

# Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs

Wendy M. Olivas, Denise Muhrad and Roy Parker\*

Department of Molecular and Cellular Biology and Howard Hughes Medical Institute, University of Arizona, Tucson, AZ 85721, USA

Received July 10, 1997; Revised and Accepted October 8, 1997

## ABSTRACT

The genome sequences from increasing numbers of organisms allow for rapid and organized examination of gene expression. Yet current computational-based paradigms for gene recognition are limited and likely to miss genes expressing non-coding RNAs or mRNAs with small open reading frames (ORFs). We have utilized two strategies to determine if there are additional transcripts in the yeast *Saccharomyces cerevisiae* that were not identified in previous analyses of the genome. In one approach, we identified strong consensus polymerase III promoters based on sequence, and determined experimentally if these promoters drive the expression of an RNA polymerase III transcript. This approach led to the identification of a new, non-essential 170 nt non-coding RNA. An alternative strategy analyzed RNA expression from large sequence gaps >2 kb between predicted ORFs. Fifteen unique RNA transcripts ranging in size from 161 to 1200 nt were identified from a total of 59 sequence gaps. Several of these RNAs contain unusually small potential ORFs, while one is clearly non-coding and appears to be a small nucleolar RNA. These results suggest that there are likely to be additional previously unidentified non-coding RNAs in yeast, and that new paradigms for gene recognition will be required to identify all expressed genes from an organism.

## INTRODUCTION

The genome sequencing projects provide an opportunity to identify the complete spectrum of genes in an organism and to utilize that information to understand gene expression and function. A key step in this process will be to determine the chromosomal locations of genes and the diversity of gene products. Current approaches primarily focus on identifying gene locations based on open reading frame criteria. For example, in the yeast *Saccharomyces cerevisiae*, the first eukaryote whose genome has been completely sequenced, 5885 genes containing open reading frames (ORFs) are predicted (1). However, genes that express transcripts that either contain short ORFs (<300 nt)

or are non-coding RNAs would be missed in this type of sequence analysis and would need to be identified by other means.

Non-coding RNAs, or ncRNAs, have received considerable attention in recent years as it has become apparent that there is a striking diversity of these molecules in all cell types. Much work has focused on understanding the roles of RNAs in RNase P (2) and telomerase (3) action, of small nuclear RNAs (snRNAs) in mRNA splicing (4,5), and of small nucleolar RNAs (snoRNAs) in rRNA maturation (6–8). Yet ncRNAs have also been implicated in the processes of transcription (9), translation (10), transport (11), RNA editing (12), mRNA stability (13), differentiation (14) and protein degradation (15). For example, the Xist ncRNA of mammals is critical for X chromosome inactivation in females (16), while the *Drosophila* roX1 ncRNA localizes to the male X chromosome and is potentially involved in dosage compensation (17). In *Caenorhabditis elegans*, the *lin-4* ncRNA negatively regulates translation of the Lin-14 protein by duplex formation at repeated sequence elements within the *lin-14* RNAs 3' untranslated region (UTR) (18,19). The 10Sa RNA of *Escherichia coli* contains both transfer and messenger domains that enable the RNA to associate with ribosomes and tag aberrant polypeptides with a degradation signal (15). In addition, the *bic* RNA of birds is preferentially activated in metastatic tumors in conjunction with *c-myc* activation, suggesting that *bic* RNA may collaborate with *c-myc* in late stages of tumor progression (20).

The wide range of roles played by these ncRNAs and the large number of organisms in which they have been detected suggests that eukaryotic cells are likely to contain several ncRNAs that have yet to be identified. Moreover, the roles of these new ncRNAs will need to be determined. The yeast *S.cerevisiae* is an ideal model system for this type of analysis due to the availability of the genome sequence and the ability to easily test gene function by genetic analysis. For these reasons, we have asked if there are additional unidentified and unpredicted RNAs in yeast that might have been missed in ORF-based methods of genome analysis. Specifically, we utilized two strategies. One strategy assayed RNA expression from consensus polymerase III promoters, while the second strategy analyzed RNA expression from genomic sequences lacking predicted ORFs. We report here the identification of 16 unique RNA transcripts. Sequence analysis of these transcripts reveal ncRNAs and potential mRNAs that contain small unpredicted ORFs. These results provide evidence that

\*To whom correspondence should be addressed. Tel: +1 520 621 9347; Fax: +1 520 621 4524; Email: rrparker@u.arizona.edu

yeast and likely other organisms express many more RNAs than previously predicted.

## MATERIALS AND METHODS

### RNA polymerase III promoter analysis

Oligonucleotides were designed to anneal to RNA just 3' of the B box sequence from each candidate promoter. These oligonucleotides were end labeled with [ $\gamma$ - $^{32}$ P]ATP. Total yeast RNA was isolated as previously described (21) from yeast strain yRP683 (*MATa*, *leu2*, *lys2*, *his4*, *trp1*, *ura3*) grown to mid-log phase in rich (YEPD) media at 14, 30 and 37°C or in YEP + 3% glycerol at 30°C. Northern blot analysis was performed by loading 40  $\mu$ g of RNA in each lane of 1.25% agarose/6.7% formaldehyde gels or 6% polyacrylamide/7 M urea gels, blotting to Zeta-probe (BioRad), and hybridizing to the radiolabeled oligonucleotides 15 h at 42°C. The resulting blots were washed at 50°C and imaged using a Molecular Dynamics PhosphorImager.

