Curr Genet (2005) 48: 77–87
DOI 10.1007/s00294-005-0001-x

RESEARCH ARTICLE

Zhihong Zhang · Fred S. Dietrich

Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in *Saccharomyces cerevisiae*

Received: 25 January 2005 / Revised: 23 May 2005 / Accepted: 30 May 2005 / Published online: 13 July 2005
© Springer-Verlag 2005

Abstract We have taken advantage of recently sequenced hemiascomycete fungal genomes to computationally identify additional genes potentially regulated by upstream open reading frames (uORFs). Our approach is based on the observation that the structure, including the uORFs, of the post-transcriptionally uORF regulated *Saccharomyces cerevisiae* genes *GCN4* and *CPA1* is conserved in related species. Thirty-eight candidate genes for which uORFs were found in multiple species were identified and tested. We determined by 5′ RACE that 15 of these 38 genes are transcribed. Most of these 15 genes have only a single uORF in their 5′ UTR, and the length of these uORFs range from 3 to 24 codons. We cloned seven full-length UTR sequences into a luciferase (LUC) reporter system. Luciferase activity and mRNA level were compared between the wild-type UTR construct and a construct where the uORF start codon was mutated. The translational efficiency index (TEI) of each construct was calculated to test the possible regulatory function on translational level. We hypothesize that uORFs in the UTR of *RPC11, TPK1, FOL1, WSC3*, and *MKK1* may have translational regulatory roles while uORFs in the 5′ UTR of *ECM7* and *IMD4* have little effect on translation under the conditions tested.

Keywords Yeast · Translation · Expression · Post-transcription · mRNA

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00294-005-0001-x

Communicated by S. Hohmann

Z. Zhang · F. S. Dietrich (✉)
Department of Molecular Genetics and Microbiology,
Duke University Medical Center, Durham, NC, 27710 USA
E-mail: dietr003@mc.duke.edu
Tel.: +1-919-6842857
Fax: +1-919-6811035

Introduction

The regulatory role of upstream open reading frames (uORFs) found in the 5′ UTR of mRNAs have been shown for both prokaryotes and eukaryotes. In prokaryotes, it has long been known that upstream genes in a polycistronic mRNA can regulate the expression of downstream genes (Schumperli et al. 1982; Gerstel and McCarthy 1989). In eukaryotes, uORFs have been identified in many species including human (*mdm-2*; *bcl-2*; *c-mos*; *ATF4*) (Harigai et al. 1996; Steel et al. 1996; Brown et al. 1999; Vattem and Wek 2004), *Neurospora crassa arg-2* (Wang and Sachs 1997), *Aspergillus nidulans cpcA* (Hoffmann et al. 2001), and *S. cerevisiae CLN3* (Polymenis and Schmidt 1997). Among these uORF-containing genes, the best understood regulatory models are the *S. cerevisiae GCN4* and *CPA1* genes, having four small uORFs and one relatively long uORF in respectively in their 5′ UTR, respectively. In the case of the *GCN4* transcript, four uORFs of 3~4 codons, including the stop codon, confer translational regulation by altering the efficiency of translational reinitiation in response to amino acid starvation (Hinnebusch 1997). First, uORF1 is translated, and then the translational initiation complex continues scanning until reaching an appropriate ATG where translation is reinitiated. In the presence of abundant aminoacylated tRNA, the turnover rate of eIF2·GDP to eIF2·GTP is rapid. Thus reinitiation begins upstream of uORF2, 3, and 4, where translation of uORF4 leads to translation termination. In media lacking amino acids, a 4~20-fold increase in Gcn4p results from a reduced eIF2·GDP to eIF2·GTP turnover rate and reinitiation begins downstream of uORF4, resulting in translation of *GCN4*. In this derepression transition, the *GCN4* mRNA level only increases by a factor of 2, indicating that translational regulation is the major regulator of Gcn4p levels (Albrecht et al. 1998).

In the case of *CPA1*, the single uORF encodes a 25-amino acid peptide, the arginine attenuator peptide

(AAP), which can reduce Cpa1p synthesis in response to arginine surplus (Werner et al. 1987). It has been shown that the translation of *CPA1* does not depend on reinitiation as in the case of *GCN4*, but instead depends on leaky scanning. Reduced leaky scanning results in translation of AAP and thus reduces translation of the main *CPA1* ORF (Gaba et al. 2001). Translation of the *CPA1* uORF may also trigger nonsense mediated decay (NMD) of the transcript (Messenguy et al. 2002).

In *Saccharomyces cerevisiae*, only 17 genes have been reported having uORF(s) (Vilela and McCarthy 2003). The major obstacle in identifying uORF-containing genes in *S. cerevisiae* is that for most genes, the 5′ UTR length is unknown, and thus it is likely that most uORF modulated genes are yet undiscovered. It has been predicted that there may be 200 uORF regulated *S. cerevisiae* genes (Vilela et al. 1998).

Conservation of uORFs among related species is an alternative way to find unknown uORFs. For instance, the uORFs in the vertebrate ATF4 gene (Vattem and Wek 2004), and in the fungal *GCN4* and *CPA1* genes (supplementary material) have conserved coding sequences, lengths, and approximate location relative to the main translational start codon. Recently, whole genome sequences of several hemiascomycetous fungal species have been published, including *Saccharomyces paradoxus*, *Saccharomyces bayanus, Saccharomyces mikatae, Saccharomyces kudriavzevii, Saccharomyces kluyveri, Saccharomyces castelii* (Cliften et al. 2003; Kellis et al. 2003), *Ashbya gossypii* (Dietrich et al. 2004), *Candida glabrata*, *Kluyveromyces lactis* (Dujon et al. 2004), and *Kluyveromyces waltii* (Kellis et al. 2004). All these species share significant protein similarity and extensive synteny. Using the available whole genome sequence data, we compared the upstream sequence of known *S. cerevisiae* uORF(s) containing gene orthologs from these recently sequenced species. It is noteworthy that the extent of conservation of the previously reported uORFs varies considerably, with the uORFs for *GCN4* and *CPA1* having broad conservation among species, while other uORFs are conserved only among the most closely related species, the sensu stricto group, including *S. cerevisiae*, *S. paradoxus*, *S. bayanus, S. mikatae*, and *S. kudriavzevii*. We postulate that sequence conservation can be used to identify additional uORF containing genes in *S. cerevisiae*. The existence of conservation suggests a conserved role of these uORFs among these species.

