA	AAAG	TGATGA	CAGGGCCTTG	CTGA	GCGGTGATCG
Pho-sR18	TTTATTTTTA	ATGATGA	AACAGCCAGGACC	CTGA	TGGGA	TGATGA	GTGGTGGGCTTAG	CTGA	TGTTGCGGTA
Pho-sR19	TGGCGGGCTC	GTGATGA	GCTTTCCCTACGGC	CCGA	GCTTAGG	CGATGA	GGAATACAGCCAGGG	CTGA	TTTTGGTGAT
Pho-sR20	ATGAATGGCG	ATGATGA	GGCCTCGATTGG	CTGA	AT	CGATGA	TTGAGAGGGACTTGG	CTGA	GCGGTGATTA
Pho-sR21	TACCATGCCG	ATGATGA	GACCGGTACTGG	CCGA	AGTA	TGATGA	GCACTCGGTTAG	CTGA	GGCCTGAAAA
Pho-sR22	GGAACATCCG	ATGATGG	GAACAGGGTAGCTG	CCGA	GT	TGATGA	GGAAGCCGTTCCAGA	CTGA	GGAAAGAAAA
Pho-sR23	TCCTATTAAG	ATGATGA	ATCTGGAGCCCC	CTGA	TCGG	TGATGA	TCAGTCTGCGGAG	CTGA	TAACATGATT
Pho-sR24	AAGTTACCCT	ATGATGA	GAGAGCTGTTAA	ATGA	G	CGGTGA	TTAAAGGATGGCTGG	CTGA	GGGTGAGATA
Pho-sR25	TGAAGGTTCA	ATGATGA	AAACTCCCTGAT	CTGA	AAAATAAA	TGATGA	AGACCGGTTCA	CTGA	GATTTCTGCT
Pho-sR26	TAAAAAGGCG	ATGATGA	GTGATGGGCGAA	CTGA	AAACA	TGATGA	AGGTAGGTGATCT	CTGA	GCTATCATCG
Pho-sR27	TAGCTTCAGA	GTGATGA	GCCCGCGCAGCG	CTGA	TAGA	TGATGA	AGATTTTAAGTG	CTGA	CTTCTAACCT
Pho-sR28	ATACTTAGGA	ATGATGA	CCGGTTTCGGGA	CTGA	ACGA	TGATGA	CACCAGCTATCG	CTGA	CCTTGTCTAA
Pho-sR29	TTGTGGGCGG	ATGATGA	TCCTTGCCCAGC	CTGA	GGAG	TGATGA	AGTCGGTATTAG	CTGA	CGCTGTGTTT
Pho-sR30	ATTATTGTGG	ATGATGA	GTAGTCGGAAGG	CTGA	GCTGA	TGATGA	AGGACGCCCATGT	CTGA	TCCTTATCTT
Pho-sR31	GGGATTAAAG	ATGATGA	ACTCGGCAGGTC	CTGA	TTCG	TGATGA	GGATCTTGAGTG	CTGA	TTTACCTTAC
Pho-sR32	TAAAATGGAG	ATGATGA	GCTTGGCCGCTAC	CCGA	G	TGATGA	GGGCCAAACTCCGGTG	CTGA	TCACATTGCA
Pho-sR33	CCTCTTTGAC	GTGATGA	GGAATCGGGGAG	CTGA	AGCTA	TGATGA	GATCCGCCAACG	CTGA	TTTTCCAAAT
Pho-sR34	ATCCAAGCCG	ATGAGGA	TCGTTAGCCACG	CTGA	GGA	TGATGA	TAAGAGGGTTAG	CCGA	GGCTTATTTT
Pho-sR35	AGACGAAAAA	ATGATGA	GTACGGGGCCAC	CTGA	GCGG	TGATGA	GGTTTCTCCCAAGT	CTGA	TTACCTAACC
Pho-sR36	GGTTTGGTGG	ATGAAGA	GAGGGTAGGTAG	CTGA	TCTG	TGATGA	ACTAATCGGCCC	CTGA	CACGGGGTGA
Pho-sR37	GGTTAGAGCG	ATGATGT	AGATTAGCCCCGA	CTGA	GCGG	TGATGA	GGCTGGCCCATCG	CTGA	GTCCCAAAAA
Pho-sR38	GGAGTAGGCG	ATGATTT	GACTCCGGAAAAG	CTGA	ATGA	TGATGA	AGTCCAGCCCGA	ATGA	GCCGGGGTGA
Pho-sR39	AAAGGGGGTG	ATGAGGA	AAATTTCGCTAG	CTGA	AGTGAAAGA	TGATGA	ATACCCTCGCAA	CCGA	GCCTATTTTA
Pho-sR40	TCCAGAGGCC	ATGAGGA	TAGGCGGGTTTG	CTGA	CCTCGGGGCG	TGATGA	ACCTTTGGAGCC	CCGA	GGGCCGGAGA
Pho-sR41	TTGAAGGCGG	ATGATGA	AGGCTAATTT	CGGA	TTGG	TGAAGA	GCCCTATGAGCG	CTGA	CGCTATGATT
Pho-sR42	CTACTTTGTG	ATGATGA	TCTTCACGGTACC	CTGA	GAGGCGCTCGA	TGATGA	GCCCCCTCTCAA	CTGA	TCATAGCTTT
Pho-sR43	AGTAATTGAG	ATGAAGA	AAAAGCACCTCCA	CTGA	GTGA	TGATGA	CACGCCCACGGT	CTGA	TCAAACTGTT
Pho-sR44	CGTCAAGGCG	ATGAGGA	GATCGGTCTTGT	CTGA	AGAA	TGATGA	AAGTGGTCACCG	CTGA	GCCTTAAGAA
Pho-sR45	AAAAAACGGG	ATGAGGA	AGCCGAGGACAC	CTGA	AGGA	TGATGA	CTCTTCGCTTCG	CCGA	CCGGCGGTGA
Pho-sR46	TCAAAAGGCA	ATGAGGA	ATGAATCCAATG	CTGA	GCTTAGG	CAATGA	TTGGCCCCAGAGTGG	CCGA	GCCTTCATAT
Pho-sR47	AGTTCAACCG	ATGAAGA	GCCTTTTGGGCC	CTGA	GAGA	TGATGA	AAGCCCCCCTAT	CTGA	GGGGGATAAA
Pho-sR49	TTCCTGGGCG	ATGAAGA	ACTGGATCCGAG	CTGA	TCTCA	TGATGA	AGGAGGTTTAAC	CTGA	GCCTCCATGA
Pho-sR53	AGTCCGGCCG	ATGACGA	AGGTGGGCACTCT	CCGA	TT	TGATGA	GGATGTGGGGCGAGGAGC	CAGA	GGCTTTATAA
Pho-sR54	ATTAGTTGGG	ATGATGA	AGTTCGCTAGAA	CTGA	GAGG	TGATGA	CGCCCAGGGTAT	CTGA	CTTTTCCAAT