### Gap analysis

Chromosomal sequence locations of all known and predicted ORFs within the yeast genome were obtained from the MIPS (Munich Information Centre for Protein Sequences) computer database. Gap regions  $\geq 2$  kb between ORFs were amplified by PCR for 30 cycles with primers designed with 5' restriction sites to anneal ~150 nt away from each flanking 5' and 3' ORF. PCR products were radiolabeled by random priming and extending with [ $\alpha$ - $^{32}$ P]dATP. Northern blot analysis of total yeast RNA was performed as described above using the radiolabeled PCR products as hybridization probes.

### Mapping of RNAs

PCR products of gap regions that expressed RNAs on northern blots were cloned into pBluescript (Stratagene), then the gap sequences were digested into at least three fragments. Each fragment was random-prime radiolabeled with [ $\alpha$ - $^{32}$ P]dATP and used to probe northern blots of total yeast RNA as described above. Once an expressed RNA was isolated to a distinct restriction fragment, oligonucleotides were designed complementary to both Watson and Crick strands within the restriction fragment. These oligonucleotides were end labeled with [ $\gamma$ - $^{32}$ P]ATP and used in northern blot analyses as described above. The 5'-3' orientation of an expressed RNA was determined as the complement of the oligonucleotide to which it hybridized. These complementary oligonucleotides were then used in primer extension reactions to determine the 5'-end of each RNA. The oligonucleotide that annealed to RNA170 from the polymerase III promoter analyses, was also used for primer extension of RNA170. Primer extension was performed by annealing 10  $\mu$ g of total yeast RNA with  $1 \times 10^6$  c.p.m. of oligonucleotide and extending with avian myeloblastosis virus reverse transcriptase (AMV-RT) as described in Current Protocols (22). Extension products were purified by phenol:CHCl<sub>3</sub> extraction and ethanol precipitation, then electrophoresed on a 6% polyacrylamide/7 M urea sequencing gel. Dideoxy sequencing reactions of the respective gap fragments using the same oligonucleotides as from primer extension analysis were run beside extension products to determine the 5' nucleotide(s).

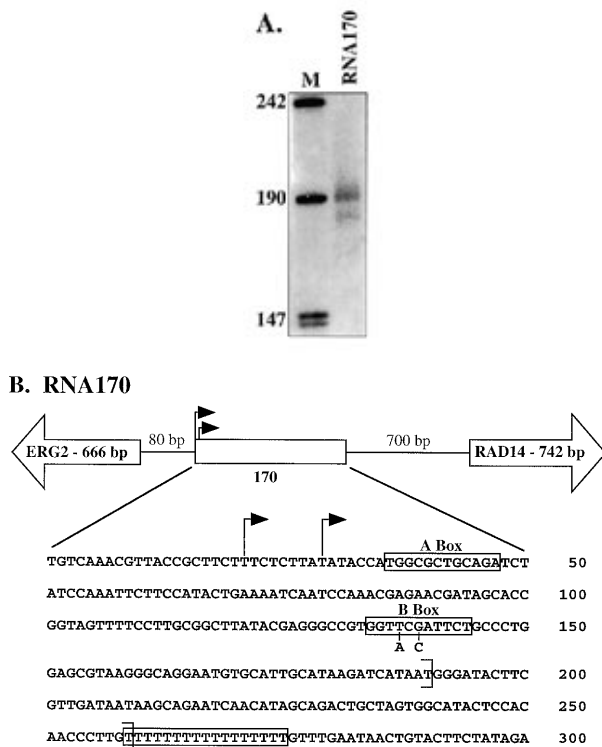
RNase protection and RNase H cleavage methods were utilized to estimate the 3'-ends of the RNAs. RNase protections were performed with RNase One (Promega) as per the manufacturer's instructions by the use of radiolabeled antisense transcripts produced by T7 or T3 transcriptions with [ $\alpha$ - $^{32}$ P]UTP from plasmids containing the gap sequences or the RNA170 gene. Protection products were purified by ethanol precipitation and analyzed on polyacrylamide gels as described above. RNase H cleavage reactions were performed as previously described (23) with 10  $\mu$ g of total yeast RNA and 500 ng of oligo dT or an oligonucleotide complementary to the identified RNA. Cleavage products were purified by phenol:CHCl<sub>3</sub> extraction and ethanol precipitation, then electrophoresed on 6% polyacrylamide/7 M urea gels that were blotted and probed with radiolabeled gap restriction fragments. The precise 3' nucleotides of RNA170 and RNA161 were determined by a ligase-mediated RT-PCR method. T4 RNA ligase was used to ligate a DNA oligonucleotide onto the 3'-end of a gel-purified RNA. The product was reverse transcribed using a primer complementary to the DNA oligonucleotide, PCR amplified using the reverse transcription primer and an oligonucleotide homologous to an internal RNA sequence, cloned into pBluescript and sequenced.

