Systematic Discovery of New Genes in the *Saccharomyces cerevisiae* Genome

Marco M. Kessler, Qiandong Zeng, Sarah Hogan, et al.

Genome Res. 2003 13: 264-271
Access the most recent version at doi:10.1101/gr.232903

Systematic Discovery of New Genes in the *Saccharomyces cerevisiae* Genome

Marco M. Kessler, Qiandong Zeng, Sarah Hogan, Robin Cook, Arturo J. Morales,[1] and Guillaume Cottarel

Genome Therapeutics Corporation, Waltham, Massachusetts 02453, USA

We used genome-wide comparative analysis of predicted protein sequences to identify many novel small genes, named smORFs for small open reading frames, within the budding yeast genome. Further analysis of 117 of these new genes showed that 84 are transcribed. We extended our analysis of one smORF conserved from yeast to human. This investigation provides an updated and comprehensive annotation of the yeast genome, validates additional concepts in the study of genomes in silico, and increases the expected numbers of coding sequences in a genome with the corresponding impact on future functional genomics and proteomics studies.

Current consensus suggests the number of yeast genes to be very close to 6000 (Goffeau et al. 1996; Mewes et al. 1997; Winzeler and Davis 1997). The number of protein-coding genes with open reading frames (ORFs) longer than 100 codons is predicted to be between 5300 and 5400 (Mackiewicz et al. 2002). Two recent studies suggest that the determination of gene numbers by stringent gene identification methods may underestimate the number of genes in human and other organisms (Gopal et al. 2001; Reboul et al. 2001). The initial annotation of the budding yeast genome did not include many short open reading frames (ORFs; Andrade et al. 1997; Olivas et al. 1997). Recent experimental studies designed to catalog all genome transcripts using SAGE technology (Velculescu et al. 1997; Basrai et al. 1999) and the analysis of a collection of transposon insertions (Ross-Macdonald et al. 1999) have discovered new ORFs that were not previously identified in silico. This pool of genes includes some that code for putative proteins that are shorter than 100 amino acids (Velculescu et al. 1997; Basrai et al. 1999; Ross-Macdonald et al. 1999). Two recent studies described the use of comparative sequence analysis between the genome of *Saccharomyces cerevisiae* and those of other hemiascomycetous yeasts (Blandin et al. 2000) or other *Saccharomyces* genomes (Cliften et al. 2001) to discover small nonannotated protein-coding and nonprotein-coding genes in chromosomal regions previously considered intergenic. In addition, transposon insertion was recently used to identify yeast genes that were previously overlooked (Kumar et al. 2002).

Here we describe a systematic in silico method to identify new small genes in *S. cerevisiae* that is an extension of the searches conducted by Blandin and Cliften (Blandin et al. 2000; Cliften et al. 2001) by using a more comprehensive database of fungal sequences. In addition we provide a comprehensive demonstration that the majority of the genes predicted are actually transcribed. Our findings from comprehensive database searches and experimental studies suggest that the number of coding genes in *S. cerevisiae* is substantially higher than currently believed.

[1]Corresponding author.
E-MAIL arturo.morales@genomecorp.com; FAX (781) 398-2476.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.232903.

RESULTS

Identification of New ORFs

While comparing whole fungal genomes, we (Zeng et al. 2001) as well as others (Kupfer et al. 1997; Tzung et al. 2001) made the observation that fungal translated ORFeomes are very diverse. For example, many *S. cerevisiae* proteins do not have homologs in *Candida albicans* (Tzung et al. 2001). However, comparison of *S. cerevisiae* predicted translated ORFs with a comprehensive fungal database that includes predicted protein sequences from *C. albicans*, *Schizosacharomyces pombe*, *Aspergillus nidulans* and *fumigatus*, *Cryptococcus neoformans*, *Fusarium sporotrichioides*, *Neurospora crassa*, and *Pneumocystis carinii* suggests that most budding yeast translated ORFs have homologs in one or more other fungal genomes (Zeng et al. 2001). We used this observation to identify novel protein-coding sequences in the budding yeast genome.