In this study, we exploit the conservation of uORFs to find new uORFs in *S. cerevisiae* using multiple alignments of the upstream regions from orthologous genes. For this project we used for a definition of "uORF" an open reading frame encoding at least two residues, starting with an ATG start codon, present within the transcribed 5′ leader sequence, and not overlapping the protein-coding region by more than one nucleotide. Whether these candidate uORFs are transcribed was tested by 5′ RACE and primer extension; a reporter system is used to identify the potential regulatory role of these uORFs.

Materials and methods

Strains and cultures

S. cerevisiae W303-1A (*MAT*a *ura3-1 leu2-3,112 trp1-1 can1-100 ade2-1 his3-11,15 [psi+]*) was the strain used in this study. Yeast was cultured in synthetic dropout media (SD-Trp) at 30°C for plasmid selection. *Escherichia coli* strain TOP10 (Invitrogen) was used in plasmid construction. All primers and constructs used in this study are listed in supplementary material.

Comparative search of uORF

The Saccharomyces sequence data were downloaded from Saccharomyces Genome Database (SGD) (Cherry et al. 1997), *A. gossypii* genome sequence was from AGD (Dietrich et al. 2004), *C. glabrata* and *K. lactis* are from Génolevures (Dujon et al. 2004; Sherman et al. 2004). Protein annotation from *S. cerevisiae* was used as the dataset to identify orthologous genes from the DNA sequence of the other species. A perl script was used to find orthologous genes in these other species using tblastn (Altschul et al. 1990). In this study it is necessary to identify corresponding 5′ UTR region in all species used, thus only the N-terminal 60 amino acids was used to identify orthologues. In order to properly align start codons, and thus properly align the untranslated regions, several relatively stringent criteria were applied: first, the tblastn alignment need to be >70% identical on amino acid level; second, the alignment need to extend more than 80% (48 amino acids); third, the starting methonine should be at similar positions (1–5 amino acids difference) between the *S. cerevisiae* gene and each orthologue. Only genes with orthologues in at least three species by these criteria, including *S. cerevisiae*, were further analyzed.

Based on this UTR alignment, UTR regions of 30~210 bp (30 bp increments) were extracted from each corresponding genome and translated in three forward frames to identify putative uORFs. uORFs were defined as having a minimal of three codons including the start and stop codons. The upstream sequence translation results were clustered by their orthologous relationship. The four sensu stricto species in addition to *S. cerevisiae* (*S. paradoxus*, *S. bayanus, S. mikatae*, and *S. kudriavzevii*) were used in this clustering. The data from two other *Saccharomyces* species, *S. kluyveri*, and *S. castelii*, as well as with *A. gossypii*, *C. glabrata, K. lactis*, and *K. waltii*, were used to validate candidate genes. The candidate selection process is based on the similarity of uORF number, length, position, and coding sequence. We looked for uORFs with lengths differing by no more than ±3 amino acids, and position differences relative to the start codon of no more than ±10 bp. To determine similarity among these uORFs in each cluster, the translated sequences of these uORFs were aligned. The relative position of the starting ATG in each cluster, relative to the ATG of the main open reading frame, and the lengths of the uORFs were also determined. For coding sequence similarity, in the case of long uORFs (>6 codon), we confirmed that the uORFs were more than 50% identical. To account for possible sequencing errors, we allowed one exception, so that, for each uORF one sensu stricto species might not match the above criteria. For each *n*-member orthologue group, in addition to at least $n-1$ of the sensu stricto species containing a conserved uORF by the above criteria, at least one nonsensu stricto species had a uORF in the orthologous position, though similarity of the coding sequence was not required.

5′ RACE

5′ RACE for each candidate gene was done using the SMART-RACE kit (BD Bioscience Clontech). Briefly, 1 µg total RNA purified by RNeasy Mini kit (Qiagen) from *S. cerevisiae* was reverse transcribed into cDNA by 5′ CDS primer (oligo-dT). A special template switching oligo adapter was added to 3′ of first strand cDNA. A gene specific CDS oligo and a universal adapter oligo were used to amplify a fragment containing 5′ end sequence of mRNA. The PCR product was cloned to pCR2.1-TOPO (Invitrogen) and sequenced (ABI genetic analyzer 310).

Primer extension

The AMV Primer Extension Kit (Promega) was used. Following the manufacturer's protocol, 30 µg total RNA purified by RNeasy Mini kit (Qiagen) from *S. cerevisiae* was reverse transcribed by annealing with gene specific ^{32}P labeled oligo under 42°C for 30 min. The extension product was separated on 8% denaturing polyacrylamide gel containing 7 M urea by electrophoresis and detected on auto-radiographic film.

Calculation of the AUG context adaptation index (A_{UG}CAI)

The method and scoring matrix of Miyasaka (1999) was used to calculate the A_{UG}CAI.

Cloning and site-directed mutagenesis

Standard molecular biology protocols were used in cloning UTR sequences into *Bam*HI/*Nde*I linearized YCpFL′ vector. After verification by PCR and sequencing, the site-directed mutagenesis were done by QuickChange Site-directed mutageneisis kit (Stratagene) or recombinant PCR (Ho et al. 1989).

Luciferase (LUC) assay

The protocol is adapted from previous reports (Oliveira et al. 1993) and manufacturer protocols. Briefly, *S. cerevisiae* cells carrying various reporter plasmid constructs were harvested from 5 ml SD-Trp culture (Qbiogene) at $OD_{600}=1.0$ and washed twice with autoclaved deionized water. The cells were resuspended in 400 µl extraction buffer (0.1 M KH_2PO_4, 1 mM DTT, pH 7.8). About 400 mg of glass beads (425–600 micron, Sigma) were added and cells were broken by vortexing four times for 30 s with 30 s interval at 4°C. The cell extracts were centrifuged at 14,000 rpm for 5 min at 4°C. Each 5 µl of supernatant was added to 50 µl of LUC reaction solution (Promega). The LUC activities were measured by luminometer (Turner Biosystem Inc, Model TD-20/20). The total protein concentration of cell extracts was measured by Quick Start Bradford Protein Assay (BioRad) for normalization purpose.

Northern blot

The CDP-star nonradioactive northern kit (Amersham) was used. Probes matching to the firefly LUC gene (LUC, GenBank accession number M15077) coding region and *S. cerevisiae ACT1* (GenBank accession number L00026) were generated from PCR and labeled according the protocol. The 30 µg total RNA of each strain purified at the same time of LUC assay were loaded on 1% formaldehyde agarose gel, and transferred to nylon membrane which was described previously (Zhang and Dietrich 2003). The hybridization and detection procedure followed manufacturer protocols.

