

Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics

John P. McCutcheon and Sean R. Eddy*

Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

Received March 13, 2003; Revised April 14, 2003; Accepted May 6, 2003

DDBJ/EMBL/GenBank accession nos*

ABSTRACT

We screened for new structural non-coding RNAs (ncRNAs) in the genome sequence of the yeast *Saccharomyces cerevisiae* using computational comparative analysis of genome sequences from five related species of *Saccharomyces*. The screen identified 92 candidate ncRNA genes. Thirteen showed discrete transcripts when assayed by northern blot. Of these, eight appear to be novel ncRNAs ranging in size from 268 to 775 nt, including three new H/ACA box small nucleolar RNAs.

INTRODUCTION

In addition to protein-coding genes that produce messenger RNAs, there are also genes for non-coding RNAs (ncRNAs) that function directly as structural, regulatory or even catalytic molecules in cells (1,2). Some ncRNAs are well known, such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs). Other ncRNAs were discovered more recently, and some are in surprisingly numerous gene families, such as the small nucleolar RNAs (snoRNAs), which guide the site-specific modification of various other RNAs (3), and microRNAs, which appear to be a large class of post-transcriptional regulators (4).

Historically, both experimental gene discovery efforts and computational genome sequence annotation have been biased against non-coding genes, since ncRNAs lack open-reading frames (ORFs) and often lack other features such as poly(A) tails. More recently, a number of computational and experimental efforts have been undertaken to systematically identify new ncRNA genes, especially in model systems amenable to genetic and biochemical analysis (5–17).

The budding yeast *Saccharomyces cerevisiae* is one of the most powerful model systems and was the first completely sequenced eukaryotic genome (18). Systematic discovery of new ncRNA genes in the yeast genome has to date focused on finding new members of known ncRNA gene families, such as tRNAs (19,20) or C/D box methylation guide snoRNAs (21). An exception is a pioneering study from Roy Parker's laboratory, where transcriptional activity was probed in 'gray holes' (unusually large intergenic regions) and downstream of predicted polymerase III promoters. One new

snoRNA gene and one novel ncRNA of unknown function were discovered (22).

Our laboratory has described a computational approach to ncRNA gene-finding using comparative genome sequence analysis, implemented in the program QRNA (23). QRNA identifies conserved sequences that show mutational patterns consistent with a conserved RNA secondary structure, as opposed to a conserved coding frame or other conserved genomic features. Since QRNA only looks for conserved intramolecular RNA secondary structure, it fails to identify ncRNAs that lack significant structure (such as many *trans*-acting antisense RNAs). Also, it identifies *cis*-regulatory mRNA structures in addition to independent ncRNA transcripts, and has a significant apparent false positive rate, so experimental characterization of predicted loci is essential. We have previously employed QRNA in screens for new ncRNA genes in the genome of the bacterium *Escherichia coli*, using comparisons to four other gamma proteobacterial genome sequences (5), and in the archaeon *Pyrococcus furiosus*, using comparisons to two other *Pyrococcus* species (14).

Mark Johnston and colleagues have recently assembled draft genome sequences of five *Saccharomyces* species for comparative genomics purposes (24). They suggested that one use of their data would be to identify new ncRNA genes (24). We used these comparative *Saccharomyces* genomic data to perform a QRNA screen for new structural ncRNA genes in *S.cerevisiae*.

MATERIALS AND METHODS

Computational screen

Saccharomyces cerevisiae genome sequence and annotation were downloaded from the *Saccharomyces* Genome Database (SGD), <ftp://genome-ftp.stanford.edu/pub/yeast/> on 29 November 2001 (25). Non-mitochondrial intergenic sequences other than long terminal repeats (LTRs) and autonomously replicating sequences (ARSs) were extracted according to SGD annotation. Draft genome sequences from *Saccharomyces mikatae* (2.7× coverage in contigs), *Saccharomyces kudriavzevii* (3.3× coverage in contigs), *Saccharomyces bayanus* (2.4× coverage in contigs), *Saccharomyces castellii* (3.8× coverage in contigs) and *Saccharomyces kluyveri* (3.4× coverage in contigs) were obtained from <http://genome.wustl.edu/projects/yeast/> and RepeatMasked (A. F. A. Smit

*To whom correspondence should be addressed. Tel: +1 314 362 7666; Fax: +1 314 362 7855; Email: eddy@genetics.wustl.edu

†AY253285–AY253289

and P. Green, <http://ftp.genome.washington.edu/RM/RepeatMasker.html>) using the SGD-annotated *S.cerevisiae* LTR and ARS sequences as the RepeatMasker library. Sequence comparisons were performed with WU BLASTN 2.0 (W. Gish, <http://blast.wustl.edu>) using the low-complexity sequence filters dust and seg. QRNA version 1.2b (23) (<http://www.genetics.wustl.edu/eddy/software/#qrna>) was used with a window size of 150 nt (-w 150), a window slide increment of 50 nt (-x 50) and a cut-off of 5 bits for the posterior log-odds score of the RNA model. At the time of data download, the SGD identified 6357 protein-coding genes and 363 non-mitochondrial ncRNA genes (276 of which were non-mitochondrial tRNAs) (25).

Northern blots and transcript sequencing

Total yeast RNA was harvested from strain BY4743 (Research Genetics) as described (26) in five different growth conditions (all growths were performed at 30°C unless otherwise noted): logarithmic growth in YPD (OD₆₀₀ = 0.5), YPD grown for 3 days (saturated), logarithmic growth in YPD followed by heat shock for 30 min at 37°C, logarithmic growth in minimal media and logarithmic growth in YPGalactose. For northern blots, 10 µg of total RNA was electrophoresed on 6% acrylamide gels, electroblotted to Zeta-probe nylon membranes (BioRad), and hybridized to 5' end-labeled 40–50 nt oligonucleotide probes, as previously described (14), except the hybridization temperature was 65°C. The 100 bp DNA size standard ladder was from New England Biolabs. To facilitate comparison of results, all northern blots were consistently exposed for 18–20 h on a Phosphorimager screen (Molecular Dynamics). Rapid amplification of cDNA ends (RACE) (27) and DNA sequencing were carried out as described (14). A synthetic poly(A) tail was added to total RNA for use in first-strand synthesis. A list of the oligos used in the northern blot and RACE experiments is in Supplementary Material (Table S2).