Figure 5.7: **Alignment of 50 *Pyrococcus horikoshii* sRNA predictions.**

Pyrococcus Homolog Groups

	5' flanking	C box	Compl R1	D' box	C' box	Compl R2	D box	3' flanking
Pho-sR01	AAGTAAAG AAGGCG	ATGATGA	A GCCTTCCGCAC	CTGA	ATGA	TGAGGA	GTGGACGGCTTC	CTGA	GCCT ACTCCTTA
Pab-sR01	AAATATAAAT GGCA	ATGATGA	A GCCTTCCGCAC	CTGA	ACGG	TGAGGA	GTGGACGGCTTC	CTGA	GCCT CACTCCTT
Pfu-sR01	AGGGAGGT AAGGCG	ATGATGA	T GCCTTCCGCAC	CTGA	TTGG	TGAGGA	GTGGACGGCTTC	CTGA	GCCT ACTCCTTA
Pho-sR02	AAATAAAAAG AGGG	ATGATGA	GTTTTTCCCTCACT	CTGA	GGAG	TGATGA	GGAGCCGATGC A	CTGA	CCTC GATCATTG
Pab-sR02	GAAAGAGAAA AGGG	ATGATGA	GTTTTTCCCTCACT	CTGA	GCCG	TGATGA	GGAGCCGATGC T	CTGA	CCTC TGCCATAA
Pfu-sR02	AAGAGAAATGG GGG	ATGATGA	GTTTTTCCCTCACT	CTGA	TTAG	TGATGA	GGAGCCGATGC A	CTGA	CCTC GAGCATTA
Pho-sR03	ATATAATTGT GGCG	ATGATGA	A TAGCAAGCCA G	CTGA	AGAG	TGATGA	A GTGAACACCCC	CTGA	GCC TACCTAATC
Pab-sR03	AAACTATGAA GGCG	ATGATGA	G TAGCAAGCCA C	CTGA	CCTA	TGATGA	G GTGAACACCCC	CTGA	GCC AATTTCATC
Pfu-sR03	AATCACCTCC GGCG	ATGATGA	A TAGCAAGCCA C	CTGA	AGAG	TGATGA	G GTGAACACCCC	CTGA	GCC TATTCCATG
Pho-sR04	TTCCCCGAGT TGGG	ATGATGA	G GGAGATTTCGG	CCGA	GTGG	TGAGGA	G ACTCGCATGGG	CTGA	CC TTTCTAAGGG
Pab-sR04	CTCTGAAGGA TGGG	ATGATGA	A GGAGATTTCGG	CCGA	AAGG	TGAGGA	A ACTCGCATGGG	CTGA	CC TTTCTCAGAG
Pfu-sR04	TTAGAGCATAA GGG	ATGATGA	A GGAGATTTCGG	CAGA	GGTG	TGAAGA	G ACTCGCATGGG	CTGA	CC ACCACCTTTA
Pho-sR05	GATAATAAGGTGT G	ATGATGA	A CGCCATC-GAT A	CTGA	GATA	TGATGA	CCGGATTCCTGG	CTGA	TTTCT TTTATT T
Pab-sR05	AGATAAGACAATA G	ATGATGA	G CGCCATC-GAT G	CTGA	GGAG	TGATGA	CCGGATTCCTGG	CTGA	TCT TT ATT CTTT
Pfu-sR05	TGAGCTTATTTTGA	ATGATGA	G CGCCATCCGAT A	CTGA	GGGCA	TGATGA	CCGGATTCCTGG	CTGA	TCT CATT TTCTC
Pho-sR06	AGATTGGAAAT GGG	ATGATGA	AG TTTGCTACC C	CTGA	AGAA	TGATGA	A CCCTGCCGTTA	CTGA	CC ATTTAGCTCG
Pab-sR06	GACTCCAATG AGGA	ATGATGA	GA TTTGCTACC A	CTGA	GCAG	TGATGA	G CCCTGCCGTTA	CTGA	CC TTTTACTATT
Pfu-sR06	CTTGAAAAAT AGGG	ATGATGA	GA TTTGCTACC T	CTGA	AAATAA	TGATGA	A CCCTGCCGTTA	CTGA	CC GTTTAACTCC