Our approach to identify candidate ORFs for new genes in the *S. cerevisiae* genome is outlined in Figure 1. Briefly, we

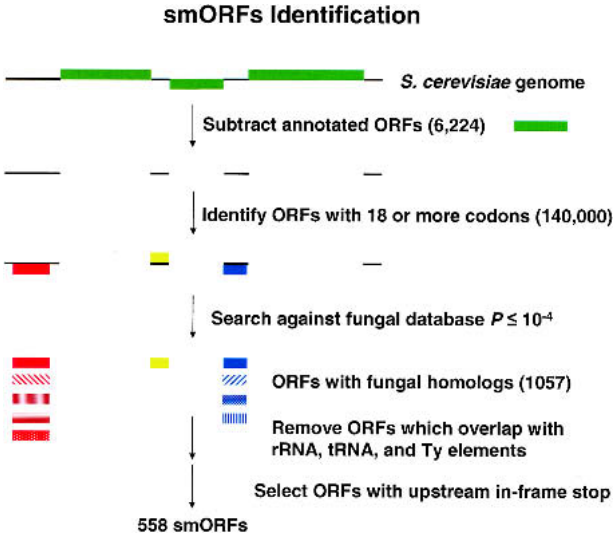

Figure 1 Overall strategy for smORF identification. Computational method used to identify new ORFs not identified by conventional methods.

13:264–271 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. List of *S. cerevisiae* smORFs for Which a Transcript was Detected

smorf	Original name	Chromososme	Coordinates	Size (aa)	CAI	RT-PCR product (bp)	SGD annotation	Notes
3	SR13	chr1	124491–124306	62	0.243	39		
13		chr2	18437–18177	87	0.181	99		
18[a]	smORF2	chr2	89973–90221	82	0.154	205	YBL071W-A	
19[c]	SR12	chr2	91789–92025	79	0.241	200	YBL069W	embedded ORF
24	SR11	chr2	209370–209609	79	0.102	142		
44	SR10	chr2	382978–382817	54	0.134	93		
46		chr2	391575–391309	89	0.091	72		
53	SR9	chr2	614131–613982	50	0.125	115		
54	SR8	chr2	622941–623105	55	0.082	113		
57[a,b]	SR7	chr2	684935–685219	94	0.105	170	YBR233W-A	
68		chr3	24325–24032	98	0.149	99	YCL057C-A	removed by S GD
82[a]	SR6	chr3	113722–114018	98	0.102	247	YCL001W-B	
93	SR5	chr3	316181–315990	63	0.121	161		
98	SR4	chr4	229171–229431	87	0.175	38		
101	SR3	chr4	410052–409804	83	0.151	68		
104		chr4	410052–409804	83	0.108	77		
109[a,b]		chr4	603805–603587	73	0.119	31	YDR079C-A	
118	SR2	chr4	794716–794567	50	0.164	124/98[e]		
121	SR1	chr4	829143–829346	67	0.160	155		
122		chr4	830646–830389	86	0.149	31		
123		chr4	848063–848215	51	0.067	64		
127		chr4	955124–955324	67	0.066	99		
139[a]	smORF8	chr4	1233506–1233267	79	0.161	227	YDR379C-A	
144		chr4	1283788–1284033	82	0.153	34		
154		chr9	96725–96522	78	0.135	67		
167		chr5	59549–59740	77	0.099	41		
171		chr5	117380–117183	66	0.216	87		
189		chr5	318642–318806	86	0.065	97		
201		chr6	48734–48925	64	0.173	63		
217		chr7	93078–93308	77	0.078	99		
226		chr7	418882–418700	61	0.097	40		
247		chr7	733407–733622	72	0.118	31		
250		chr7	974773–974573	67	0.091	92		
274		chr8	280232–280393	54	0.095	79		
283		chr8	453556–453705	50	0.084	75		
286		chr8	467058–466954	55	0.102	74		
288		chr8	466928–467110	61	0.079	62		
298		chr10	181408–181250	53	0.158	99		
301[a,b]		chr10	316419–316676	86	0.106	99	YJL062W-A	
303[a]		chr10	411124–410927	66	0.183	94	YJL012C-A	
313		chr10	637622–637861	80	0.081	99		
318		chr10	717273–717323	74	0.132	42		
324		chr11	98609–98397	71	0.186	25		
337		chr11	447674–447432	81	0.156	27		
352		chr12	136344–136520	59	0.197	85		
363		chr12	455434–455637	68	0.105	34		
382		chr12	455402–455566	55	0.078	99		
392		chr12	454697–455071	125	0.170	98		
398		chr12	454393–454557	55	0.059	99		
421		chr12	462672–462523	50	0.045	99		
439		chr12	452439–452624	62	0.086	99		
483		chr12	490407–490595	63	0.131	65		
494[d]	SR15	chr12	468959–468831 472611–472483 482191–482063 485843–485715	44	0.052	100		
499		chr12	582236–582006	77	0.140	37		
505		chr12	708338–708168	57	0.105	45		
509	SR17	chr12	760642–760355	96	0.102	190/82[e]		
511	SR16	chr12	815810–815983	58	0.105	125/61[e]		
514[a]	SR18	chr12	853460–853717	85	0.282	230	YLR363W-A	
519		chr12	950468–950262	69	0.135	39		
526		chr13	167781–167623	53	0.129	99		
530	SR25	chr13	337312–337602	97	0.082	160/58[e]		
532		chr13	550910–551134	75	0.121	94		
540		chr13	733267–733455	63	0.158	64		
543		chr13	804455–804691	79	0.101	80		
544		chr13	887925–887731	65	0.094	59		