For the candidate 39 and CUP1 expression experiments, total RNA was harvested from cells grown to midlog in YPD in the presence of 1 mM tetraethylenepentamine (TEPA) (Sigma-Aldrich) or 1 mM copper sulfate (Sigma-Aldrich).

The G-statistic

The G-statistic was used as described (28). With three codon positions being measured, 2 degrees of freedom are available. G-statistic scores >5.99 are significant at the 0.05 confidence level with 2 degrees of freedom.

RESULTS

The screen consists of several steps. Standard pairwise sequence comparison is used to identify conserved intergenic sequences and to produce a data set of pairwise alignments. QRNA is used to analyze these alignments one at a time and to classify a subset of them as being predicted to be evolving under the constraint of a conserved RNA secondary structure. Overlapping predictions are collapsed into a single predicted RNA locus on the genome. Northern blot analysis with strand-specific oligo probes is used to identify loci which express discrete RNA transcripts and to identify the size and strand of the transcript. Loci that are not visibly or reproducibly expressed under any tested growth condition are considered to

be false positive predictions and eliminated. 5' and 3' RACE-PCR clones of overlapping halves of these RNAs are sequenced to obtain full-length transcript sequences. These full-length RNA sequences are reanalyzed computationally to eliminate loci in UTRs of mRNAs and previously unannotated ORFs miscalled by QRNA. The result is a set of RNAs which are known to be expressed, evolutionarily conserved at the sequence level, apparently conserved at the RNA secondary structural level, do not overlap previously annotated genes, and apparently devoid of a significant or evolutionarily conserved ORF. The details of these steps follow.

Identification of conserved intergenic sequence alignments

For QRNA analysis, we want the comparative genomes to be distant enough that structural RNAs have accumulated enough compensatory base pair mutations for QRNA to differentiate them from other conserved regions, but close enough that BLASTN sequence alignments detect conserved RNAs and align them reasonably correctly, even in base paired regions. The 'sweet spot' is at ~75–85% sequence identity. Because different genes evolve at different rates, having multiple genomes at different evolutionary distances is an advantage. Ideally, no comparative genome would be so close that neutrally evolving, nonfunctional sequence would contribute significantly to the BLASTN alignments, so QRNA is given only significantly conserved (and therefore likely functional) sequences to analyze.

The *Saccharomyces* genus is partitioned into three subgroups: *sensu stricto*, *sensu lato* and petite negative, in order of increasing evolutionary distance from *S.cerevisiae* (29). Draft sequences of three *sensu stricto* species (*S.mikatae*, *S.kudriavzevii*, *S.bayanus*), one *sensu lato* (*S.castelli*), and one petite negative (*S.kluyveri*) were available. The average percent identity of alignments between *S.cerevisiae* and these five other genomes indicated they were at an appropriate evolutionary distance for QRNA analysis (Table 1).

6065 annotated intergenic regions (3 098 859 nt) of *S.cerevisiae* were searched against the five draft *Saccharomyces* genomes with WU-BLASTN. Alignments that were at least 50 nt long and 65–90% identical with E-values of ≤0.001 were kept for QRNA analysis, a total of 23 716 alignments (Table 1).

As examples of how QRNA would perform in these genomes, we looked at results with three known ncRNAs: the RNA component of the ribonuclease RNaseP, the RNA component of the signal recognition particle (SRP) and the splicingosomal RNA U1. RNase P is found by QRNA in each of the *sensu stricto* species with an average score of 19.14 bits (well above our significance threshold of 5 bits) and an average percent identity of ~87%. QRNA predicts an RNA structure at positions 16–314 of the 369 nt RNA. U1 RNA is found in *S.kudriavzevii*, *S.bayanus* and *S.kluyverii* with an average score of 14.58 bits and an average percent identity of ~83%. QRNA predicts an RNA structure at positions 185–460 of the 668 nt RNA. Finally, SRP RNA is found in each of the *sensu stricto* species with an average score of 14.86 bits and an average percent identity of ~85%. QRNA predicts an RNA structure at positions 244–466 of the 675 nt RNA. These results indicated that the *sensu stricto* species would be most informative in finding new ncRNAs and that while QRNA

Table 1. BLASTN alignment and QRNA statistics

Species	BLAST alignments		Length	QRNA No. of predictions	No. of known RNAs identified	No. of new RNAs identified	Which candidates identified
	No.	% Identity					
<i>S.mikatae</i>	6942	76.1	257	203	132	4	10, 19, 24, 74
<i>S.kudriavzevii</i>	6828	75.9	247	185	128	4	29, 39, 71, 92
<i>S.bayanus</i>	5004	76.5	214	160	126	3	10, 29, 74
<i>S.castellii</i>	2564	81.5	152	100	91	0	None
<i>S.kluyveri</i>	2378	81.2	141	92	87	0	None
Overall	23 716	77.2	222	305	162	8	10, 19, 24, 29, 39, 71, 74, 92

The species in column 1 are listed in decreasing relatedness to *S.cerevisiae*. Column 2 shows the number of BLASTN alignments that met the criteria described in Materials and Methods for input into QRNA. Column 3 shows their average percent identity, and average length is shown in column 4. Column 5 shows the number of QRNA predictions made from the pairwise comparison alone (the total number of 305 accounts for overlapping predictions). Column 6 shows the number of QRNA predictions from column 5 that overlap known ncRNA genes. Column 7 shows the number of eight new ncRNAs reported in this study that each of the pairwise comparisons would have found using that genome alone; column 8 gives the candidate numbers.

finds known RNAs with high scores, it underpredicts their real sizes and structures. QRNA typically detects only part of the conserved secondary structure, and often imperfectly. We do not consider the details of these structure predictions to be of much biological use, as they are compromised by the poor signal/noise inherent in this difficult computational problem; we therefore have not included raw QRNA structure predictions in this report.