Pho-sR07	CGAGGTAATCC GGG	ATGATGA	A CCTCCATCCCA A	CTGA	ATAAA	TGATGA	A TGCACATCAGG	CTGA	CATTACCTTTTT
Pab-sR07	TGAAATTATAA GGG	ATGATGA	G CCTCCATCCAA G	CTGA	GCAAG	TGATGA	G TGCACATCAGG	CTGA	ACCTTTTTATCC
Pfu-sR07	TTTGGAGATTTTC G	ATGATGA	G CCTCCATCCCA TG	CTGA	AGGG	TGATGA	G TGCACATCAGG	CTGA	G CTTTATAACTT
Pho-sR08	GATTC GAA TAGGCG	ATGATGA	G CTCCATCCCTA C	CTGA	GTTG	TGATGA	A TGTAGCGCGC T	CTGA	GC AACCCTTATT
Pab-sR08	AAATTTGAA AGGCG	ATGATGA	G CTCCATCCCTA T	CTGA	GTTG	TGATGA	C TGTAGCGCGC T	CTGA	GC TACAGGCTCT
Pfu-sR08	GGGTTC AA TGAGCA	ATGATGA	G CTCTATCCCTA T	CTGA	CCCA	TGATGA	C TGTAGCGCGC G	CTGA	GC TACTCCTTTA
Pho-sR09	CGGGCCGACAG GGG	ATGAAGA	G CTTTTGCTTT G	CTGA	GCAGA	TGATGA	C CACGCCCTTCG	CTGA	CCTGCT ATTTGA
Pab-sR09	CCGGGCCTACA GTT	ATGATGA	A CTTTTGCTTT G	CTGA	TGTGG	TGATGA	G CACGCCCTTCG	CTGA	TACT CTCTCGTC
Pfu-sR09	GGGCCCACAAT GGG	ATGATGA	C CTTTTGCTTT A	CTGA	ACACA	TGATGA	C CACGCCCTTCG	CTGA	CC TAAATATTTG
Pho-sR10	ATCATTAGCGACT A	ATGATGA	A CTACTCCCGG	CTGA	GGGG	TGATGA	ACCACCTACCGG	CTGA	GGT GAAAGCATG
Pab-sR10	TTAAAACGGTA GC G	ATGATGA	A CTACTCCCGG	CTGA	GCTG	TGATGA	ACCACCTACCGG	CTGA	GGT GGAAACATG
Pfu-sR10	AAAAATTATG GCAG	ATGATGA	G CTACTCCCGG	CTGA	AAGA	TGATGA	ACCACCTACCGG	CTGA	GGT TATGGAAAA

Figure 5.8: **Alignment of 10 *Pyrococcus* sRNA homolog families.** Pho = *Pyrococcus horikoshii*, Pab = *Pyrococcus abyssi*, and Pfu = *Pyrococcus furiosus* sRNAs. Note strong conservation within box features and complementary regions, and lack of conservation in intervening and flanking sequences. Sequences are roughly aligned in flanking regions to show where conservation ends.

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.*, 25, 3389–3402. Available at http://www.ncbi.nlm.nih.gov/BLAST/.

Amiri, K. (1994). Fibrillarin-like proteins occur in the domain Archaea. *J Bacteriol*, 176, 2124–7.

Bachellerie, J. & Cavaille, J. (1997). Guiding ribose methylation of rRNA. *Trends Biochem Sci*, 22, 257–61.