Table 1. (*Continued*)

smorf	Original name	Chromososme	Coordinates	Size (aa)	CAI	RT-PCR product (bp)	SGD annotation	Notes
556[a]	SR22	chr14	330326–330544	73	0.130	170/37[e]	YNL162W-A	
561	SR21	chr14	381388–381242	48	0.102	90		
564		chr15	4130–4312	61	0.121	55		
570	smORF31	chr14	586816–386598	72	0.124	216		
577		chr15	4130–4312	61	0.153	75		
580		chr15	9179–9352	58	0.222	77		
590		chr15	355651–355857	78	0.135	29		
591		chr15	371684–371956	63	0.139	99		
598		chr15	461792–462040	59	0.054	99		
601		chr15	467391–467624	70	0.103	67		
625		chr15	907927–907718	70	0.085	97		
626		chr15	939340–939549	70	0.111	67		
631		chr15	1045189–1045344	52	0.117	78		
632		chr15	1058416–1058583	56	0.154	73		
640		chr16	75702–75989	96	0.158	99		
643		chr16	188512–188306	69	0.133	84		
655		chr16	480177–480368	64	0.151	30		
667		chr16	744172–744384	71	0.087	81		
672		chr16	883234–883452	73	0.129	95		

This table indicates the chromosomal location based on SGD nucleotide coordinates. smORFs marked with an ([a]) correspond to those identified by Blandin et al. (2000). smORFs marked with a ([b]) sign correspond to those identified by Cliften et al. (2001). smorf19 ([c]) was originally located downstream of YBL069W (*AST1*), which was recently extended (*Saccharomyces* Genome Database, 2001.http://www.genome-www.standford.edu/Saccharomyces). smorf494 ([d]) is found four times in chromosome 12, downstream of the *ASP3* gene in a 3.6-kb repeat (Johnston et al. 1997). For smORFs marked with ([e]), RT-PCR was carried with two sets of primers, one set shown in Fig. 2D and a second in Fig. 2E.

started with the *S. cerevisiae* genomic sequence (12.07 mb total) and removed the nucleotide sequences of the previously identified 6,224 coding ORFs (Saccharomyces Genome Database, December 5 1997, http://www.genome-www.stanford.edu/Saccharomyces). We then used the remaining sequences (3.45 mb) to identify all stop-to-stop ORFs that encode proteins with an arbitrary size of 18 amino acids or longer, based on the observation in *E. coli* that the majority of genes code for proteins with 18 or more amino acids (*E. coli* Genome Center, University of Wisconsin, Madison. http://www.genetics.wisc.edu/).

This approach resulted in approximately 140,000 predicted ORF products, most of them shorter than 100 residues. These ORF products were next searched against a comprehensive fungal protein sequence database to identify those with potential homologs. This fungal database consists of all NCBI entries listed under "fungi" (August 20, 2000, excluding any *S. cerevisiae* sequences), plus the genomic sequences from *C. albicans* (Stanford University) and *A. fumigatus* (PathoGenome; http://www.LabOnWeb.com), EST sequences from *A. nidulans*, *C. neoformans*, *F. sporotrichioides*, and *N. crassa* (University of Oklahoma Health Sciences Center), and *P. carinii* EST sequences (University of Georgia). Using a cutoff score of $P \leq 10^{-4}$ (this score was chosen because it is reasonably stringent for short translated ORFs), we found 1057 *S. cerevisiae* predicted ORF products with potential homologs in the fungal database. ORFs were considered similar when the region of sequence similarity between the small open reading frame (smORF) and the predicted protein(s) from our database extends over the entire coding region. We then removed the ORFs which were annotated after 1997 and while this work was in progress, those that overlap with rRNA, tRNA, and Ty elements, and selected the smORFs with a start-to-stop ORF and with an upstream in-frame stop. This approach resulted in 558 smORFs that code for predicted proteins with potential fungal homologs and are located in chromosomal sections previously identified as intergenic. The 558 smORFs range in length from 18–190 codons with a mean of 64.99 codons, a median of 64 codons, and a mode of 62 codons. A unique aspect of our search is that a comprehensive database of fungal genomic and cDNA sequences was used in the sequence similarity searches as opposed to the databases containing hemiascomycetous and *Saccharomyces* species used in the studies of Blandin, Cliften, and coworkers (Blandin et al. 2000; Cliften et al. 2001).