Prediction of structural RNA loci by QRNA

QRNA classified a total of 305 candidate structural RNA loci in *S.cerevisiae* (genomic coordinates, QRNA score and functional classification are given in Supplementary Material Table S1). 177 of these loci overlapped 162 known RNA genes and were removed from the candidate list (Table 1). This allowed us to estimate the sensitivity of the screen; 162/363 (~45%) of known and annotated non-mitochondrial ncRNA genes were detected. Most of the known ncRNAs that QRNA missed are tRNAs or snoRNAs; U4, U5 and the telomerase RNA were also not found. To estimate the false positive rate of the screen, we randomly shuffled the input pairwise alignments (preserving their percent identity, while destroying any specific pattern of mutation) and rescored the shuffled alignments with QRNA. Seventy-nine loci were predicted as RNAs in shuffled data, indicating that a significant but not unreasonable number of false positives exist, accounting for about a quarter of the loci (e.g. about half of the unknown loci). This test only accounts for false positives that arise from QRNA's inherent rate of misclassifying alignments with random patterns of mutation. Other biological sources of false positives exist, such as inverted repeat DNA structures, *cis*-regulatory RNA structures and RNA pseudogenes, all of which have an increased likelihood of looking like ncRNA genes to QRNA.

To conservatively eliminate sequences likely to be *cis*-regulatory mRNA elements and reduce the number of candidates that would be screened by northern blots, 34 loci that fell within 50 nt of an annotated (ORF) start or stop codon were removed from the set of candidates.

BLASTN and BLASTX searches with the remaining candidates against the NCBI non-redundant nucleotide and protein databases were done to detect any obvious unannotated protein or protein pseudogene sequences. One likely pseudogene of ribosomal protein S8 was removed.

Candidates 39 and 40 were identical predictions in two identical tandem copies of the CUP1 locus, CUP1-1 and CUP1-2. They are treated as a single prediction (arbitrarily chosen as candidate 39).

Detection of transcription by northern blot

The 92 remaining candidates were screened for transcriptional activity by northern blot, assaying total RNA from yeast grown in five different growth conditions (see Materials and Methods). For each candidate, two oligonucleotide probes (one per strand) were selected because QRNA gives no strand information; a conserved structure on one strand is also conserved on the other (oligo sequences are given in Table S2). Thirteen candidates were found to be transcribed, with sizes ranging from ~280 to >1000 nt (Fig. 1). Two of these (4 and 72, Fig. 1A and 1B) were then inferred to be part of adjacent coding genes based just on size; their transcripts were larger than the intergenic region, and large enough to encompass one of the flanking ORFs (Table 2).

One might wonder whether the yeast genome might have so many uncharacterized small transcripts that we would get similar results simply by probing randomly selected intergenic regions, or intergenic regions with conserved sequence but no significant QRNA score. That is, did the QRNA screen enrich for new RNAs? To test this, twenty 200 nt intergenic regions were selected randomly and treated exactly as QRNA predictions were (oligo sequences are given in Table S2). Thirteen of these 20 controls were conserved with an *E*-value of ≤ 0.001 between *S.cerevisiae* and at least one of the other *Saccharomyces* species. None showed a discrete transcript indicative of an ncRNA or an mRNA by northern blot (data not shown).

Transcript sequencing by RACE

QRNA predictions do not accurately predict the bounds of the gene. To determine the exact location of the transcribed locus in the genome, and also to confirm the presence of the transcript and check the mature RNA sequence against the genome sequence, we attempted to clone and sequence 5' and 3' RACE-PCR products for the 11 candidates that showed positive northern blots and that were not already attributed to a protein-coding transcript. At least partial RACE data and sequence were obtained for 9 of the 11 (Table 2). Two more candidates, 65 and 66, were seen to be part of transcripts for