### Phenotype analysis of RNAs

Deletions of RNA716, RNA515, RNA530 and RNA487 were made by PCR amplifying ~500 bp sequences flanking the RNA genes and cloning these on either side of a *URA3* gene inserted into the *Bam*HI site of pBluescript. A *Sac*I-*Kpn*I digest of this plasmid was used to transform haploid cells of yRP683 using the LiOAc method (24). Deletion of RNA161 was accomplished by transformation of a PCR product that contains the neomycin (*neo*) gene, which confers resistance to the drug G418. Specifically, PCR primers were designed at their 5'-ends with homology to 50 bp sequences flanking the RNA gene, and at their 3'-ends with ~20 bp homologous to sequences flanking the *neo* gene. The *neo* gene was amplified from plasmid pRP665, in which expression of *neo* is under the control of the GPD promoter (25) and the terminator sequence from the *PGK1* gene (26). The PCR product was then transformed into yRP683 haploid cells using the LiOAc method. Deletion of RNA170 was done by amplifying ~500 bp sequences flanking the RNA gene and cloning these on each side of either the *URA3* or *neo* gene inserted into the *Bam*HI site of pBluescript. A *Sac*I-*Kpn*I digest of this plasmid was used to transform haploid cells of yRP683 using the LiOAc method.

Overexpression of the above RNAs was accomplished by cloning respective restriction fragments into the polylinker of p426 (27), a 2 $\mu$  vector containing the *URA3* gene in a pBluescript backbone. These plasmids were transformed into yRP683 using the LiOAc method. Levels of overexpression were examined by northern blot analysis as described in the above sections. A 2 $\mu$  plasmid expressing RNA170 with mutations in its B Box was made by PCR amplifying the gene using primers containing the mutated nucleotides (Fig. 1B) and cloning the products into p426. Plasmids expressing TDS4 were generously donated by Stephen Buratowski.

Deletion and overexpression strains were streaked for single colonies on a range of media (YEPD, minimal, synthetic complete) with various nutrient (sucrose, acetate, galactose, raffinose, succinate, glycerol, maltose, lactic acid, ethanol), salt (450 or 900 mM NaCl) or drug (hydroxyurea, methylmethanesulfonate,



**Figure 1.** Identification and mapping of a new RNA polymerase III transcript. (A) Northern blot analysis shows the detection of RNAs expressed from a polymerase III promoter on chromosome XIII, denoted as RNA170. Sizes of DNA bands in the marker lane (M) are labeled in nucleotides. Total yeast RNA was electrophoresed in a polyacrylamide gel, blotted and hybridized to an oligonucleotide complementary to the polymerase III transcript. (B) Mapping of RNA170. A segment of chromosome XIII is depicted to show the orientation of the expressed RNA170. Large open arrows are labeled for the genes flanking the RNA gene, while the rectangle represents the coding region of RNA170. Sizes of the genes and distances between genes are indicated. Filled arrows denote major transcription start sites of the RNA. The schematics are not drawn to scale. In the RNA170 gene sequence, arrows denote transcription start sites and brackets denote 3'-ends. The A Box sequence, B Box sequence and poly-T tract are boxed. Point mutations made in the B Box are indicated.

cycloheximide) conditions; grown at 14, 30 or 37°C; and monitored for differences in colony size. Deletion strains were mated with strain yRP684 (isogenic to the parental yRP683 strain) on YEPD agar. Diploid cells were starved to promote sporulation. The resulting tetrads were dissected and monitored for growth differences.

## RESULTS AND DISCUSSION

### Identification of a new 170 nt RNA polymerase III transcript

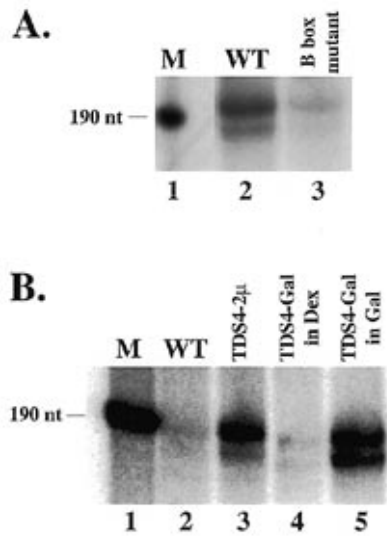
The first method we used to search for functional RNAs was designed to specifically identify new RNA polymerase III transcripts. Such RNA transcripts are typically not translated and function in various cellular processes such as translation and RNA processing. Promoters of polymerase III genes are characterized by two conserved domains (28). Given this organization, we used the conserved A and B box domains of the tRNA type 2 promoter to search for new polymerase III

promoters in the yeast genome. Specifically, a computer search for the consensus B box sequence (GTTCRANYC) was performed allowing only one mismatch. The resulting list of candidates was then searched for at least 50% conservation of the A box sequence (TRGCNNAGYNGG) within 300 nt upstream of the B box, and a poly-T termination signal prior to the next downstream gene. Sequences located within known or predicted ORF regions were eliminated, as well as sequences that were, or had homology to, known polymerase III transcripts. For the 10 candidates that met the above criteria, oligo hybridization probes located just 3' of the B box were used to probe northern blots to determine if any RNA transcripts were expressed. In this analysis one probe identified an RNA doublet near 190 nt (Fig. 1A).