Bachellerie, J. & Cavaille, J. (1998). Small nucleolar RNAs guide the ribose methylations of eukaryotic rRNAs. In H. Grosjean & R. Benne (Eds.), *Modification and Editing of RNA* (pp. 255–272). ASM Press.

Bachellerie, J.-P., Michot, B., Nicoloso, M., Balakin, A., Ni, J., & Fournier, M. J. (1995). Antisense snoRNAs: A family of nucleolar RNAs with long complementarities to rRNA. *Trends Biochem. Sci.*, 20, 261–264.

Bakin, A. & Ofengand, J. (1995). Mapping of the 13 pseudouridine residues in *Saccharomyces cerevisiae* small subunit ribosomal RNA to nucleotide resolution. *Nucl. Acids Res.*, 23, 3290–3294.

Balakin, A., Smith, L., & Fournier, M. (1996). The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell*, 86, 823–34.

Barrett, C., Hughey, R., & Karplus, K. (1997). Scoring hidden Markov models. *Comput. Applic. Biosci.*, 13, 191–199.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., & Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 21, 3329–30.

Beltrame, M. & Tollervey, D. (1995). Base pairing between U3 and the pre-ribosomal RNA is required for 18S rRNA synthesis. *EMBO J*, 14, 4350–6.

Bennetzen, J. & Hall, B. (1982). Codon selection in yeast. *J Biol Chem*, 257, 3026–31.

Benson, D., Boguski, M., Lipman, D., Ostell, J., Ouellette, B., Rapp, B., & Wheeler, D. (1999). Genbank. *Nucleic Acids Res*, 27, 12–7.

Billoud, B., Kontic, M., & Viari, A. (1996). Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res*, 24, 1395–403.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, 277, 1453–1462.

Boguski, M., Lowe, T., & Tolstoshev, C. (1993). dbEST – database for expressed sequence tags. *Nature Genet.*, 4, 332–333.

Brannan, C. I., Dees, E. C., Ingram, R. S., & Tilghman, S. H. (1990). The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, 10, 28–36.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S., & Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71, 515–526.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., & Willard, H. F. (1992). The human Xist gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71, 527–542.

Brown, J. (1999). The Ribonuclease P database. *Nucleic Acids Res*, 27, 314. Available at http://www.mbio.ncsu.edu/RNaseP/home.html.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., Venter, J. C., et al. (1996). Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. *Science*, 273, 1058–1073. Current annotation at http://www.tigr.org/tdb/.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268, 78–94.

Caffarelli, E., Losito, M., Giorgi, C., Fatica, A., & Bozzoni, I. (1998). In vivo identification of nuclear factors interacting with the conserved elements of box C/D small nucleolar RNAs. *Mol Cell Biol*, 18, 1023–8.

Cavaille, J. & Bachellerie, J. (1998). SnoRNA-guided ribose methylation of rRNA: structural features of the guide RNA duplex influencing the extent of the reaction. *Nucleic Acids Res*, 26, 1576–87.

Cavaille, J., Nicoloso, M., & Bachellerie, J. (1996). Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature*, 383, 732–5.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282, 2012–8.

Chamberlain, J., Lee, Y., Lane, W., & Engelke, D. (1998). Purification and characterization of the nuclear RNAse P holoenzyme complex reveals extensive subunit overlap with RNAse MRP. *Genes Dev*, 12, 1678–90.

Chanfreau, G., Legrain, P., & Jacquier, A. (1998a). Yeast RNAse III as a key processing enzyme in small nucleolar RNAs metabolism. *J Mol Biol*, 284, 975–88.

Chanfreau, G., Rotondo, G., Legrain, P., & Jacquier, A. (1998b). Processing of a dicistronic small nucleolar RNA precursor by the RNA endonuclease Rnt1. *EMBO J*, 17, 3726–37.

Crick, F. (1966). Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol*, 19, 548–55.

Dandekar, T. & Hentze, M. W. (1995). Finding the hairpin in the haystack: Searching for RNA motifs. *Trends Genet.*, 11, 45–50.

Daniels, G. R. & Deininger, P. L. (1985). Repeat sequence families derived from mammalian tRNA genes. *Nature*, 317, 819–822.

Deininger, P. L. (1989). SINEs: Short interspersed repeated DNA elements in higher eucaryotes. In D. E. Berg & M. M. Howe (Eds.), *Mobile DNA*. American Society for Microbiology.

Dong, H., Nilsson, L., & Kurland, C. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol*, 260, 649–63.

Dunbar, D. & Baserga, S. (1998). The U14 snoRNA is required for 2'-O-methylation of the pre-18S rRNA in *Xenopus* oocytes. *RNA*, 4, 195–204.

Dunbar, D., Wormsley, S., Lowe, T., & Baserga, S. (1999). Fibrillarin-associated box C/D snoRNAs in *Trypanosoma brucei*: sequence conservation and implications for 2'-O-ribose methylation of rRNA. Submitted.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge UK: Cambridge University Press.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6, 361–365.