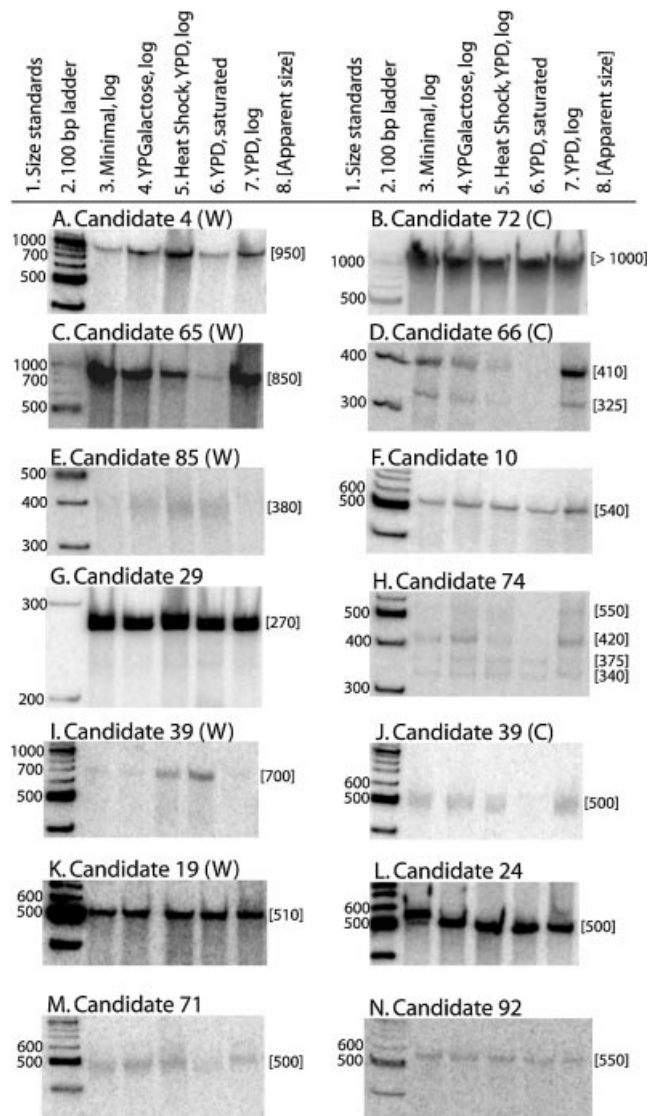


Figure 1. Northern blots for the 13 transcripts found in this screen. The candidate name is followed by either (W) or (C), indicating that the northern displayed was probed with an oligonucleotide targeting transcripts originating from the Watson or Crick strand. Flanking bands in the 100 bp ladder (which is in lane 2) are indicated in lane 1. Lanes 3–7 contain total RNA from the following growth conditions (all grown at 30°C to mid-log except where noted): lane 3, minimal media; lane 4, YPGalactose; lane 5, heat-shock 30 min at 37°C in YPD; lane 6, saturated YPD growth; lane 7, YPD. Lane 8 is the estimated size of each of the transcripts, shown as a bracketed number.

known coding genes (see Discussion). This left nine potential ncRNAs in the candidate list.

Sequence analysis

We do not know that these transcripts are ncRNAs. While QRNA has called each one a structural RNA based on its pattern of conservation between *S.cerevisiae* and at least one other species of *Saccharomyces*, it is difficult to exclude the possibility that they contain short translated ORFs. QRNA has a significant false positive rate, and coding RNAs may have structural RNA regulatory regions. We therefore re-examined

each sequence and its conservation, looking for conserved ORFs, now using knowledge of the full-length sequence of the transcript from the RACE experiments. In particular, we analyzed multiple sequence alignments, whereas QRNA only looks at pairwise alignments. A simple scoring scheme was developed. For all ORFs >18 nt (six amino acids) in a given *S.cerevisiae* candidate locus, a multiple alignment was constructed from the homologous regions of *S.mikatae*, *S.kudriavzevii* and *S.bayanus* using WU BLASTN followed by CLUSTALW (30). In some instances, sequence from one of the species was not present, possibly due to insufficient sequencing coverage or gene loss. Using the frame established by the ATG start codon of the *S.cerevisiae* sequence, completely identical first, second and third codon columns were tallied and normalized by the total number of perfectly conserved columns (see Fig. 2 for example). The null hypothesis is that conservation is independent of reading frame, whereas a conserved coding region will show a non-random distribution, mostly because of third position wobble. The significance of a deviation from randomness can be estimated by standard tests such as the *G*-statistic (see Materials and Methods).

We applied this test to 3323 annotated, non-hypothetical, conserved, intronless ORFs in *S.cerevisiae*, and found that 90% of real ORFs significantly reject the null hypothesis at $P \leq 0.05$. On average, completely conserved columns are in the first, second and third position 36.9, 40.5 and 22.6% of the time, respectively. Fifty-one of the 87 annotated, non-mitochondrial and non-tRNA ncRNAs contained an 'ORF' that gave a significant BLASTN alignment to at least one other *Saccharomyces* genome (tRNAs were excluded because of their short length). None of these 51 false ORFs rejected the null hypothesis. On average, completely conserved columns in false ORFs in ncRNA alignments were in the first, second and third positions 33.8, 33.3 and 32.9% of the time, respectively.

We tested the nine candidates and found one, candidate 85, which had a conserved 66 aa ORF and a bias towards third position mutations (Fig. 2). We therefore call candidate 85 a putative small ORF.

This leaves a final count of eight putative ncRNAs discovered by this screen (Table 2). The sequences and genomic coordinates for the five candidates where complete RACE is available have been deposited at SGD with the three-letter name RUF (RNA of unknown function) and at GenBank. Candidate 10 has been given SGD identifier RUF1 and GenBank accession no. AY253285; candidate 29, RUF2, AY253286; candidate 74, RUF3, AY253287; candidate 19, RUF4, AY253288; candidate 39, RUF5, AY253289. The three others have only been deposited at SGD: candidate 92, RUF6; candidate 24, RUF7; and candidate 71, RUF8. Although RUF1, RUF2 and RUF3 all seem to be H/ACA box snoRNAs, our evidence is not definitive and therefore we have not assigned them SNR gene names, the standard snoRNA nomenclature in yeast. Additional gene names are possible for these RNAs in the future as more information as to their function become available, but we have chosen RUF because it explicitly states that the function of the RNA is unknown (and it's easy and fun to pronounce). Our group and others will do more screens for ncRNAs in yeast, and the RUF nomenclature allows a systematic naming convention for future RNAs of unknown function.

Table 2. The candidate ncRNAs

Candidate	Strand	Chromosome	QRNA prediction	QRNA % identity	Approximate northern blot size	RACE size	ncRNA coordinates	Species present in (by BLASTN)	Final classification
4	W	4	654 475–654 623		950	ND			3' UTR of BMH2 transcript
72	C	13	477 821–478 107		>1000	ND			3' UTR of YKU80 or 5' UTR of PGM2
65	W	12	673 632–673 807		850	864			3' UTR of RPS28 transcript
66	C	12	795 049–795 231		410	387			5' End of SNR57 transcript (possible snoRNA)
85	W	2	680 349–680 498		380	387			Apparent 198 nt ORF
RUF1 (10)	C	4	1 492 936–1 492 651	76	540	550	1 493 029–1 492 480	<i>Smik, Skud, Sbay</i>	H/ACA snoRNA
RUF2 (29)	W	7	316 761–316 957	81	270	268	316 786–317 053	All 5 species	H/ACA snoRNA
RUF3 (74)	W	13	626 190–626 532	76	550/420, 375/340	531/431	626 239–626 669	All 5 but <i>Sklu</i>	H/ACA snoRNA
RUF4 (19)	W	5	378 376–378 574	72	510	788	377 955–378 742	<i>Smik, Skud, Sbay</i>	Likely ncRNA
RUF5 (39)	W	8	212 783–213 027	77	700	775 ^a	212 379–213 153	<i>Skud, Sbay</i>	Likely ncRNA
RUF6 (92)	C	16	857 235–857 037	79	550	? ^b	857 621–856 571 ^b	<i>Smik, Skud, Sbay</i>	Likely ncRNA
RUF7 (24)	W	6	123 202–123 350	78	500	? ^c	123 086–123 585 ^c	<i>Smik, Skud, Sbay</i>	Possible ncRNA
RUF8 (71)	W	2	462 588–462 685	88	500	? ^b	462 137–463 087 ^b	All 5 but <i>Scas</i>	Possible ncRNA