The transcribed region corresponding to this RNA was determined by a number of techniques and is shown in Figure 1B. Primer extension reactions revealed two major transcriptional start sites for the RNA, located 7 and 15 bases upstream of the A Box. These distances are typical for tRNA polymerase III promoters. Surprisingly, two 3'-ends were determined by sequencing cDNAs (see Materials and Methods). A minor end was identified at a poly-T tract that presumably corresponds to an RNA polymerase III transcriptional terminator. The major 3'-end was identified as being located 70 nt upstream of this large poly-T tract. These results raise the possibility that the mature transcripts are produced by an RNA processing event from the primary transcripts that arise by termination at the poly-T tract. Together, the mapping data reveal that the mature transcripts deriving from the two major start sites are 170 and 162 nt in length. The difference in size between the RNA doublet visualized on northern blots (190 and 183 nt) and the mapping data may reflect gel mobility anomalies of the RNA due to strong structural elements within the transcript.

Two observations provide experimental evidence that these transcripts, termed together as RNA170, are transcribed by RNA polymerase III. First, when wild-type or mutant copies of the RNA170 gene are introduced into yeast on plasmids, point mutations within the consensus B box element (Fig. 1B) decrease expression to the levels seen in an untransformed wild-type strain (Fig. 2A). In addition, expression levels of the RNA increased >5-fold in yeast strains that overexpressed TDS4 (Fig. 2B), a limiting component of the polymerase III transcription machinery (29). Consistent with RNA170 being produced by RNA polymerase III, RNase H cleavage of the RNA with oligo dT showed that RNA170 is not polyadenylated (data not shown).

As a first step towards determining the function of RNA170, we constructed yeast strains that either overexpress or are deleted for the RNA170 gene. Overexpression of RNA170 from high copy plasmids in the presence of increased TDS4 protein led to yeast strains that produce ~10-fold more RNA170, but show no obvious phenotype under a variety of conditions (see Materials and Methods). In order to create a null mutation in this gene, the entire RNA170 transcript-coding region was replaced with either the neomycin gene or *URA3* (see Materials and Methods). Strains carrying the *rna170Δ* are viable and produce no detectable RNA170, indicating that the RNA is not essential for growth under a variety of conditions (see Materials and Methods). Although not essential, low stringency northern blots of total RNA from a variety of related yeast species (*Saccharomycopsis capsularis*, *Saccharomyces kluyveri*, *Schizosaccharomyces pombe*, *Pichia canadiensis* and *Zygosaccharomyces florentinus*) using an RNA170 probe identified cross hybridizing transcripts



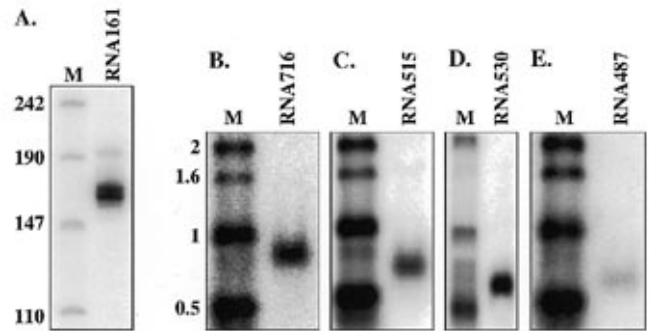
**Figure 2.** Northern blot analysis of RNA170. (A) Inhibition of RNA170 expression by B Box mutations. Wild-type RNA (lane 2) or RNA with point mutations in the B Box (see Fig. 1B; lane 3) are expressed from a 2 $\mu$  plasmid in a wild-type background yeast strain. DNA markers (M) are shown in lane 1. (B) Overexpression of RNA170 in the presence of the polymerase III component, TDS4. Wild-type RNA is expressed from a 2 $\mu$  plasmid in a wild-type background yeast strain (lane 2) in the presence of TDS4 from a 2  $\mu$  plasmid (lane 3) or a galactose-driven TDS4 grown in dextrose (lane 4) or galactose (lane 5). DNA markers (M) are shown in lane 1.

of between 125 and 275 nt in several species (data not shown), suggesting that RNA170 may be evolutionarily conserved.

Knowing that RNA170 is a potential RNA polymerase III transcript, this RNA could be involved in any number of cellular functions. RNA polymerase III transcribes both nuclear and cytoplasmic RNAs involved in a variety of cellular processes. For example, tRNAs and 5S RNA are involved in translation, 7SL RNA is a component of the signal-recognition particle (30), RNase P functions in tRNA processing (31) and U6 RNA is involved in mRNA splicing (4). In addition, Y RNAs are a class of small RNAs, originally characterized in vertebrates, that fold into distinct stem-loop structures and are associated with the Ro 60 kDa protein, forming the cytoplasmic Ro ribonucleoprotein complex of unknown function (32). The discovery of Y RNAs in *C.elegans* (33) suggests that these RNAs may be conserved in other eukaryotes, including yeast. Although RNA170 shows no significant sequence similarity to Y RNAs or other known polymerase III transcripts, it shows 61% sequence identity over 141 nt to a sequence of unknown function from a cosmid of *C.elegans*.