Results

Identification of potential uORF containing genes

For each species, upstream regions of length 30, 60, 90, 120, 150, 180, and 210 bp were extracted from genes with orthologues in *S. cerevisiae*. The number of putative orthologous gene pairs between *S. cerevisiae* and the other sensu stricto species was *S. cerevisiae*–*S. paradoxus* 5,272 pairs, *S. cerevisiae*–*S. bayanus* 5,031 pairs, *S. cerevisiae*–*S. mikatae* 4,836 pairs; *S. cerevisiae*–*S. kudriavzevii* 4,921 pairs; *S. cerevisiae*–*S. castelii* 2,944 pairs, and *S. cerevisiae*–*S. kluyveri* 2,642 pairs. The number of orthologous pairs used in this study is less than the total number of orthologous pairs, as in some cases alignment of the start codon was problematic. For *S. cerevisiae*, 5,542 genes passed the tblastn filter described in the material and methods section. From these upstream regions, uORFs were identified. The percentage of genes containing one or more putative uORF in the upstream region of each length for each species was calculated. We

Table 1 uORF containing genes verified by 5′ RACE

ORF	Name	5′ UTR length (bp) (by 5′ RACE)	Number and size of uORFs[a]	Position	A_{UG}CAI[b]	PE[c]	Expression levels (mRNA copies/cell)[d]	Gene product
YLR242C	*ARV1*	156	uORF1(12), uORF2(3), uORF3(7)	−125, −108, −40	0.208, 0.452, 0.192, *0.456*	No	0.4	Protein involved in sterol distribution
YLR443W	*ECM7*	152	uORF(5)	−15	0.344, *0.464*	Yes	0.9	Nonessential protein of unknown function
YDL205C	*HEM3*	176	uORF(9)	−129	0.193, *0.225*	No	0.3	Heme biosynthesis
YDR045C	*RPC11*	154	uORF(4)	−60	0.261, *0.281*	Yes	5.3	TFIIS-like small Pol III subunit C11
YEL064C	*AVT2*	34	uORF(4)	−11	0.308, *0.426*	No	0.4	Amino acid vacuolar transport
YJL164C	*TPK1*	232	uORF(5)	−42	0.179, *0.259*	Yes	0.9	cAMP-dependent protein kinase catalytic subunit (putative)
YKL093W	*MBR1*	90	uORF(7)	−70	0.237, *0.336*	No	0.1	Involved in mitochondrial biogenesis
YLR127C	*APC2*	52	uORF(5)	−27	0.335, *0.361*	No	0.2	Anaphase promoting complex (APC) subunit
YLR146C	*SPE4*	44	uORF(6)	−41	N/A[2], *0.408*	No	1.5	Spermine synthase
YLR313C	*SPH1*	39	uORF(4)	−25	0.399, *0.556*	No	0.3	SPa2-Homolog; protein involved in shmoo formation and required for bipolar bud site selection
YML056C	*IMD4*	120	uORF(14)	−99	0.660, *0.363*	Yes	7.7	Inosine monophosphate dehydrogenase, catalyzes the first step of GMP biosynthesis
YNL047C	*SLM2*	176	uORF1(24), uORF2(6), uORF3(4)	−110, −84, −70	0.146, 0.240, 0.273, *0.403*	No	0.6	Uncharacterized ORF
YNL256W	*FOL1*	91	uORF(4)	−65	0.302, *0.320*	Yes	0.7	Folic acid biosynthesis pathway
YOL105C	*WSC3*	248	uORF(7)	−50	0.228, *0.622*	No	0.3	Cell wall integrity and stress response component 3
YOR231W	*MKK1*	102	uORF(10)	−71	0.297, *0.382*	Yes	0.8	MAP kinase kinase (MEK)
YKL109W	*HAP4*	256	uORF1(10), uORF2(4)	−249, −60	0.221, 0.323, *0.280*	Yes	1.0	Transcriptional activator protein of *CYC1*
YEL009C	*GCN4*	575	uORF(4), uORF2(2), uORF3(4), uORF4(4)	−361, −293, −176, −151	0.751, 0.335, 0.273, 0.493, *0.707*	Yes	22.3	Transcriptional activator of amino acid biosynthetic genes

[a] uORF is numbered from 5′ to 3′ of the transcript, and the length (codon) is shown in the parenthesis. The size of uORFs includes stop codon

[b] A_{UG}CAI: The AUG codon adaptation index was calculated as described (Miyasaka 1999). For *SPE4*, since the uORF start at only three nucleotids downstream of the transcript 5′ end, the CAI calculation is not applicable. The number in italic is the main coding region A_{UG}CAI

[c] PE: Primer extension verified or not

[d] Expression level data were obtained from whole genome gene expression microarray analysis (http://web.wi.mit.edu/young/expression/) (Holstege et al. 1998)