ND, not determined; ?, either one or both of the 3' and 5' RACE reactions did not work and so the size of the transcript is unknown; *Smik, S.mikatae; Skud, S.kudriavzevii; Sbay, S.bayanus; Scas, S.castellii; Sklu, S.kluyveri*: The candidate names are shown as RUF names followed by the candidate number in parenthesis. The % identity column shows the percentage identity reported for the best BLASTN hit to *S.bayanus*, using a query sequence defined by the ncRNA coordinates in column 8. Note that because of heterogeneous conservation across a locus, a simple overall identity measure does not reflect significance; for example, RUF4 shows low overall identity, but the BLAST alignment encompasses a smaller strongly conserved region.

^aRUF5-1 and RUF5-2 are two copies of the same ncRNA that arise from a duplicated region of the genome.

^bNo RACE data are available, so the coordinates are only the entire range of potential coordinates calculated by adding the size of the transcript seen on the northern blot to either side of the oligonucleotide position. The real transcript bounds are somewhere in this range.

^cOnly 3' RACE data are available, the coordinates are calculated by adding the northern blot size to the 3' coordinates. Candidates 4, 72, 65, 66 and 85 were attributed genomic features other than ncRNAs. Candidates RUF1–RUF8 are classified as new ncRNAs from this screen.

The following summarizes our results for all 13 detected transcripts, starting with the five that we do not believe are novel ncRNAs.

Candidates 4 and 72—possible UTR structures

Two candidates showed large transcripts on northern blots, and we did not pursue them further.

Candidate 4 is probably in the 3' UTR of BMH2, a 14-3-3 protein family member (31). BMH2 has an 822 nt ORF. The QRNA prediction falls 54 nt downstream of the BMH2 stop codon, in a 585 nt intergenic region. The candidate 4 northern blot showed a 900–1000 nt transcript on the same strand as BMH2 (Fig. 1A).

Candidate 72 is probably either in the 3' UTR of YKU80 (a DNA-binding Ku80 homolog) or the 5' UTR of PGM2 (phosphoglucomutase), which are transcribed from chromosome 13 in the same orientation. The QRNA prediction is in a 695 nt intergenic region, 193 nt from the 3' end of YKU80 and 216 nt from the 5' end of PGM2. The ORFs for PGM2 and YKU80 are 1710 and 1890 nt, respectively. The observed transcript for 72 is large and unresolved on the northern blot, >1000 nt (Fig. 1B). We assume it contains either the PKU80 or PGM2 ORF.

Candidate 65—ribosomal protein RPS28B 3' UTR

RACE experiments showed that candidate 65 is in the 3' UTR of ribosomal protein RPS28B. The prediction is in the 1093 nt intergenic region between RPS28B and NEJ1 on chromosome 12, 298 nt 3' from the RPS28B stop codon. The RPS28B

transcript appears to be ~850 nt long on the northern blot (Fig. 1C), while the size of the ORF is only 204 nt. This implies that the RPS28B 3' UTR is large, ~600 nt. It is plausible that RPS28B is post-transcriptionally regulated by a mechanism involving a conserved RNA structure in the 3' UTR, but no evidence for this exists presently.

Candidate 66—in a transcript with C/D box snoRNA SNR57

Candidate 66 is just 25 nt upstream from the C/D box snoRNA SNR57 on chromosome 12. The northern blot showed a major band at 410 nt and a minor band at 325 nt (Fig. 1D). RACE yielded a 387 nt transcript (Table 2) that included the entire SNR57 gene, predicted to be an 88 nt C/D box methylation guide snoRNA (21). A second northern blot probed with an oligonucleotide for SNR57 showed a strong transcript at ~90 nt and a minor transcript at ~410 nt (data not shown). 'Siamese' snoRNAs which contain both H/ACA and C/D snoRNA moieties are known (32), and we have seen such sequences processed into separate snoRNAs in Archaea (14,33). We are therefore suspicious that candidate 66 is a new snoRNA (or some other functional RNA); however, we cannot find the consensus features of either C/D or H/ACA snoRNAs in its sequence. On the other hand, some snoRNAs are known to be transcribed as larger precursor RNAs and then processed to their final functional sizes (34). Therefore, conservatively, we are not classifying 66 as a new ncRNA, because of the possibility it is just part of an SNR57 precursor.