Eddy, S. R. & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22, 2079–2088. COVE software available at http://www.genetics.wustl.edu/eddy/software/.

El-Mabrouk, N. & Lisacek, F. (1996). Very fast identification of RNA motifs in genomics DNA. application to tRNA search in the yeast genome. *J. Mol. Biol.*, (pp. 46–55).

Fichant, G. A. & Burks, C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, 220, 659–671.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496–512. Current annotation at http://www.tigr.org/tdb/.

Fraser, C., Gocayne, J., White, O., Adams, M., Clayton, R., Fleischmann, R., Bult, C., Kerlavage, A., Sutton, G., Kelley, J., Fritchman, J., Weidman, J., Small, K., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T., Saudek, D., Phillips, C., Merrick, J., Tomb, J.-F., Dougherty, D., Bott, K., Hu, P.-C., Lucier, T., Peterson, S., Smith, H., Hutchison, C., & Venter, J. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270, 397–403. Current annotation at http://www.tigr.org/tdb/.

Gannot, P., Bortolin, M.-L., & Kiss, T. (1997). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, 89, 799–809.

Gautheret, D., Major, F., & Cedergren, R. (1990). Pattern searching/alignment with RNA primary and secondary structures: An effective descriptor for tRNA. *Comput. Applic. Biosci.*, 6, 325–331.

Gautier, T., Berges, T., Tollervey, D., & Hurt, E. (1997). Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol Cell Biol*, 17, 7088–98.

Gish, W. (1998). WU-BLAST. Available from http://blast.wustl.edu/.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274, 546–567.

Grate, L. (1995). Automatic RNA secondary structure determination with stochastic context-free grammars. In C. Rawlings & others (Eds.), *Proc. Third Int. Conf. on Intelligent Systems in Molecular Biology* (pp. 136–144). Menlo Park, California: AAAI Press.

Green, C. & Vold, B. (1993). *Staphylococcus aureus* has clustered tRNA genes. *J Bacteriol*, 175, 5091–6.

Greenwood, S., Schnare, M., & Gray, M. (1996). Molecular characterization of U3 small nucleolar RNA from the early diverging protist, *Euglena gracilis*. *Curr Genet*, 30, 338–46.

Gribskov, M. (1994). Profile analysis. In A. Griffin & H. Griffin (Eds.), *Methods in Molecular Biology: Computer Analysis of Sequence Data Part II* (pp. 247–267). Totowa NJ: Humana Press.

Gribskov, M., Luthy, R., & Eisenberg, D. (1990). Profile analysis. *Meth. Enzymol.*, 183, 146–159.

Gutell, R. (1994). Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucl. Acids Res.*, 22, 3502–3507.

Gutell, R., Gray, M., & Schnare, M. (1993). A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. *Nucleic Acids Res*, 21, 3055–74.

Gutell, R. R. (1993). Collection of small subunit (16S and 16S-like) ribosomal RNA structures. *NAR*, 21, 3051–3054.

Guthrie, C. & Abelson, J. (1982). Organization and expression of tRNA genes in *Saccharomyces cerevisiae*. In J. Strathern, E. Jones, & J. Broach (Eds.), *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.

Guthrie, C. & Patterson, B. (1988). Spliceosomal snRNAs. *Ann. Rev. Genet.*, 22, 387–419.

Hadjiolov, A. (1985). *The Nucleolus and Ribosome Biogenesis*. New York: Springer–Verlag.

Hani, J. & Feldmann, H. (1998). tRNA genes and retroelements in the yeast genome. *Nucleic Acids Res*, 26, 689–96.

Hatlen, L. & Attardi, G. (1971). Proportion of HeLa cell genome complementary to transfer RNA and 5S RNA. *J Mol Biol*, 56, 535–53.

Hayes, W. & Borodovsky, M. (1998). How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res*, 8, 1154–71.

Henras, A., Henry, Y., Bousquet-Antonelli, C., Noaillac-Depeyre, J., Gelugne, J., & Caizergues-Ferrer, M. (1998). Nhp2p and Nop10p are essential for the function of H/ACA snoRNPs. *EMBO J*, 17, 7078–90.

Higgins, D., Bleasby, A., & Fuchs, R. (1992). ClustalV: Improved software for multiple sequence alignment. *Comput. Applic. Biosci.*, 8, 189–191.

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*, 158, 573–97.

Jarmolowski, A., Zagorski, J., Li, H., & Fournier, M. (1990). Identification of essential elements in U14 RNA of *Saccharomyces cerevisiae*. *EMBO J*, 9, 4503–9.

Johnson, P. & Abelson, J. (1983). The yeast tRNA-Tyr gene intron is essential for correct modification of its tRNA product. *Nature*, 302, 681–7.