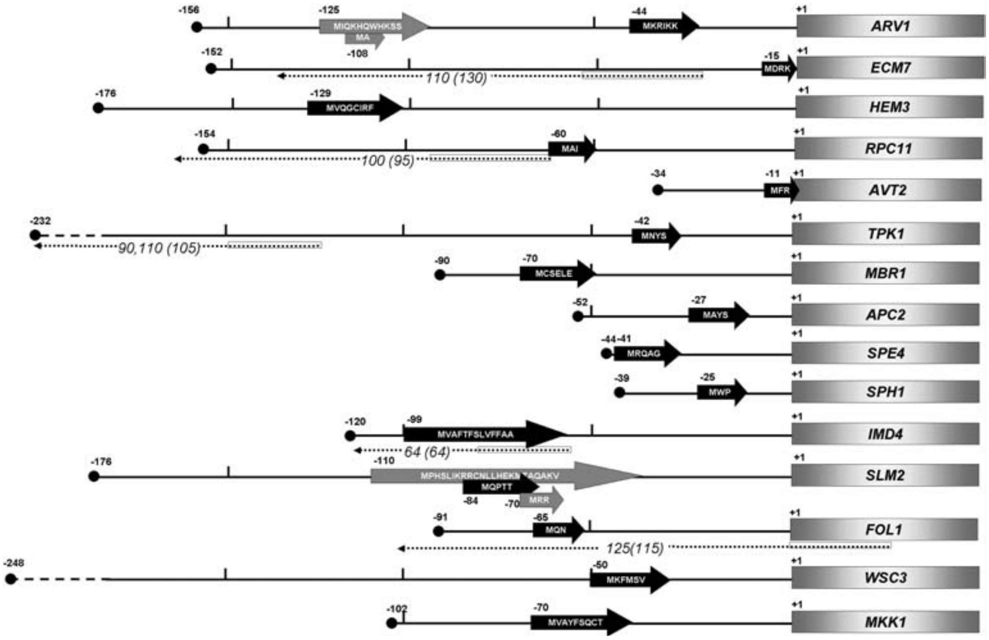

Fig. 1 Graphic presentations of 15 newly found uORF-containing *S. cerevisiae* genes. The position and the amino acid sequence of each uORF are shown drawn to scale. The *black arrows* indicate that the uORF is conserved in at least four of the five sensu stricto species. *Gray arrows* indicate uORF in *S. cerevisiae* is not conserved in two or more of the other sensu stricto species. The *dotted line* and the *rectangular box* below the transcript represent primer extension products and primer extension primers, respectively. The *number under the primer extension arrow* is the observed length (nt) of the primer extension product and the *number in the parenthesis* is the expected size based on the 5′ RACE results

observed that uORFs are present in over 95% of 250 bp 5′ upstream regions of *S. cerevisiae*. Since most previously identified gene transcription start sites (TSS) are less than 200 bp from the start codon (Hampsey 1998), and with the high percentage of genes with presumably spurious potential uORFs when longer upstream regions were selected, a 210-bp 5′ upstream region was chosen as the upper boundary for this comparative uORF analysis. The sensu stricto uORF translation results were clustered based on orthologous relationship and extracted upstream sequence length. Each gene cluster contains at least two sensu stricto species genes in addition to *S. cerevisiae*, and each cluster contains at least one putative uORF. A total of 19 gene clusters containing uORFs were found using a 30-bp upstream window, 116 for 60 bp, 445 for 90 bp, 1,012 for 120 bp, 1,676 for 150 bp, 2,354 for 180 bp, and 2,957 for 210 bp (supplementary material). Alignment of noncoding regions is more difficult than alignment of coding regions as they are more diverged, and contain more small insertions/deletions, making precise comparison of upstream positions between species difficult. “Conservation” here is defined as uORFs sharing similar length, position relative to the main ATG, and amino acid sequence; the first two factors were weighted more heavily than sequence conservation of the translated product. The primary candidates were validated through comparing uORF pattern to the orthologous gene of two other *Saccharomyces* species, *S. kluyveri* and *S. castelii*, as well as *A. gossypii*, *C. glabrata*, *K. lactis*, and *K. waltii*. For these non*Saccharomyces* species, the orthologous relationship was according to published data (Brachat et al. 2003; Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004). In addition to conservation among the sensu stricto species, for a uORF to be considered a “good” candidate, the length and position of the uORF should be similar in one of these species. Finally, a total of 38 *S. cerevisiae* gene candidates were selected according to the criteria described above (supplementary material).

Experimental verification of candidate genes

We performed 5′ RACE to determine if the predicted uORFs are encoded within the 5′ UTR. For this purpose, gene specific primers, close to the 5′ end of the coding region were designed to clone the 5′ transcript ends by 5′ RACE. The cloned 5′ RACE products of the candidate genes were sequenced to determine the TSS. Among the 38 candidates, we successfully cloned the 5′ end of 22 genes and verified that 15 genes of these 22 have real uORFs on their transcripts. The GenBank submission numbers are listed in supplementary material.

Among these 15 newly verified uORF containing genes (Table 1, Fig. 1), most of them have only a single uORF in the 5′ UTR, with only *ARV1*, and *SLM2* having multiple uORFs. The length of these uORFs ranges from 3 to 24 codons. The lengths of confirmed 5′ UTRs ranges from 34 to 248 nucleotides. The distance from the transcript 5′ end to the first uORF start codon varies from 198 nucleotides of *WSC3* to 3 nucleotides of *SPE4*.

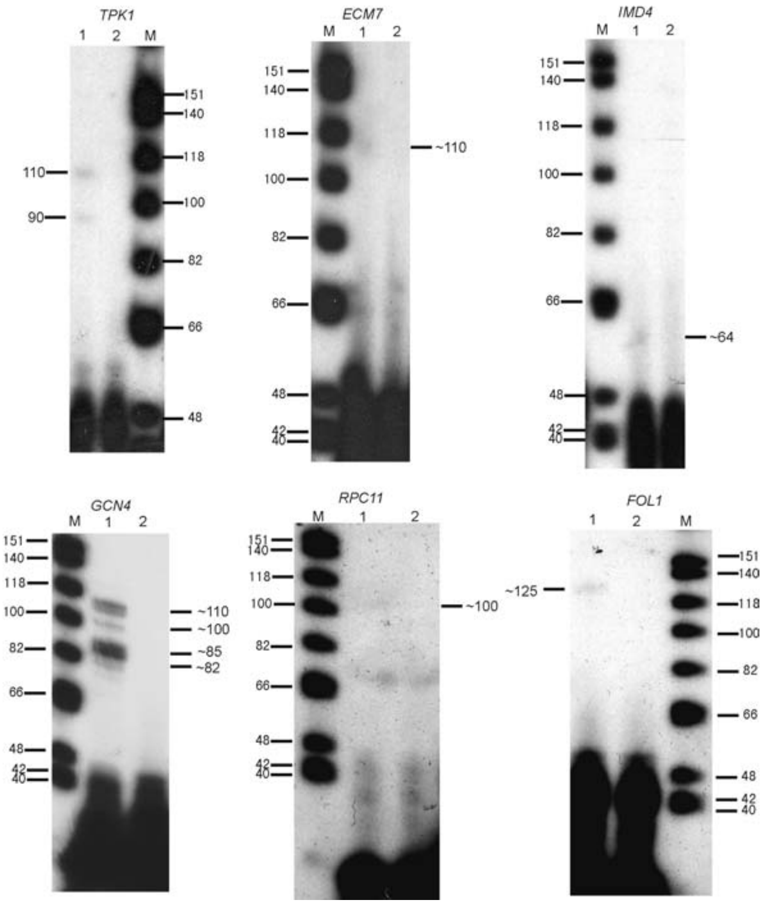

Fig. 2 Transcription start site mapping by primer extension. The primer extension products of five candidates gene along with one positive control, *GCN4*, are shown. *Lane M* Φ 174 *Hin*fI DNA markers; *lane 1* Primer extension reaction; *lane 2* Reaction without RNA (Negative control). All numbers shown in figures are length of single strand DNA (nt)