### Analysis of genome gaps

A second strategy we used to identify new RNA transcripts took advantage of the observation that the yeast genome has a very compact distribution of genes (1). In fact, the majority of predicted ORFs in yeast are oriented <1 kb apart, allowing adequate sequence space for promoters and 5'-3' UTRs. This dense packing of genes suggests that the rare  $\geq 2$  kb sequence gaps that are located between some ORFs are not simply random nucleotides, but are functionally important, possibly encoding gene products. An example of this arrangement is a 2.1 kb gap



**Figure 3.** Identification of RNA transcripts expressed from chromosomal gaps. Shown are examples of five northern blots (A-E) representing detection of RNAs expressed from gap regions. RNAs are denoted as RNAX, where X is the size of the transcript. (A) displays total yeast RNA electrophoresed in a polyacrylamide gel, while (B-E) display total yeast RNA electrophoresed in agarose gels. All gels were blotted and hybridized to a radiolabeled gap sequence. Sizes of bands in DNA marker lanes (M) are labeled in nucleotides for A, and in kilobases for B-E.

between predicted ORFs on chromosome II that contains the gene for the untranslated TLC1 RNA, a component of telomerase (34). To test whether other gaps also express RNAs, we examined several large gaps in the yeast genome by northern analysis using probes specific to those regions. In contrast to the promoter based strategy described above, this approach should identify transcripts expressed by any of the RNA polymerases and could also identify mRNAs which were not predicted due to the small size of their ORFs.

Our analysis proceeded in the following steps. First, computer searches of all 16 yeast chromosomes identified a total of 58 sequence gaps  $\geq 2$  kb located between known and hypothetical ORFs (Table 1). For the purposes of this work, we have defined gaps as the first nucleotide downstream of an ORFs start/stop codon to the last nucleotide upstream of the next ORFs start/stop codon. Telomeric and centromeric regions contain numerous non-ORF elements and so were avoided in our analysis. Polymerase chain reaction (PCR) amplification of each gap region was performed using primers that annealed  $\sim 150$  nt away from flanking ORFs. This distance should avoid overlap with the vast majority of 5' and 3' UTRs of the flanking ORFs. PCR products were then radiolabeled and used as hybridization probes for northern blot analysis of total yeast RNA. RNA was prepared from cells growing under various conditions (see Materials and Methods). As a positive control for this method, a PCR probe made from the 2.1 kb gap of chromosome II containing the TLC1 RNA was used. Surprisingly, the TLC1 gap probe not only detected the expected 1300 nt RNA transcript, but also hybridized to a unique 161 nt doublet (Fig. 3A). Sequence analysis of this RNA indicated it is likely to be a new snoRNA (see below). From the 58 identified sequence gaps, 14 new RNA transcripts ranging in size from 450 to 1200 nt were found, thus, with the 161 nt transcript, giving a total number of 15 new RNA transcripts (Table 1). Examples of these RNAs are shown in Figure 3B-E. Together, >20% of the  $\geq 2$  kb gaps expressed RNAs, with some gaps expressing two or three unique transcripts.

**Table 1.** Gaps  $\geq 2$  kb between known and predicted yeast ORFs

Chromosome	Gap size (kb)	Gap location	RNA from gap (nt)
I	3.6	27969–31573	-
	2.3	177008–179269	-
	2.3	69531–71798	-
II	2.1	163956–166091	-
	TLC1-2.1	306911–303084	1300; 161
	2.4	340707–343055	-
	2.3	362470–364744	-
	3.3	416857–420156	716
	2.2	554263–556502	-
III	2.5	28920–31433	-
IV	3.1	80414–83548	-
	3.4	504735–508185	~700
	3.3	543410–546681	~500; ~800; ~1200
	2.3	793928–796233	-
V	1.9	173338–175246	-
	2.4	259639–262053	515
VI	NONE		
VII	3.0	73900–76890	-
	2.3	165096–167355	-
	2.1	323226–325328	-
	2.5	420553–423088	~1000 <sup>a</sup>
	2.0	809415–811440	~700
	2.1	993517–995633	~800 <sup>b</sup>
	3.3	201302–204597	-
VIII	3.6	521733–525386	-
	2.4	336895–339340	-
IX	4.2	385344–389568	-
	2.7	394555–397290	-
	2.2	179801–181998	-
X	2.6	605345–607997	-
	2.6	150692–153274	-
XI	3.4	229526–231873	-
	1.6	259173–260777	-
	2.1	302668–304759	-
	2.0	307862–309844	-
	2.5	321163–322873	-
	2.1	92647–94745	-
	2.0	107899–109903	-
	1.9	223060–224923	-
XII	1.9	474058–475973	-
	1.9	489927–491869	-
	1.9	758833–760750	-
	2.0	949184–951148	-
	1.9	305593–307487	-
	1.9	370517–372443	-
	2.7	432125–434786	~700; ~1000
	2.3	480693–483012	~600
XIII	2.5	511075–513591	530
	1.7	334864–336543	-
	1.8	355042–356795	-
	1.7	691281–693018	487
XIV	1.8	724303–726129	-
	2.7	346949–349676	-
	1.9	384931–386822	-
	2.3	436346–438642	-
XV	2.6	461813–464448	-
	1.9	672409–674351	-
	2.1	213962–216013	-
	2.8	587514–590280	-
XVI	2.0	769657–771649	-

<sup>a</sup>RNA most prevalent at 30°C.<sup>b</sup>RNA most prevalent at 14°C.