Kano, A., Ohama, T., Abe, R., & Osawa, S. (1993). Unassigned or nonsense codons in *Micrococcus luteus*. *J Mol Biol*, 230, 51–6.

Keeney, J. B., Chapman, K. B., Lauermann, V., Voytas, D. F., Astrom, S. U., von Pawel-Rammingen, U., Bystrom, A., & Boeke, J. D. (1995). Multiple molecular determinants for retrotransposition in a primer tRNA. *Mol. Cell. Biol.*, 15, 217–226.

Kiss-Laszlo, Z., Henry, Y., Bachellerie, J., Caizergues-Ferrer, M., & Kiss, T. (1996). Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85, 1077–88.

Kiss-Laszlo, Z., Henry, Y., & Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J*, 17, 797–807.

Klootwijk, J. & Planta, R. J. (1973). Analysis of the methylation sites in yeast ribosomal RNA. *Eur. J. Biochem.*, 39, 325–333.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, 235, 1501–1531.

Krzyzosiak, W., Denman, R., Nurse, K., Hellmann, W., Boublik, M., Gehrke, C., Agris, P., & Ofengand, J. (1987). In vitro synthesis of 16S ribosomal RNA containing single base changes and assembly into a functional 30S ribosome. *Biochemistry*, 26, 2353–64.

Laferriere, A., Gautheret, D., & Cedergren, R. (1994). An RNA pattern matching program with enhanced performance and portability. *Comput Appl Biosci*, 10, 211–2.

Lafontaine, D. & Tollervey, D. (1998). Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem Sci*, 23, 383–8.

Leader, D., Clark, G., Watters, J., Beven, A., Shaw, P., & Brown, J. (1997). Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *EMBO J*, 16, 5742–51.

Levitan, A., Xu, Y., Ben-Dov, C., Ben-Shlomo, H., Zhang, Y., & Michaeli, S. (1998). Characterization of a novel trypanosomatid small nucleolar RNA. *Nucleic Acids Res*, 26, 1775–83.

Liang, W. & Fournier, M. (1995). U14 base-pairs with 18S rRNA: a novel snoRNA interaction required for rRNA processing. *Genes Dev*, 9, 2433–43.

Lisacek, F., Diaz, Y., & Michel, F. (1994). Automatic identification of group I intron cores in genomic DNA sequences. *J. Mol. Biol.*, 235, 1206–1217.

Lowe, T. & Eddy, S. (1999). A computational screen for methylation guide snoRNAs in yeast. *Science*, 283, 1168–71.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25, 955–964.

Lowe, T. M. & Eddy, S. R. (1998). The Eddy Lab snoRNA Database. Available at http://rna.wustl.edu/snoRNAdb/.

Maden, B. & Hughes, J. (1997). Eukaryotic ribosomal RNA: the recent excitement in the nucleotide modification problem. *Chromosoma*, 105, 391–400.

Maden, B. E. H. (1990). The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucl. Acids Res. Mol. Biol.*, 39, 241–303.

Maden, B. E. H., Corbett, M. E., Heeney, P. A., Pugh, K., & Ajuh, P. M. (1995). Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA. *Biochimie*, 77, 22–29.

Marvel, C. C. (1986). A program for the identification of tRNA-like structures in DNA sequence data. *Nucl. Acids Res.*, 14, 431–435.

Mattaj, I., Tollervey, D., & Seraphin, B. (1993). Small nuclear RNAs in messenger RNA and ribosomal RNA processing. *FASEB J*, 7, 47–53.

Mattaj, I. W. (1993). RNA recognition: A family matter? *Cell*, 73, 837–840.

Maxwell, E. & Fournier, M. (1995). The small nucleolar RNAs. *Annu Rev Biochem*, 6, 897–934.

Mizuta, K., Hashimoto, T., & Otaka, E. (1995). The evolutionary relationships between homologs of ribosomal YL8 protein and YL8-like proteins. *Curr Genet*, 28, 19–25.

Ni, J. (1998). *The Major Function of Eukaryotic Small Nucleolar RNAs is Nucleotide Modification in Ribosomal RNA.* PhD thesis, University of Massachusetts, Amherst.

Ni, J., Tien, A., & Fournier, M. (1997). Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, 89, 565–73.

Nicoloso, M., Caizergues-Ferrer, M., Michot, B., Azum, M., & Bachellerie, J. (1994). U20, a novel small nucleolar RNA, is encoded in an intron of the nucleolin gene in mammals. *Mol Cell Biol*, 14, 5766–76.

Nicoloso, M., Qu, L., Michot, B., & Bachellerie, J. (1996). 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, 178–95.

Noon, K. R., Bruenger, E., & McCloskey, J. A. (1998). Posttranscriptional modifications in 16S and 23S rRNAs of the Archaeal hyperthermophile *Sulfolobus solfataricus*. *J. Bacteriol.*, 180, 2883–2888.