For these 15 genes, we performed primer extension experiments to verify the TSS. Gene specific primers close to the 5′ RACE predicted end were synthesized to carry out reverse transcription using AMV reverse transcriptase. Among these 15, five genes generated bands and the molecular marker loading on the same TBE-Urea PAGE gel to confirm that the band sizes were similar to the expected sizes based on the 5′ RACE (Fig. 2), with two genes primer extension results suggesting a longer 5′ UTR than that found by 5′ RACE. Despite several attempts, primer extension determination of the TSS was not successful for the ten genes with putative uORFs but with expression levels reported as less than 0.7 copies/cell (Holstege et al. 1998). In order to confirm the reliability of the methods used, we also tested three previously reported uORF(s)-containing genes (Vilela and McCarthy 2003) by 5′ RACE, which includes *GCN4*, *HAP4*, and *TIF4631*.

Investigation of roles of uORF(s) in gene expression

In order to elucidate the role of the uORFs identified in this study, a LUC reporter system combined with site-directed mutagenesis of uORF start codons was used (Fig. 3a). We focused on seven genes plus *GCN4* and *HAP4* as the positive control. Among these seven genes, in five cases the 5′ UTR was verified by both 5′ RACE and primer extension, the exceptions being *MKK1*, which was verified only by 5′ RACE, and *WSC3*, which was verified twice by 5′ RACE.

In addition to replacing the original 5′ UTR of LUC gene of the reporter system with each gene's full-length 5′ UTR, we also mutated the start codon of each uORF by site-directed mutagenesis or recombinant PCR. For *GCN4*, a truncated 5′ UTR containing uORF3 and uORF4 was also cloned into reporting system (*GCN4′*) and the start codon of uORF4 was mutated in a construct (*GCN4**).

The LUC activity and mRNA level were measured and compared by LUC assay and Northern blot (Fig. 3b), respectively. The LUC assay was repeated three times and the Northern Blot twice. The standard deviation was calculated after normalization with the vector-only control. The translational efficiency index (TEI) was used to evaluate the ability of translation provide with same amount of messenger RNA (Fig. 4). The TEI of full length GCN4UTR-LUC is 5%, the TEI of the truncated uORF3-and-4-only GCN4UTR-LUC (*GCN4′*) is 11%, and uORF3-only GCN4UTR-LUC (*GCN4**) is 26% of positive control, consistent with published data.

The Northern blot result in Fig. 3b provides mRNA levels of each UTR-LUC gene. After site-directed

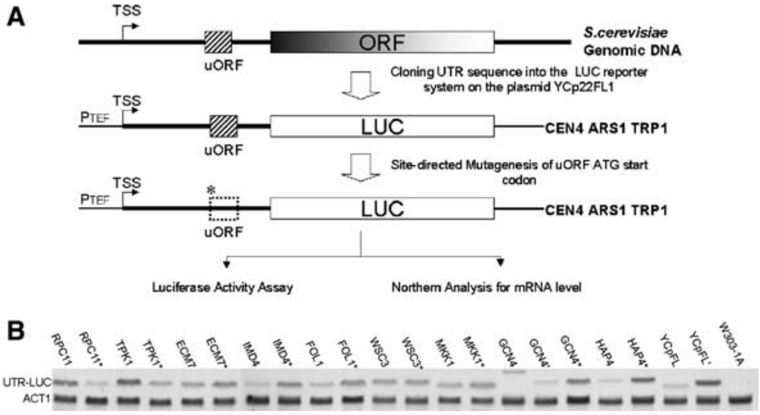

Fig. 3 Functional studies of uORF roles by LUC reporter gene and Northern blot. **a** YCp22FL1 (YCpFL′) system (Rajkowitsch et al. 2004) was used to find potential regulatory function of uORFs. Full-length uORF containing UTR sequences were cloned into YCpFL′ system, containing firefly LUC reporter gene. Site-direct mutagenesis was used to eliminate the uORF by mutating the ATG start codon. Each pair of constructs was transformed into *S. cerevisiae* W303-1A strain to measure the LUC enzyme activities and mRNA levels. *Cross-hatched boxes* indicate uORF; *dotted boxes* indicate absence of uORF resulting from mutagenesis of start codon. *TSS* Transcription start site; **b** Northern blot result. The *ACT1* gene was used as internal control in Northern blot. Each *GENE** indicates the uORF start codon in the full length UTR has been completely mutated except that *GCN4** contains truncated *GCN4*-UTR with uORF3 ATG intact but uORF4 ATG mutated. The *GCN4′* construct contains truncated *GCN4* UTR with uORF3 and 4 only

mutagenesis of the ATG start codon of each uORF, most genes have higher mRNA level detected and increased LUC activity. In particular, for the results shown in Fig. 4, mRNA level and LUC activity can be classified as follows:

1. *FOL1*, *MKK1*, and *GCN4*. Both mRNA and protein increased after uORF start codon mutation, but the protein activities increased more than that of mRNA level (TEI > 1), which indicated both steady state mRNA level and translation enhanced.
2. *ECM7*, *IMD4*, and *HAP4*. Both mRNA and protein increased after uORF start codon mutation, while the TEI was about the same between original and mutated construct.
3. *WSC3*. The mRNA level is unchanged while protein activity increased, thus translation is enhanced. This is similar to the category I.
4. *RPC11*, *TPK1*. The mRNA level decreased though protein level was unchanged (TEI > 1).

Discussion

uORF conservation suggests they are functional entities

In this study, we have identified a total of 15 new uORF containing gene in *S. cerevisiae*, nearly doubling the number of published uORFs. The uORFs of these genes are conserved among the sensu stricto species, and are conserved in at least one other hemiascomycete. These uORFs tend to be conserved in fewer species than *GCN4* and *CPA1*. For example, the *RPC11* uORF was found in *S. kluyveri*, but not in *S. castelii* (supplementary data).