ORF Validation

We chose a subset of 117 smORFs for further characterization and validation (Table 1). As a first step we determined whether smORFs were expressed in yeast cells. Primers were designed to amplify smORFs 2, 8, and 31 as well as the *ACT1* gene (actin) as control (see Methods), and used for polymerase chain reaction (PCR) amplification with *S. cerevisiae* genomic DNA as a template to test the PCR amplification conditions. Products of the predicted size were obtained for all three smORFs as well as the actin control (Fig. 2A, lanes 2,6,10,14). No PCR products were obtained in reactions without the template (Fig. 2A, lanes 1,5,9,13), or using as a template RNA isolated from *S. cerevisiae* grown on rich (YPD) or complete synthetic minimal (CSM) media (Fig. 2A, lanes 3,4,7,8,11, 12,15,16). This indicates that these RNA samples were not contaminated with genomic DNA. We then tested for the presence of RNA transcripts originating from these smORFs as well as from the actin control, using RT-PCR. Products of the expected sizes were obtained for actin, as well as all three smORFs (Fig. 2B, lanes 2,3,5,6,8,9,11,12). This indicates that actin and the three smORFs are indeed expressed in yeast cells grown in rich and in minimal media. No RT-PCR product

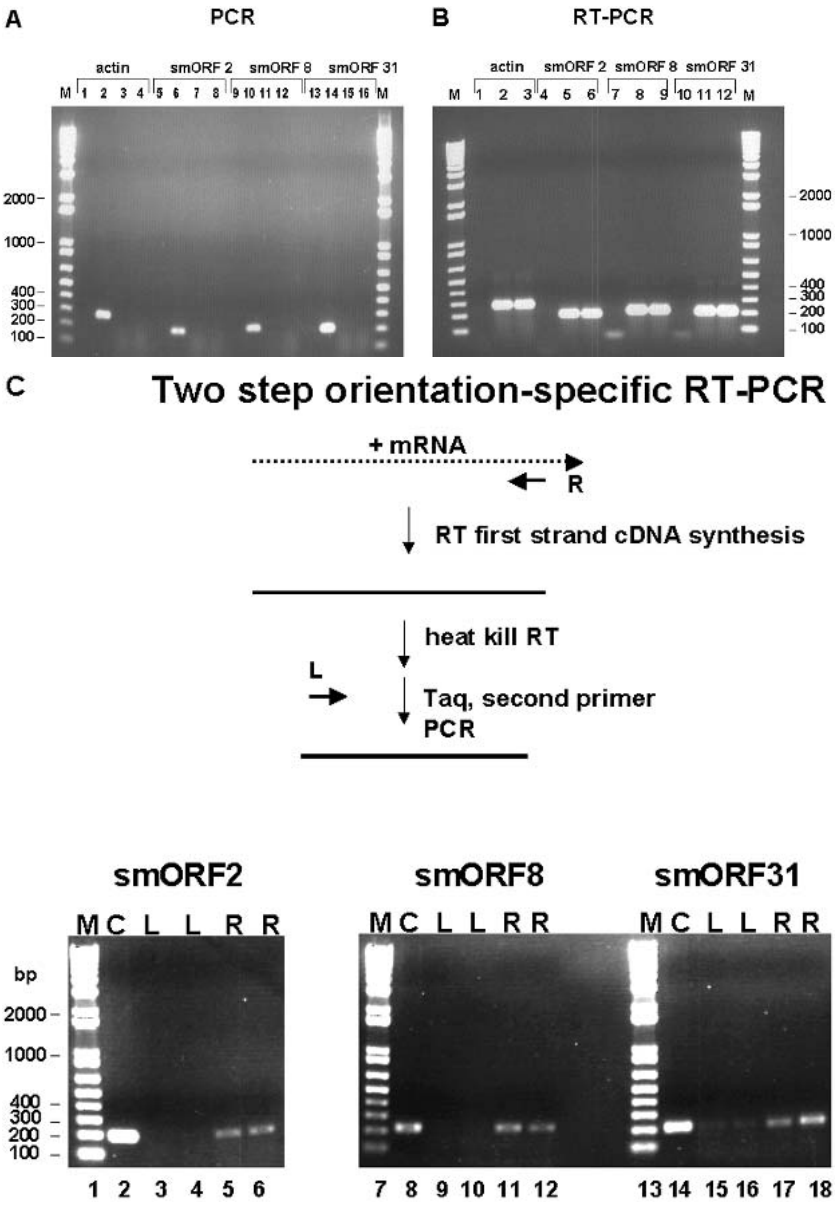

Figure 2 (Continued on next page)

was obtained in reactions without a template (Fig. 2B, lanes 1,4,7,10). The identity of the RT-PCR products was confirmed by cloning them, followed by restriction mapping and dideoxy sequencing (data not shown).

We then wanted to make sure that the smORF transcripts we identified were transcribed from the predicted DNA strand. To do this, we performed a variation of the RT-PCR experiment, first adding a primer complementary to the predicted mRNA and the reverse transcriptase. After first-strand cDNA synthesis, the reverse transcriptase was inactivated with heat and then Taq polymerase and both smORF-specific primers were added (Fig. 2C). Under these conditions we observed PCR products only when first-strand synthesis was conducted with primers complementary to the predicted mRNA (Fig. 2C, lanes 5,6,11,12,17,18). No PCR product was observed when first-strand synthesis was done with primers that have the same sequence as the mRNA (lanes 3,4,9,10,15,16). These results indicate that the transcripts observed for smORFs 2, 8, and 31 are made from the predicted strand.