Oba, T., Andachi, Y., Muto, A., & Osawa, S. (1991). CGG: an unassigned or nonsense codon in *Mycoplasma capricolum*. *Proc Natl Acad Sci U S A*, 88, 921–5.

Ofengand, J. & Fournier, M. (1998). The pseudouridine residues of rRNA: Number, location, biosynthesis, and function. In H. Grosjean & R. Benne (Eds.), *Modification and Editing of RNA* (pp. 229–254). ASM Press.

Paolella, G. & Russo, T. (1985). A microcomputer program for the identification of tRNA genes. *Comput Appl Biosci*, 1, 149–51.

Parker, R., Simmons, T., Shuster, E., Siliciano, P., & Guthrie, C. (1988). Genetic analysis of small nuclear RNAs in *Saccharomyces cerevisiae*: viable sextuple mutant. *Mol Cell Biol*, 8, 3150–9.

Pavesi, A., Conterlo, F., Bolchi, A., Dieci, G., & Ottonello, S. (1994). Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucl. Acids Res.*, 22, 1247–1256.

Pearson, W. & Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85, 2444–8.

Percudani, R., Pavesi, A., & Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol*, 268, 322–30.

Petfalski, E., Dandekar, T., Henry, Y., & Tollervey, D. (1998). Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. *Mol Cell Biol*, 18, 1181–9.

Planta, R. & Mager, W. (1998). The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast*, 14, 471–7.

Qu, L., Henras, A., Lu, Y., Zhou, H., Zhou, W., Zhu, Y., Zhao, J., Henry, Y., Caizergues-Ferrer, M., & Bachellerie, J. (1999). Seven novel methylation guide small nucleolar RNAs are processed

from a common polycistronic transcript by Rat1p and RNase III in yeast. *Mol Cell Biol*, 19, 1144–58.

Raue, H., Klootwijk, J., & Musters, W. (1988). Evolutionary conservation of structure and function of high molecular weight ribosomal RNA. *Prog Biophys Mol Biol*, 51, 77–129.

Reynolds, W. (1995). Developmental stage-specific regulation of *Xenopus* tRNA genes by an upstream promoter element. *J Biol Chem*, 270, 10703–10.

Rice, C. M., Fuchs, R., Higgins, D. G., Stoehr, P. J., & Cameron, G. N. (1993). The EMBL data library. *Nucl. Acids Res.*, 21, 2967–2971.

Riedel, N., Wise, J., Swerdlow, H., Mak, A., & Guthrie, C. (1986). Small nuclear RNAs from *Saccharomyces cerevisiae*: unexpected diversity in abundance, size, and molecular complexity. *Proc Natl Acad Sci U S A*, 83, 8097–101.

Roberts, T., Sturm, N., Yee, B., Yu, M., Hartshorne, T., Agabian, N., & Campbell, D. (1998). Three small nucleolar RNAs identified from the spliced leader-associated RNA locus in kinetoplastid protozoans. *Mol Cell Biol*, 18, 4409–17.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C., & Haussler, D. (1994a). Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, 22, 5112–5120.

Sakakibara, Y., Brown, M., Underwood, R. C., Mian, I. S., & Haussler, D. (1994b). Stochastic context-free grammars for modeling RNA. In L. Hunter (Ed.), *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences: Biotechnology Computing*, volume V (pp. 284–293). Los Alamitos, CA: IEEE Computer Society Press.

Salzberg, S., Delcher, A., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated markov models. *Nucleic Acids Res*, 26, 544–8.

Samarsky, D. & Fournier, M. (1999). A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 27, 161–164. Available from http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html.

Saurin, W. & Marliere, P. (1987). Matching relational patterns in nucleic acid sequences. *Comput. Applic. Biosci.*, 3, 115–120.

Schiestl, R., Manivasakam, P., Woods, R., & Gietz, R. (1993). Introducing DNA into yeast by transformation. *Methods*, 4, 79–85.

Sharp, P., Stenico, M., Peden, J., & Lloyd, A. (1993). Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans*, 21, 835–41.

Shaw, P., Beven, A., Leader, D., & Brown, J. (1998). Localization and processing from a polycistronic precursor of novel snoRNAs in maize. *J Cell Sci*, 111, 2121–8.

Shortridge, R., Pirtle, I., & Pirtle, R. (1986). IBM microcomputer programs that analyze DNA sequences for tRNA genes. *Comput Appl Biosci*, 2, 13–7.

Sibbald, P., Sommerfeldt, H., & Argos, P. (1992). Overseer: a nucleotide sequence searching tool. *Comput Appl Biosci*, 8, 45–8.

Smith, C. M. & Steitz, J. A. (1997). Sno storm in the nucleolus: New roles for myriad small RNPs. *Cell*, 89, 669–672.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G. M., Daniels, C. J., Mao, J., Rice, P., Nolling, J., & Reeve, J. N. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. *J Bacteriol.*, 179, 7135–7155.