Analysis of *S. cerevisiae* upstream sequence and that of related species indicates that the putative uORF occurrence rate is quite high particularly for longer hypothetical UTR's. However, most of these putative uORF's are spurious as they do not fall within the transcribed regions of the genome. Comparative analysis was able to exclude many of these spurious potential uORFs, though likely excluded some real uORFs as well. Thus it is likely that *S. cerevisiae* has more uORFs yet to be discovered. As shown in Table 1, the A_{UG}CAI of the uORF start codons is similar to that of coding region start codons of genes expressed at low levels (Miyasaka 1999; Gaba et al. 2001). An initial analysis of the uORF start codons revealed no apparent contextual pattern to distinguish them from coding region start codons (Table 1).

The primary confirmation of these uORFs involved determination of the TSS, something not currently known for most *S. cerevisiae* genes. A thorough analysis of TSS in *S. cerevisiae* would be useful in identifying more uORF containing genes. Recently, a new whole genome TSS mapping methodology, 5′ SAGE, has been developed (Shiraki et al. 2003; Hashimoto et al. 2004; Wei et al. 2004), and a similar method has recently been applied in *S. cerevisiae* (Zhang and Dietrich 2005). 5′ SAGE data, combined with sequence comparison of multiple species, and confirmation by gene specific RACE and primer extension experiment, has the potential to greatly increase our understanding of the regulatory potential of uORFs.

5′ RACE and primer extension verification results are mostly consistent

We selected 38 candidate genes to verify the predicted uORF by TSS mapping. Initial mapping of the TSS was carried out by 5′ RACE with further verification by primer extension. For 15 genes, 5′ RACE confirmed that

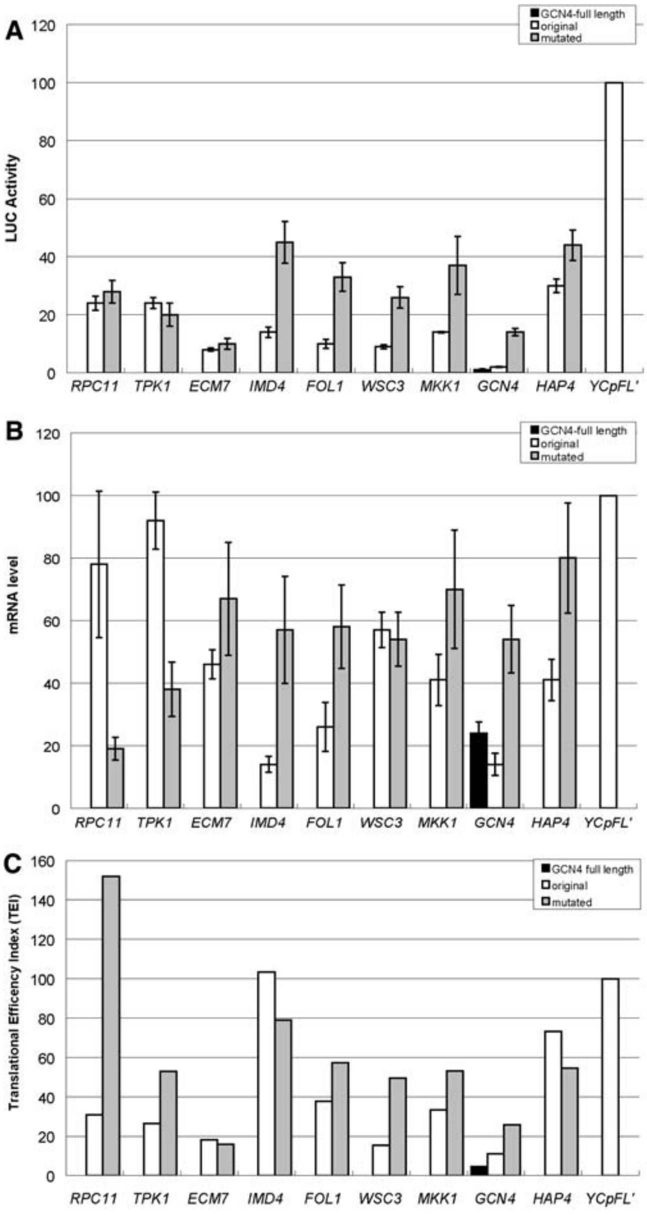

Fig. 4 UTR-LUC constructs have variable mRNA level and LUC activities. **a** The LUC activity was normalized relative to the YCpFL′ control by total protein concentration of cell lysate. **b** *ACT1* was used to normalize concentrations to adjust for differences in total RNA sample loading. UTR-LUC gene expression was normalized relative to the YCpFL′ LUC expression level. **c** The TEI was calculated with the formula: $\mathrm{TEI} = \frac{\mathrm{LUCActivity(normalized)}}{\mathrm{mRNAdensity(normalized)}} \times 100$. For *GCN4*, the full length *GCN4* UTR-LUC is shown with a *black bar*. The truncated *GCN4* uORF3, 4-only—LUC is marked as “original” (*GCN4′* in Fig. 3). And the mutated version of *GCN4′* containing uORF3-only (*GCN4** in Fig. 3) is marked as “mutated”

the TSS is upstream of the predicted uORF. Of the remaining 23 candidate genes, for 16 genes no clear 5′ RACE results were obtained, and for 7 genes the 5′ RACE results showed the TSS as downstream of the start codon of the candidate uORFs. That some of the candidate uORFs would turn out to be false positives was anticipated due to the high frequency of start and stop codons in these A-T rich noncoding regions. Among these 15 genes with uORFs based on 5′ RACE, five were supported by primer extension. The results of 5′ RACE and primer extension of these five genes were generally consistent, with primer extension determined TSS differing from those identified by 5′ RACE by less than 20 nt; these discrepancies did not change the conclusion that these uORFs exist. The reason for this discrepancy between primer extension and 5′ RACE was unclear, but may be an artifact of the multiple TSS of *S. cerevisiae* genes. Most yeast genes have multiple TSS (Zhang and Dietrich 2005), and both primer extension and 5′ RACE potentially may be biased towards different TSS. Previous reports indicate that for the *TIF4631* TSS (Goyer et al. 1993; Zhou et al. 2001) are at 508 bp and 295 bp upstream of main ATG, whereas our 5′ RACE results only detected a short UTR (~38 nt, data not shown) not containing a uORF. Our result is consistent with a more recently report (Verge et al. 2004). This may be a case of 5′ RACE bias and thus needs further investigation. It is also possible that our sequenced 5′ RACE clones and those of Verge et al. arise from premature termination of reverse transcription.