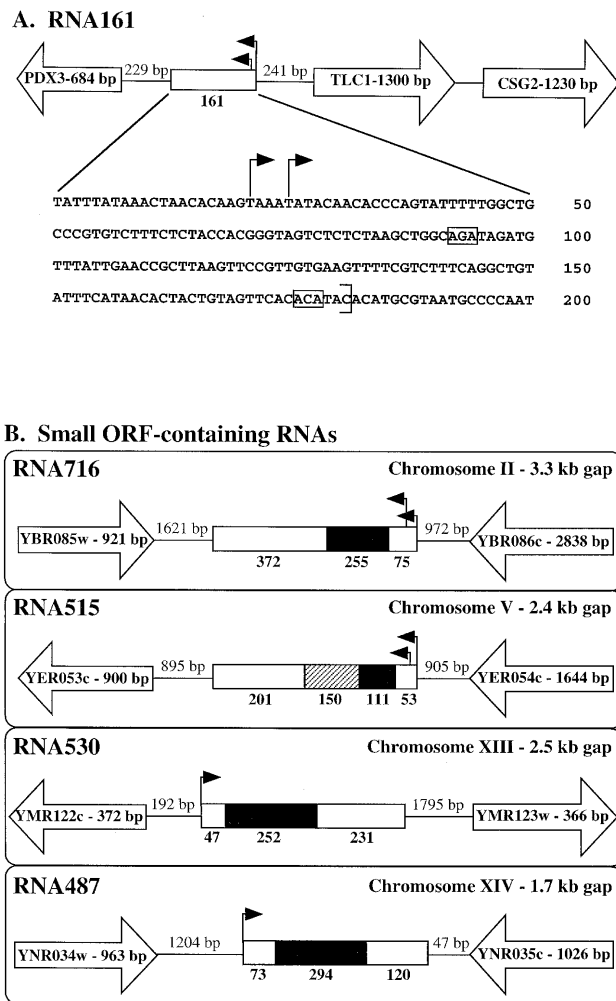
## RNA mapping

A variety of techniques was employed to map the location of the 15 unique RNAs identified in the northern analysis described above. First, all RNAs were approximately mapped by hybridizing northern blots with probes derived from different restriction fragments from the respective gap regions. In all cases, the RNA was localized to a distinct region within the gap. This observation indicated that these RNAs are not derived from the neighboring mRNAs of predicted or known ORFs. Five transcripts were then analyzed further. These included the 161 nt doublet RNA expressed from near the TLC1 gene, termed RNA161, and four of the more abundant RNAs expressed from other gaps (RNA716, RNA515, RNA530 and RNA487, where each number represents the transcript size). These RNAs were precisely mapped using a combination of primer extension, RNase protection, RNase H cleavage and cloning of RT-PCR products. The results of these analyses are shown in Figure 4.

Sequence analysis of RNA161 provided evidence that this transcript may be a snoRNA. There are two major classes of snoRNAs: those that contain two consensus sequence elements termed boxes C and D, and those that contain an ACA triplet located 3 nt from the RNA 3'-end (35). Many snoRNAs of the ACA class also contain an AGA triplet positioned between 5' and 3' stem-loop structures (35). RNA161 contains an ACA triplet 3 nt from its 3'-end (Fig. 4A). Moreover, computer folding programs predict stable 5' and 3' stem-loop structures on either side of an AGA triplet within the RNA. RNase H cleavage of RNA161 with oligo dT showed that the RNA is not polyadenylated. From this evidence, it is likely that this 161 nt transcript is a new member of the ACA snoRNA family. Overexpression of the RNA polymerase III component TDS4 had no effect on RNA161 expression, suggesting that the RNA is instead transcribed by RNA polymerase II (data not shown).

An important question for the other four transcripts is whether or not these transcripts are ncRNAs or mRNAs. The mapping of RNA515 is unusual in this regard. In this case, the transcribed region contains a very small potential ORF of 37 aa near the 5'-end (Fig. 4B). Following the UGA stop codon of this ORF, there is another 50 aa before the next in-frame stop codon (Fig. 4B, hatched box). It is possible that this stop codon is read through to produce an 87 aa ORF. If so, then this would be the first example of stop codon read-through in yeast. We do not believe this stop codon is an artifact of a mutant allele only found in our yeast strain, because the stop codon is also present in the *S.cerevisiae* genome database. In support of the role of RNA515 as an mRNA, the RNA is polyadenylated based on RNase H cleavage with oligo dT. Alternatively, this transcript may function as a ncRNA. As a precedence, the Xist RNA of mammals and the roX1 RNA of *Drosophila* are both large polyadenylated RNAs that are believed to act as ncRNAs.

The structure of RNA716, RNA530 and RNA487 is most consistent with those transcripts functioning as mRNAs and encoding small polypeptides. For each of these RNAs, a potential ORF of 84–98 aa is found within the RNA (Fig. 4B). Moreover, in each case the ORF is located near the 5'-end of the RNA. Since these potential ORFs are all <100 aa, they would not have been predicted by the standard computational analysis where the minimum ORF length is set at 100 aa. Consistent with the possibility that these transcripts are mRNAs, the RNAs also have poly(A) tails based on RNase H cleavage with oligo dT.