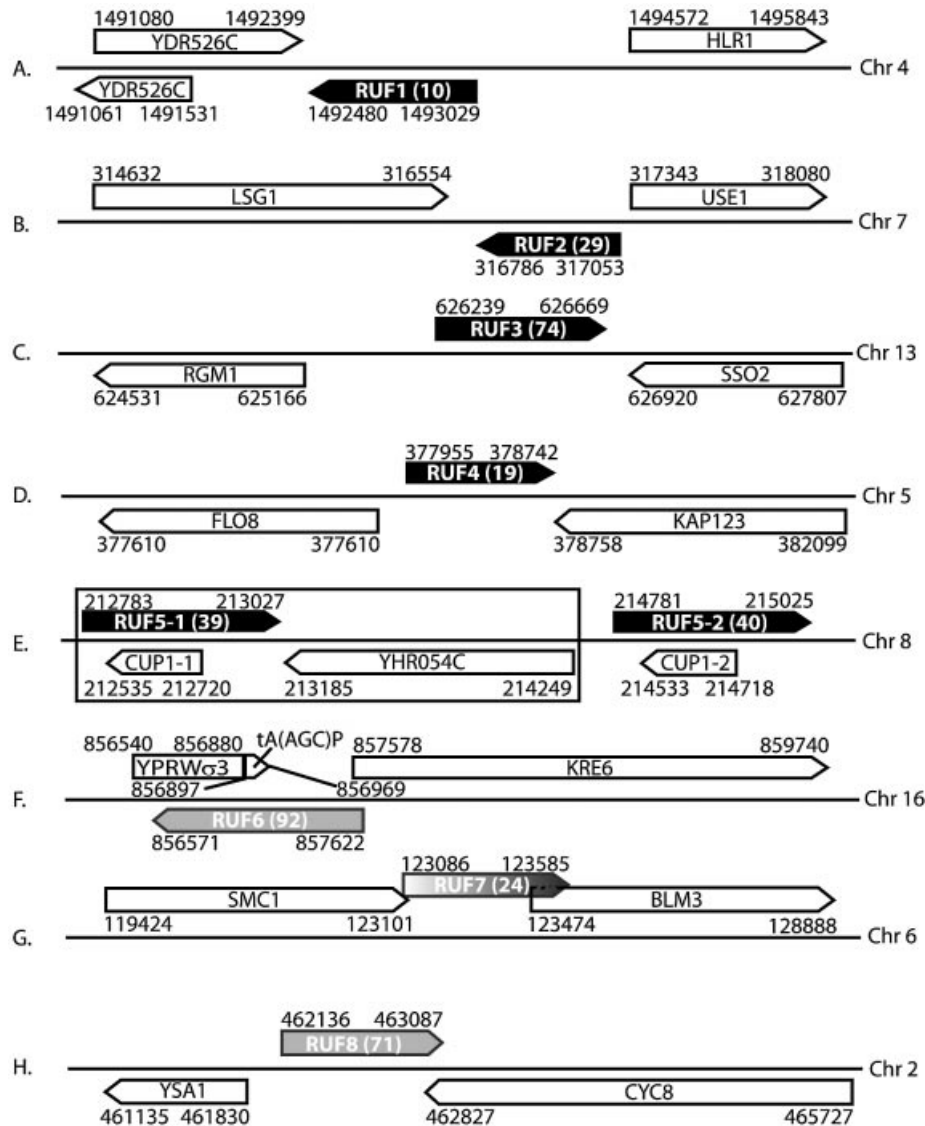


Figure 4. Schematic of the genomic context for the eight ncRNA candidates. SGD annotation is represented by open boxes. The ncRNAs for which complete RACE data are available are shown as black boxes, and the candidates for which there is incomplete or no RACE data are shown as gray or black-to-gray gradient boxes. The coordinates for the bounds of the genes are noted. For each RNA, the RUF names are given followed by the candidate number in parenthesis. (A) RUF1 (candidate 10). (B) RUF2 (candidate 29). (C) RUF3 (candidate 74). (D) RUF4 (candidate 19). (E) RUF5-1 (candidate 39) and RUF5-2 (candidate 40), and the CUP1 tandem array. The repeating unit is denoted by the box. (F and H) RUF6 (candidate 92) and RUF 8 (candidate 71). No RACE data are available, so the potential bounds of the gene are calculated by adding the size observed on the northern to each side of the bounds of the oligonucleotide probe. (G) RUF7 (candidate 24). Only 3' RACE data are available (as indicated by the darker end of the box), so the bounds of the transcript are calculated by adding the size observed on the northern to the bounds of the oligonucleotide probe.

sequenced (39). [RUF5-2 (candidate 40) fell in the identical sequence 63 nt upstream of the other copy, CUP1-2.]

The northern blot for RUF5 also showed a diffuse 500 nt transcript present on the same strand as CUP1 in all conditions except stationary phase (Fig. 1J). The transcript size for CUP1 is ~500 nt (38), so we assume this 500 nt transcript is CUP1 poly(A)⁺ mRNA. However, CUP1 is supposed to be induced by heat-shock treatment (40), and this transcript is, if anything, slightly lower in the heat-shock condition as compared to the standard growth condition (Fig. 1J, heat shock compared to YPD). Multiple RACE attempts have so far failed to isolate this transcript.

It is interesting that the putative CUP1 transcript and the antisense RUF5 RNA appear to be reciprocally regulated in stationary phase (Fig. 1I and J). We imagined that maybe copper levels are depleted in stationary phase, and RUF5 RNA comes on and represses CUP1 mRNA. We looked at northern blots of total RNA from exponentially growing cells in YPD plus 1 mM copper sulfate (high copper) or 1 mM TEPA, a copper chelator (low copper). Levels of the 500 nt CUP1 mRNA were induced in high copper (as expected), but unaffected by low copper (e.g. present at levels comparable to the YPD lane in Fig. 1J); whereas RUF5 transcription was not affected by either condition (e.g. no visible band, as in the

YPD lane in Fig. 1I) (data not shown). We still suspect RUF5 is an antisense regulator of CUP1, but according to these preliminary results, it does not appear to be involved in simple copper-responsive regulation of CUP1 transcript level.

RUF4, RUF6, RUF7 and RUF8—novel ncRNAs of unknown function

The remaining four transcripts appear to be novel ncRNAs, for which we have no clues as to their function.

RUF4 (candidate 19) shows a 510 nt transcript on a northern blot (Fig. 1K), but strangely, RACE sequences, though variable at both ends, supported a substantially larger transcript of 700–788 nt (Table 2). We cannot adequately explain this. We hypothesize that the transcript we can RACE is a rare precursor, whereas the abundant transcript we see on a northern blot is a processed form that is resistant to RACE. The first step of our RACE procedure uses poly(A) polymerase to tail total RNA; this would select against RNAs without 3' hydroxyl ends, such as circular RNAs or RNAs with terminal 2'-3' cyclic phosphates.