Primer extension may be a more reliable approach to mapping TSS than 5′ RACE, as it has linear relationship to the original mRNA population and thus the diversity of multiple TSS of unequal usage may be better preserved than during the exponential amplification phase of 5′ RACE. However, primer extension experiments can only detect genes expressed at relatively high levels. The nine nondetected candidate genes are expressed at less than 0.7 copies/cell mRNA (Holstege et al. 1998) in culture conditions similar to those used in this study.

Some uORFs located on transcripts are within 20 bp of the TSS

For six of the eight genes not tested by the LUC reporter system in this study, the distances between the 5′ end of transcript and the first uORF ATG is greater than 20 bp. For *SPE4* and *SPH1*, similar to *DCD1* (McIntosh and Haynes 1986), the distance from the TSS to the uORF is 3 bp and 14 bp, respectively. The TSS to start codon distance necessary for the ATG start codon recognition by the initiation complex has been shown to be about 15~20 bp (van den Heuvel et al. 1989). Thus, it is possible the translational initiation complex does not recognize these uORFs. Further study is required to determine if uORFs within the first 20 nucleotides of a transcript are regulatory.

Primary characterization of uORF roles suggests the translation attenuation in gene expression

For most published experimental results, the existence of uORFs results in translation attenuation as long as the

ATG start codon is recognized by the scanning ribosome (Vilela and McCarthy 2003). The initial characterization of uORF function shows that for several genes (*RPC11*, *TPK1*, *FOL1*, *WSC3*, and *MKK*), the mRNA translation efficiency increased after the uORF ATG was mutated. This suggests that these uORFs function to modulate the translation of these genes, possibly through a leaky scanning mechanism (Kozak 2002). A leaky scanning mechanism is consistent with the low A_{UG}CAI values of the majority of the uORFs reported here.

The interpretation of the role of these uORFs is based on the assumption that translation occurs through the general eukaryotic mechanism in which the ribosome associates with the message in a cap-dependent fashion and scans to the first start codon (Pestova et al. 2001). By this model, leaky scanning allows the uORF to be skipped at some frequency that possibly changes in a regulated fashion. Alternatively these transcripts may contain an internal ribosome entry site (IRES) so that the ribosome would associate with the transcript after the uORF. In *S. cerevisiae* there have been several reports of IRES containing genes in including *URE2* (Komar et al. 2003), *HAP4*, TFIID (Iizuka et al. 1994), *YAP1*, and *TIF4631* (Zhou et al. 2001). It is possible that one or more of these uORF-containing genes reported here may have an IRES.

One of the genes in this study, *RPC11*, is an essential gene in *S. cerevisiae* (Chedin et al. 1998). The coding protein, Rpc11p, is a component of RNA polymerase III. The role Rpc11p is in nascent RNA hydrolysis coupled with polymerase retraction from the DNA template (Chedin et al. 1998). In *GCN4* translational regulation, under derepression condition, the accumulation of uncharged tRNA caused by amino acid starvation activates Gcn2p, a protein kinase in charge of eIF2α phosphorylation (Hinnebusch 1997). This event allows ribosome to regain reinitiation ability more slowly, thus bypassing uORF3 and uORF4 before reinitiation at the main *GCN4* start codon. As reinitiation, at least in the case of GCN4, involves levels of charged tRNAs, it is possible that charged tRNA levels may provide feedback regulation of *RPC11*.

WSC3 is a member of a family of genes responsible for maintaining cell wall integrity. Wsc3p is a stress response receptor in the cell membrane. Over-expression of the *WSC* genes can inhibit cell growth (Lodder et al. 1999), while the *WSC1ΔWSC2ΔWSC3Δ* triple mutant causes a cell lysis defect and heat shock sensitivity (Verna et al. 1997). The WSC family is believed to be upstream activator of PKC1-MAPK cascade and RAS-cAMP pathway (Verna et al. 1997; Wojda et al. 2003). Interestingly, two other uORF-containing genes, *MKK1* and *TPK1*, are major kinase components of these two pathways, respectively. The LUC assay and mRNA level detection of this study shows that for these three genes TEIs all increased after the uORF ATG was mutated. That three genes containing uORFs are involved in these pathways suggests that post-transcriptional regulation may be important in signal transduction pathways.

Previous work had shown that the uORF(s) of some genes, including *CPA1* and *YAP2*, act to destabilize mRNA (Vilela et al. 1999; Ruiz-Echevarria and Peltz 2000; Messenguy et al. 2002). In this study, in comparison to the vector only control (YCpFL′), we observed that most candidates had reduced mRNA level of the LUC gene. We also observed that the UTR-LUC construct transcripts containing the *ECM7*, *IMD4*, *FOL1*, and *MKK1* 5′ leader appeared to be more stable after the uORF ATG codon was mutated. This suggests uORFs may have a role in modulating mRNA stability, and thus influence gene expression. Characterization of the structure and regulatory capacity of the UTR regions of genes complements other approaches to investigating gene regulation. As microarray and proteomics studies suggest the importance of post-transcriptional regulation for many genes in *S. cerevisiae* (Washburn et al. 2003), understanding the role of uORFs in translational regulation is an important component of understanding overall gene regulation.

Acknowledgements We thank John E.G. McCarthy at UMIST for providing YCp22FL series plasmids. We thank Joe Heitman for providing yeast strains and Miguel Arevalo-Rodriguez, Shihua Lu, and Carol Gallione for valuable technical help and advice. We are grateful to Joe Heitman, John McCusker, Douglas Marchuk, and Bryan Cullen and for generously providing laboratory facilities and comments on the whole project. The authors also thank Mark DeLong for his careful reading and revising of the manuscript.