**Figure 4.** Mapping of new gap RNAs. (A) and (B) Segments of chromosomes are depicted to show the orientation of the expressed RNAs. Large open arrows are labeled for the ORFs or known genes flanking the RNA genes. Rectangles represent the expressed RNA coding regions, with black boxes within the rectangles denoting potential ORFs. The hatched box of RNA515 represents the additional sequence region that could be translated if the UGA stop codon of the 111 nt ORF is read through. Sizes of the RNA genes and elements within the genes are indicated, as well as distances between RNA genes and flanking ORFs. Filled arrows denote major transcription start sites of the RNAs. The schematics are not drawn to scale. (A) In the sequence, arrows denote transcription start sites and the bracket denotes the 3'-end. The AGA and ACA triplets are boxed.

Homology searches at both the nucleotide and amino acid levels failed to identify any significant matches in yeast or other organisms for these RNAs. In addition to the small ORFs, some of the mapped RNAs potentially possess unusually long 3' UTRs for yeast mRNAs (Fig. 4B). For example, the 716 nt RNA of the chromosome II 3.3 kb gap would contain a 372 nt 3' UTR if the 255 nt ORF was translated.

Together, the sequences of these newly discovered RNAs suggest that yeast may express a greater diversity of small proteins <100 aa than previously predicted. Recent analysis of the yeast transcriptome supports this prediction, as 160 polyadenylated transcripts were identified by SAGE (serial analysis of gene

expression) technology that did not correspond to predicted ORF regions in the genome (36). In brief, short diagnostic sequence tags from poly(A) RNA were isolated, concatenated, PCR amplified, cloned and sequenced. Since the 5'-3'-ends of the RNAs corresponding to the sequence tags were not mapped in the SAGE analysis, it is unclear if a given SAGE tag corresponds to a transcript with the sequence predicted to be an mRNA. However, in combination with our mapping data, the simplest prediction is that many of the SAGE tags will correspond to bona fide mRNAs. In fact, two of the more abundant transcripts we have identified in our northern blots, (RNA716 and RNA530) correspond to SAGE tags NORF10 and NORF6, respectively (36). The larger implication of this work is that the 100 aa minimum ORF size for computer-based prediction of ORFs in any genome is likely to be too large and will produce an underestimation of the actual number of ORFs in the genome.

As a first step to determining the functions of RNA161, RNA515, RNA716, RNA530 and RNA487, we constructed yeast strains that either overexpress each RNA, or are deleted for the corresponding gene. In each case, overexpression of the transcript from high copy plasmids showed no obvious phenotype under a variety of conditions (see Materials and Methods). Null mutations for each gene were created by standard methods replacing the coding region either with the URA3 gene or the neomycin resistance gene (see Materials and Methods). Strains carrying deletions of each respective RNA gene were viable and showed no obvious growth phenotype under a variety of conditions (see Materials and Methods). Furthermore, all deletion strains were able to mate and showed no defects in germination. The absence of an obvious deletion phenotype for these RNA genes is typical of most yeast genes, as 70% of the entire yeast genome shows no obvious phenotype upon disruption (37).

## CONCLUSIONS

We have utilized two strategies to methodically search for new RNA transcripts in yeast. First, consensus RNA polymerase III promoters were identified and analyzed for RNA transcription. Second, sequences  $\geq 2$  kb lacking predicted ORFs were tested for RNA expression by northern blot analysis. These strategies resulted in the identification of 16 new RNAs ranging in size from 161 to 1200 nt. Two of the RNAs are clearly non-coding, while several contain potential small ORFs. The identification of such a large number of new transcripts from only 10 candidate polymerase III promoters and 59 chromosomal gaps (including the TLC1 gap region) provides evidence that there are many RNAs expressed in yeast that cannot be predicted by standard homology searches or current open reading frame criteria. Specifically, >20% of the  $\geq 2$  kb gaps located between predicted genes express RNAs. Moreover, these RNAs can be expressed from regions not expected to be transcribed. For example, RNA170 is expressed from sequences within the assumed promoter region of the neighboring gene, ERG2. Therefore, the density of genes on chromosomes, at least in some regions, may be even higher than currently predicted (1).

We hypothesize that careful examination of other regions of the genome is likely to reveal additional new RNAs for several reasons. First, because such a large percentage of gaps  $\geq 2$  kb expressed RNAs (>20%), it is possible that a similar percentage of smaller gaps might also express RNAs. In addition, we found that the size and coding potential of our 16 new RNAs correlated

with the size of the ORF gap in which they were expressed. In particular, large gaps  $\geq 2$  kb between ORFs expressed primarily mRNAs of 487 to  $>1000$  nt. In contrast, the small ncRNAs, RNA170 and RNA161, were expressed from smaller gaps of between 980 and 631 nt, respectively. These results suggest that analysis of smaller gaps ( $<2$  kb) will reveal transcripts that are more likely to be non-coding. Next, in the polymerase III promoter search, we demanded a stringent match to the consensus B box sequence, then we utilized other criteria to narrow our northern analysis to 10 candidates. Therefore, other uncharacterized polymerase III genes may exist that simply did not meet our criteria. As the identification of novel RNAs continues, the genetic analysis of their function in yeast will be important for an understanding of the multiple roles of RNA molecules in eukaryotic cells.