The QRNA prediction for RUF7 (candidate 24) is in a 373 nt intergenic region of chromosome 6 downstream of SMC1 and upstream of BLM3 (Fig. 4G). The northern blot showed an ~500 nt transcript encoded on the Watson strand of the genome (Fig. 1L), the same strand as SMC1 and BLM3, too long to be entirely within the intergenic region but not long enough to contain either ORF. RACE for the 3' end of the transcript showed that the last 113 nt of the RUF7 transcript overlapped the 5' coding region for BLM3 (Fig. 4G). Multiple attempts at 5' RACE for the 5' end of RUF7 failed, but the transcript size combined with the 3' end position allows us to infer that the RUF7 transcript also overlaps the 3' end of SMC1. A conserved 54 nt ORF (18 amino acids) that is out of frame with BLM3 is within this overlap region, but the pattern of conservation in the putative ORF is dominated by the frame of the BLM3 ORF, indicating that there is greater selective pressure on the BLM3 ORF than on the smaller putative ORF.

Despite attempting a number of different experimental variations, RACE experiments for RUF6 (candidate 92) and RUF8 (candidate 71) consistently failed. The northern blots for these two candidates showed relatively low levels of transcription, but they are unlikely to be spurious; these levels are comparable to the low levels we observe for the CUP1 mRNA or the RUF3 (candidate 74) probable H/ACA box snoRNA, and the results were repeatable and both bands remained after more stringent washings of the blots (Fig. 1M and N). (This reproducibility contrasts to a small number of other candidates that we did not consider to be confirmed, because they showed weak northern signals that were not consistently reproducible.) The genomic region where RUF8 must lie has a conserved short ORF of length 75 nt (25 amino acid protein) that exhibits features that are weakly consistent with protein-coding genes, with slightly less conservation in putative third codon positions, but this pattern was not significant by the *G*-statistic.

DISCUSSION

We used a computational comparative genomic screen to search for ncRNA genes in *S.cerevisiae*. From a candidate list of 92 predicted ncRNAs, 13 were shown to be transcribed, and

after further characterization, eight of these are thought to be potential novel ncRNA genes (summarized in Table 2). Three of these candidates are apparently new H/ACA box snoRNAs, while the others do not appear to belong to known families of RNA genes.

The evidence that these RNAs are non-coding is that they have no significant conserved ORF by comparative sequence analysis. Every potential ORF ≥ 6 amino acids was analyzed in multiple alignments with orthologous *Saccharomyces sensu stricto* sequences. This is not definitive, but we believe it is the most powerful approach available to us. Even *in vitro* translation experiments have been misleading for distinguishing non-coding from coding transcripts; the *gas5* ncRNA, for example, was originally thought to be coding because it can be forced to express a peptide *in vitro* (41,42).

The evidence for these RNAs having biological functions is that their sequences are significantly conserved in other *Saccharomyces* species. However, at this preliminary stage, we do not have many ideas about what those functions might be, except for the three H/ACA pseudouridylation guide snoRNAs. Experimental characterization of each of the other genes will undoubtedly be necessary. We do not find significant BLASTN similarities to any of these new RNAs to anything outside the *Saccharomyces* genus, which might indicate that these RNAs are phylogenetically restricted to *Saccharomyces*. However, this is a typical result even for universally conserved RNAs since many structural RNAs are conserved at the secondary structure level, but poorly conserved at the sequence level. For example, a BLASTN search with the *S.cerevisiae* homolog of the almost universally conserved RNase P RNA (35) detects significant alignments only in the genera *Saccharomyces* [although the RNase P sequences in *Torulaspora*, *Zygosaccharomyces* and *Kluyveromyces* also show up because they have been directly sequenced to aid in RNase P secondary structure prediction (43)]. We and others are working on improved RNA similarity search algorithms that work at the structural level, which may help us find informative homologies for RNAs like these (44).

It is currently difficult to identify H/ACA box snoRNAs by sequence analysis, because the consensus sequence and structure is quite degenerate (3,36). For example, we identified several 'novel' ncRNA genes in the archaeon *Pyrococcus furiosus* (14) that have now been shown to be likely archaeal H/ACA box snoRNAs (33). In *S.cerevisiae*, there are at least 10 H/ACA box snoRNAs that remain undiscovered; the large ribosomal subunit rRNA contains 30 pseudouridyl modifications (45), all of which are likely to be H/ACA box snoRNA-directed, but only 20 of which are assigned to known H/ACA box snoRNAs (36). We have manually examined our ncRNAs for features of H/ACA box snoRNAs, but only RUF2 stood out to us. A prototype computational algorithm then detected two more, RUF1 and RUF3 (P. Schattner and T. Lowe, personal communication). We would not be surprised if more of our RNAs turn out to be H/ACA box snoRNAs.

The results presented here highlight the importance of having multiple genome sequences at various evolutionary distances available for comparative genomics. For example, if only one of the *sensu stricto* species were available to us for comparison, we would have found at most half (four of eight) of the new ncRNAs described here (Table 1). If the 'correct' combination of genomes were chosen, then two genomes

(*S.mikatae* and *S.kudriavzevii*) would have given us all of the new ncRNAs, although it is unclear how these genomes would have been chosen *a priori*. Different genes are under different evolutionary pressures; they evolve at different rates. What is an appropriate distance for one gene may not be an appropriate distance for another, and therefore to ultimately get a complete and reliable set of genes for an organism, multiple genome sequences at various evolutionary distances will be needed.

In closing, we emphasize that this screen is very unlikely to have saturated the *S.cerevisiae* genome for ncRNA genes. Although our results derive from a systematic computational screen, we cannot draw any conclusions about the total number of ncRNA genes in yeast, because our gene-finding program is imperfect and our initial approach has been conservative. QRNA has a low sensitivity (~45%, as measured on known yeast RNAs) and is particularly insensitive to ncRNAs with little or no conserved intramolecular secondary structure. We chose to screen only annotated intergenic regions of *S.cerevisiae*, even though not all of these ORF predictions are correct (46), and ncRNA genes may occur in introns (47) or even overlapping coding mRNAs (48). A negative result by northern blot does not mean that there is not a ncRNA present; it may simply be that our northern blots are not sensitive enough to detect rare transcripts or that the ncRNA is expressed in a growth condition that we did not test. We would not be surprised if in future screens, whether computational or experimental, more yeast ncRNAs are discovered.