References

Albrecht G, Mosch HU, Hoffmann B, Reusser U, Braus GH (1998) Monitoring the Gcn4 protein-mediated response in the yeast *Saccharomyces cerevisiae*. J Biol Chem 273:12696–12702

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T, Philippsen P (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. Genome Biol 4:R45

Brown CY, Mize GJ, Pineda M, George DL, Morris DR (1999) Role of two upstream open reading frames in the translational control of oncogene mdm2. Oncogene 18:5631–5637

Chedin S, Riva M, Schultz P, Sentenac A, Carles C (1998) The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination. Genes Dev 12:3857–3871

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 387:67–73

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–76

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304:304–307

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004) Genome evolution in yeasts. Nature 430:35–44

Gaba A, Wang Z, Krishnamoorthy T, Hinnebusch AG, Sachs MS (2001) Physical evidence for distinct mechanisms of translational control by upstream open reading frames. EMBO J 20:6453–6463

Gerstel B, McCarthy JE (1989) Independent and coupled translational initiation of atp genes in *Escherichia coli*: experiments using chromosomal and plasmid-borne lacZ fusions. Mol Microbiol 3:851–859

Goyer C, Altmann M, Lee HS, Blanc A, Deshmukh M, Woolford JL Jr, Trachsel H, Sonenberg N (1993) *TIF4631* and *TIF4632*: two yeast genes encoding the high-molecular-weight subunits of the cap-binding protein complex (eukaryotic initiation factor 4F) contain an RNA recognition motif-like sequence and carry out an essential function. Mol Cell Biol 13:4860–4874

Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev 62:465–503

Harigai M, Miyashita T, Hanada M, Reed JC (1996) A cis-acting element in the BCL-2 gene controls expression through translational mechanisms. Oncogene 12:1369–1374

Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K (2004) 5′-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol 22:1146–1149

van den Heuvel JJ, Bergkamp RJ, Planta RJ, Raue HA (1989) Effect of deletions in the 5′-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene 79:83–95

Hinnebusch AG (1997) Translational regulation of yeast *GCN4*. A window on factors that control initiator-tRNA binding to the ribosome. J Biol Chem 272:21661–21664

Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene 77:51–59

Hoffmann B, Valerius O, Andermann M, Braus GH (2001) Transcriptional autoregulation and inhibition of mRNA translation of amino acid regulator gene cpcA of filamentous fungus *Aspergillus nidulans*. Mol Biol Cell 12:2846–2857

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

Iizuka N, Najita L, Franzusoff A, Sarnow P (1994) Cap-dependent and cap-independent translation by internal initiation of mRNAs in cell extracts prepared from *Saccharomyces cerevisiae*. Mol Cell Biol 14:7322–7330

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Komar AA, Lesnik T, Cullin C, Merrick WC, Trachsel H, Altmann M (2003) Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells. EMBO J 22:1199–1209

Kozak M (2002) Pushing the limits of the scanning mechanism for initiation of translation. Gene 299:1–34

Lodder AL, Lee TK, Ballester R (1999) Characterization of the Wsc1 protein, a putative receptor in the stress response of *Saccharomyces cerevisiae*. Genetics 152:1487–1499

McIntosh EM, Haynes RH (1986) Sequence and expression of the dCMP deaminase gene (DCD1) of *Saccharomyces cerevisiae*. Mol Cell Biol 6:1711–1721

Messenguy F, Vierendeels F, Pierard A, Delbecq P (2002) Role of RNA surveillance proteins Upf1/CpaR, Upf2 and Upf3 in the translational regulation of yeast *CPA1* gene. Curr Genet 41:224–231

Miyasaka H (1999) The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*. Yeast 15:633–637

Oliveira CC, van den Heuvel JJ, McCarthy JE (1993) Inhibition of translational initiation in *Saccharomyces cerevisiae* by secondary structure: the roles of the stability and position of stem-loops in the mRNA leader. Mol Microbiol 9:521–532

Pestova TV, Kolupaeva VG, Lomakin IB, Pilipenko EV, Shatsky IN, Agol VI, Hellen CU (2001) Molecular mechanisms of translation initiation in eukaryotes. Proc Natl Acad Sci USA 98:7029–7036

Polymenis M, Schmidt EV (1997) Coupling of cell division to cell growth by translational control of the G1 cyclin *CLN3* in yeast. Genes Dev 11:2522–2531

Ruiz-Echevarria MJ, Peltz SW (2000) The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell 101:741–751

Schumperli D, McKenney K, Sobieski DA, Rosenberg M (1982) Translational coupling at an intercistronic boundary of the *Escherichia coli* galactose operon. Cell 30:865–871

Sherman D, Durrens P, Beyne E, Nikolski M, Souciet JL (2004) Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. Nucleic Acids Res 32(Database issue):D315–D318

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100:15776–15781

Steel LF, Telly DL, Leonard J, Rice BA, Monks B, Sawicki JA (1996) Elements in the murine c-mos messenger RNA 5′-untranslated region repress translation of downstream coding sequences. Cell Growth Differ 7:1415–1424

Vattem KM, Wek RC (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc Natl Acad Sci USA 101:11269–11274

Verge V, Vonlanthen M, Masson JM, Trachsel H, Altmann M (2004) Localization of a promoter in the putative internal ribosome entry site of the *Saccharomyces cerevisiae TIF4631* gene. RNA 10:277–286

Verna J, Lodder A, Lee K, Vagts A, Ballester R (1997) A family of genes required for maintenance of cell wall integrity and for the stress response in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 94:13804–13809

Vilela C, McCarthy JE (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Mol Microbiol 49:859–867

Vilela C, Linz B, Rodrigues-Pousada C, McCarthy JE (1998) The yeast transcription factor genes *YAP1* and *YAP2* are subject to differential control at the levels of both translation and mRNA stability. Nucleic Acids Res 26:1150–1159

Vilela C, Ramirez CV, Linz B, Rodrigues-Pousada C, McCarthy JE (1999) Post-termination ribosome interactions with the 5′ UTR modulate yeast mRNA stability. EMBO J 18:3139–3152

Wang Z, Sachs MS (1997) Ribosome stalling is responsible for arginine-specific translational attenuation in *Neurospora crassa*. Mol Cell Biol 17:4904–4913

Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR III (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100:3107–3112

Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci USA 101:11701–11706

Werner M, Feller A, Messenguy F, Pierard A (1987) The leader peptide of yeast gene *CPA1* is essential for the translational repression of its expression. Cell 49:805–813

Wojda I, Alonso-Monge R, Bebelman JP, Mager WH, Siderius M (2003) Response to high osmotic conditions and elevated temperature in *Saccharomyces cerevisiae* is controlled by intracellular glycerol and involves coordinate activity of MAP kinase pathways. Microbiology 149:1193–1204

Zhang Z, Dietrich FS (2003) Verification of a new gene on *Saccharomyces cerevisiae* chromosome III. Yeast 20:731–738

Zhang Z, Dietrich FS (2005) Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. Nucleic Acids Res 33:2838–2851

Zhou W, Edelman GM, Mauro VP (2001) Transcript leader regions of two *Saccharomyces cerevisiae* mRNAs contain internal ribosome entry sites that function in living cells. Proc Natl Acad Sci USA 98:1531–1536