## ACKNOWLEDGEMENTS

We wish to thank Stephen Buratowski for generously donating the TDS4 plasmids, Heli Roiha for providing us with the various related yeast species, and Peter Geiduschek and George Kassavetis for reagents. This work was funded by the Howard Hughes Medical Institute. W.O. is supported by a postdoctoral fellowship from HHMI.

## REFERENCES

- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.*, (1996) *Science*, **274**, 546–567.
- Kirsebom, L. A. (1995) *Mol. Microbiol.*, **17**, 411–420.
- Prescott, J. and Blackburn, E. H. (1997) *Genes Dev.*, **11**, 528–540.
- Newman, A. (1994) *Curr. Opin. Cell Biol.*, **6**, 360–367.
- Mattaj, I. W., Tollervey, D. and Séraphin, B. (1993) *FASEB J.*, **7**, 47–53.
- Maxwell, E. S. and Fournier, M. J. (1995) *Ann. Rev. Biochem.*, **35**, 897–934.
- Bachelier, J. -P., Michot, B., Nicoloso, M., Balakin, A., Ni, J. and Fournier, M. J. (1995) *Trends Biochem. Sci.*, **20**, 261–264.
- Kiss-László, Z., Henry, Y., Bachelier, J. -P., Caizergues-Ferrer, M. and Kiss, T. (1996) *Cell*, **85**, 1077–1088.
- Inouye, M. and Delihias, N. (1988) *Cell*, **53**, 5–7.
- Delihias, N. (1995) *Mol. Microbiol.*, **15**, 411–414.
- Kickhoefer, V. A., Searles, R. P., Kedersha, N. L., Garber, M. E., Johnson, D. L. and Rome, L. H. (1993) *J. Biol. Chem.*, **268**, 7868–7873.
- Simpson, L. and Thiemann, O. H. (1995) *Cell*, **81**, 837–840.
- Geck, P., Medveczky, M. M., Chou, C. S., Brown, A., Cus, J. and Medveczky, P. G. (1994) *J. Gen. Virol.*, **75**, 2293–2301.
- Rastinejad, F., Conboy, M. J., Rando, T. A. and Blau, H. M. (1993) *Cell*, **75**, 1107–1117.
- Keiler, K. C., Waller, P. R. H. and Sauer, R. T. (1996) *Science*, **271**, 990–993.
- Kay, G. F., Penny, G. D., Patel, D., Ashworth, A., Brockdorff, N. and Rastan, S. (1993) *Cell*, **72**, 171–182.
- Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. and Davis, R. L. (1997) *Cell*, **88**, 445–457.
- Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993) *Cell*, **75**, 843–854.
- Wightman, B., Ha, I. and Ruvkun, G. (1993) *Cell*, **75**, 855–862.
- Tam, W., Ben-Yehuda, D. and Hayward, W. S. (1997) *Mol. Cell. Biol.*, **17**, 1490–1502.
- Caponigro, G., Muhrad, D. and Parker, R. (1993) *Mol. Cell. Biol.*, **13**, 5141–5148.
- Ausubel, F. M. (1995) *Current Protocols in Molecular Biology*, on CD-ROM. John Wiley and Sons, Inc, New York.
- Muhrad, D. and Parker, R. (1992) *Genes Dev.*, **6**, 2100–2111.
- Gietz, R. D. and Schiestl, R. H. (1995) *Transforming Yeast with DNA*. (Invited chapter) *Methods in Molecular and Cellular Biology*. Vol. 5, pp. 255–269.
- Schena, M. and Yamamoto, K. R. (1988) *Science*, **241**, 965–967.
- Hitzeman, R. A., Hagie, F. E., Hayflick, J. S., Chen, C. Y., Seeburg, P. H. and Derynck, R. (1982) *Nucleic Acids Res.*, **10**, 7791–7808.
- Christianson, T. W., Sikorski, R. S., Dante, M., Shero, J. H. and Hieter, P. (1992) *Gene*, **110**, 119–122.
- Geiduschek, E. P. and Tocchini-Valentini, G. P. (1988) *Ann. Rev. Biochem.*, **57**, 873–914.
- Buratowski, S. and Zhou, H. (1992) *Cell*, **71**, 221–230.
- Walter, P. and Blobel, G. (1982) *Nature*, **299**, 691–698.
- Lee, J. Y., Evans, C. F. and Engelke, D. R. (1991) *Proc. Nat. Acad. Sci. USA*, **88**, 6986–6990.
- O'Brien, C. A., Margelot, K. and Wolin, S. L. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 7250–7254.
- Van Horn, D. J., Eisenberg, D., O'Brien, C. A. and Wolin, S. L. (1995) *RNA*, **1**, 293–303.
- Singer, M. S. and Gottschling, D. E. (1994) *Science*, **266**, 404–409.
- Balakin, A. G., Smith, L. and Fournier, M. J. (1996) *Cell*, **86**, 823–834.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett Jr., D. E., Hieter, P., Vogelstein, B. and Kinzler, K. W. (1997) *Cell*, **88**, 243–251.
- Goebel, M. G. and Petes, T. D. (1986) *Cell*, **46**, 983–992.