Staden, R. (1980). A computer program to search for tRNA genes. *Nucl. Acids Res.*, 8, 817–825.

Staden, R. (1988). Methods to define and locate patterns of motifs in sequences. *Comput. Applic. Biosci.*, 4(1), 53–60.

Steinberg, S., Misch, A., & Sprinzl, M. (1993). Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, 21, 3011–3015. Database available at http://www.uni-bayreuth.de/departments/biochemie/trna/.

Stenico, M., Lloyd, A., & Sharp, P. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res*, 22, 2437–46.

Strobel, M. & Abelson, J. (1986). Effect of intron mutations on processing and function of *Saccharomyces cerevisiae* SUP53 tRNA in vitro and in vivo. *Mol Cell Biol*, 6, 2663–73.

Tateno, Y. & Gojobori, T. (1997). DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res*, 25, 14–7.

Tollervey, D. & Kiss, T. (1997). Function and synthesis of small nucleolar RNAs. *Curr Opin Cell Biol*, 9, 337–42.

Tollervey, D., Lehtonen, H., Carmo-Fonseca, M., & Hurt, E. (1991). The small nucleolar RNP protein NOP1 (fibrillarin) is required for pre-rRNA processing in yeast. *EMBO J*, 10, 573–83.

Tollervey, D., Lehtonen, H., Jansen, R., Kern, H., & Hurt, E. C. (1993). Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. *Cell*, 72, 443–457.

Tranquilla, T., Cortese, R., Melton, D., & Smith, J. (1982). Sequences of four tRNA genes from *Caenorhabditis elegans* and the expression of *C. elegans* tRNALeu (anticodon IAG) in *Xenopus* oocytes. *Nucleic Acids Res*, 10, 7919–34.

Tycowski, K., Smith, C., Shu, M., & Steitz, J. (1996). A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in *Xenopus*. *Proc Natl Acad Sci U S A*, 93, 14480–5.

Tycowski, K., You, Z., Graham, P., & Steitz, J. (1998). Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol Cell*, 2, 629–38.

Veldman, G., Klootwijk, J., de Regt, V., Planta, R., Branlant, C., Krol, A., & Ebel, J. (1981). The primary and secondary structure of yeast 26S rRNA. *Nucleic Acids Res*, 9, 6935–52.

Watkins, N., Gottschalk, A., Neubauer, G., Kastner, B., Fabrizio, P., Mann, M., & Luhrmann, R. (1998a). Cbf5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA-binding protein, are present together with Gar1p in all H box/ACA-motif snoRNPs and constitute a common bipartite structure. *RNA*, 4, 1549–68.

Watkins, N., Newman, D., Kuhn, J., & Maxwell, E. (1998b). In vitro assembly of the mouse U14 snoRNP core complex and identification of a 65-kDa box C/D-binding protein. *RNA*, 4, 582–93.

Weinstein, L. & Steitz, J. (1999). Guided tours: from precursor snoRNA to functional snoRNP. *Curr Opin Cell Biol*, 11, 378–84.

Westaway, S. & Abelson, J. (1995). Splicing of tRNA precursors. In D. Soll & U. L. RajBhandary (Eds.), *tRNA: Structure, Biosynthesis, and Function* (pp. 79–92). ASM Press.

Wheelan, S. J. & Boguski, M. S. (1998). Late-night thoughts on the sequence annotation problem. *Genome Res.*, 8, 168–169.

Wilson, E., Larson, D., Young, L., & Sprague, K. (1985). A large region controls tRNA gene transcription. *J Mol Biol*, 183, 153–63.

Woese, C., Kandler, O., & Wheelis, M. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A*, 87, 4576–9.

Woolford, J. (1991). The structure and biogenesis of yeast ribosomes. *Adv Genet*, 2, 63–118.

Wozniak, P. & Makalowski, W. (1990). Searching for tRNA genes in DNA sequences–an IBM microcomputer program. *Comput Appl Biosci*, 6, 49–50.

Wu, P., Brockenbrough, J., Metcalfe, A., Chen, S., & Aris, J. (1998). Nop5p is a small nucleolar ribonucleoprotein component required for pre-18 S rRNA processing in yeast. *J Biol Chem*, 273, 16453–63.

Young, L., Takahashi, N., & Sprague, K. (1986). Upstream sequences confer distinctive transcriptional properties on genes encoding silkgland-specific tRNA-Ala. *Proc Natl Acad Sci U S A*, 83, 374–8.

Yu, Y., Shu, M., & Steitz, J. (1997). A new method for detecting sites of 2'-O-methylation in RNA molecules. *RNA*, 3, 324–31.

Zagorski, J., Tollervey, D., & Fournier, M. (1988). Characterization of an SNR gene locus in *Saccharomyces cerevisiae* that specifies both dispensible and essential small nuclear RNAs. *Mol Cell Biol*, 8, 3282–90.