SUPPLEMENTARY MATERIAL

A list of the 305 QRNA predictions (Table S1) and the oligonucleotides used in the northern blot and RACE experiments (Table S2) is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ziva Misulovin for expert assistance and advice in the laboratory portion of this work and Tom Jones for help in identifying and characterizing the snoRNA. We are very grateful to Peter Schattner and Todd Lowe for sharing unpublished results on H/ACA box snoRNA detection. We thank Paul Cliften, Mark Johnston and the Washington University Genome Sequencing Center for providing unpublished genome sequences of five *Saccharomyces* species (sequencing supported by NIH GM63803). This work was supported by NIH NHGRI-HG01363, the Howard Hughes Medical Institute and an endowment from Mr Alvin Goldfarb.

REFERENCES

- Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.*, **2**, 919–929.
- Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
- Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
- Pasquinelli,A.E. and Ruvkun,G. (2002) Control and developmental timing by microRNAs and their targets. *Annu. Rev. Cell. Dev. Biol.*, **18**, 495–513.
- Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.
- Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.
- Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.
- Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, **29**, 3928–3938.
- Lagos-Quintana,M., Rauhut,R., Yalcin,A., Meyer,J., Lendeckel,W. and Tuschl,T. (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.*, **12**, 735–739.
- Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
- Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
- Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachelierie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.
- MacIntosh,G.C., Wilkerson,C. and Green,P.J. (2001) Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.*, **127**, 765–776.
- Klein,R.J., Misulovin,Z. and Eddy,S.R. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl Acad. Sci. USA*, **99**, 7542–7547.
- Schattner,P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.*, **30**, 2076–2082.
- Tang,T.H., Bachelierie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Huttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA*, **99**, 7536–7541.
- Llave,C., Kasschau,K.D., Rector,M.A. and Carrington,J.C. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, **14**, 1605–1619.
- 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) Life with 6000 genes. *Science*, **274**, 546, 563–567.
- Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
- Fichant,G.A. and Burks,C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220**, 659–671.
- Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
- Olivas,W.M., Muhlrud,D. and Parker,R. (1997) Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.*, **25**, 4619–4625.
- Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
- Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterston,R.H. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.
- Issel-Tarver,L., Christie,K.R., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S., Fisk,D.G., Harris,M., Schroeder,M., Sethuraman,A., Tse,K., Weng,S., Botstein,D. and Cherry,J.M. (2002) *Saccharomyces* Genome Database. *Methods Enzymol.*, **350**, 329–346.
- Holstege,F.C., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
- Frohman,M.A. (1993) Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: thermal RACE. *Methods Enzymol.*, **218**, 340–356.
- Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd Edn. W.H. Freeman and Company, New York.
- Barnett,J.A. (1992) The taxonomy of the genus *Saccharomyces meyeri* ex Reess: a short review for non-taxonomists. *Yeast*, **8**, 1–23.
- Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment

- through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
31. Callejo, M., Alvarez, D., Price, G.B. and Zannis-Hadjopoulos, M. (2002) The 14-3-3 protein homologues from *Saccharomyces cerevisiae*, Bmh1p and Bmh2p, have cruciform DNA-binding activity and associate in vivo with ARS307. *J. Biol. Chem.*, **277**, 38416–38423.
 32. Jady, B.E. and Kiss, T. (2001) A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J.*, **20**, 541–551.
 33. Rozhdestvensky, T.S., Tang, T.H., Tchirkova, I.V., Brosius, J., Bachelier, J.P. and Huttenhofer, A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.*, **31**, 869–877.
 34. Fatica, A., Morlando, M. and Bozzoni, I. (2000) Yeast snoRNA accumulation relies on a cleavage-dependent/polyadenylation-independent 3'-processing apparatus. *EMBO J.*, **19**, 6218–6229.
 35. Ofengand, J. and Bakin, A. (1997) Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaeobacteria, mitochondria and chloroplasts. *J. Mol. Biol.*, **266**, 246–268.
 36. Samarsky, D.A. and Fournier, M.J. (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **27**, 161–164.
 37. Fogel, S. and Welch, J.W. (1982) Tandem gene amplification mediates copper resistance in yeast. *Proc. Natl Acad. Sci. USA*, **79**, 5342–5346.
 38. Karin, M., Najarian, R., Haslinger, A., Valenzuela, P., Welch, J. and Fogel, S. (1984) Primary structure and transcription of an amplified genetic locus: the CUP1 locus of yeast. *Proc. Natl Acad. Sci. USA*, **81**, 337–341.
 39. Johnston, M., Hillier, L., Riles, L., Albermann, K., Andre, B., Ansorge, W., Benes, V., Bruckner, M., Delius, H., Dubois, E. *et al.* (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature*, **387**, 87–90.
 40. Silar, P., Butler, G. and Thiele, D.J. (1991) Heat shock transcription factor activates transcription of the yeast metallothionein gene. *Mol. Cell Biol.*, **11**, 1232–1238.
 41. Smith, C.M. and Steitz, J.A. (1998) Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell Biol.*, **18**, 6897–6909.
 42. Coccia, E.M., Cicala, C., Charlesworth, A., Ciccarelli, C., Rossi, G.B., Philipson, L. and Sorrentino, V. (1992) Regulation and expression of a growth arrest-specific gene (gas5) during growth, differentiation and development. *Mol. Cell Biol.*, **12**, 3514–3521.
 43. Frank, D.N., Adamidi, C., Ehringer, M.A., Pitulle, C. and Pace, N.R. (2000) Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. *RNA*, **6**, 1895–1904.
 44. Eddy, S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
 45. Bakin, A., Lane, B.G. and Ofengand, J. (1994) Clustering of pseudouridine residues around the peptidyltransferase center of yeast cytoplasmic and mitochondrial ribosomes. *Biochemistry*, **33**, 13475–13483.
 46. Malpertuy, A., Tekaia, F., Casaregola, S., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., de Montigny, J. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett.*, **487**, 113–121.
 47. Weinstein, L.B. and Steitz, J.A. (1999) Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.*, **11**, 378–384.
 48. Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R. and Dennis, P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.