This same study was extended to 114 additional smORFs. RT-PCR products of the expected size were obtained for 81 of these smORFs (e.g., Fig. 2D). Therefore, 84 of the 117 smORFs are transcribed from the predicted DNA strand (Table 1). The codon adaptation index (CAI, http://www.molbio.oc.ac.uk) was calculated for the 84 smORFs (Table 1), and they range from 0.045–0.282. Similar values were obtained for the 6224 annotated yeast genes. Nine of the smORFs in this list were recently annotated by Blandin, Cliften, and coworkers (Blandin et al. 2000; Cliften et al. 2001). We originally found SR12 downstream of the *AST1* gene, which was recently reannotated and extended, and now includes SR12 (Saccharomyces Genome Database, 2001; http://www.genome- www.stanford.edu/Saccharomyces). In addition we also identified smORF 68 and had an RT-PCR product for it. This smORF maps to YCL057C-A, which was recently removed from the *Saccharomyces* Genome Database (Saccharomyces Genome Database, 2001; http://www.genome-www.stanford.edu/Saccharomyces). The majority of the smORFs were found only once in the budding yeast genome, except for SR15, which is present four times in chromosome 12, downstream of *ASP3*, in a 3.6-kb repeat region of this chromosome (Johnston et al. 1997). Even though we detected an RT-PCR product for this smORF, we do not know which copy was transcribed. To address the possibility that the observed smORF transcripts were products of read-through transcription from genes located upstream of the smORFs, the RT-PCR experiment was conducted using a primer complementary to the mRNA for first-strand synthesis (Fig. 2C) and with a second primer located 400 base pairs (bp) upstream of the smORF. With these conditions, no RT-PCR products were observed for 25 smORFs tested, indicating that the smORF transcripts are not the result of read-through transcription from upstream genes (data not shown). To confirm these observations, we measured the size of the transcripts coded for by smORFs 2, 31, and the *ACT1* gene using Northern analysis (Fig. 2E). The size of the *ACT1* mRNA was 1300 nucleotides (nt; lane 3). The mRNAs for both smORF2 and 31 have a measured size of 460 nt (Fig. 2E, lanes 5,7), in agreement with a size predicted for a 250-nt ORF, a 20–30-nt 5′ untranslated region (Hughes et al. 2000), a 100-nt 3′ untranslated region (Graber et al. 2002), and a poly (A) tail of 70–90-nt (Hector et al. 2002). Northern analysis of smORF2 transcripts shows a band of 780 nt that could correspond to alternative processing variants. It is unlikely that this band represents read-through from the gene located upstream of smORF2 (*SNR56*), because it

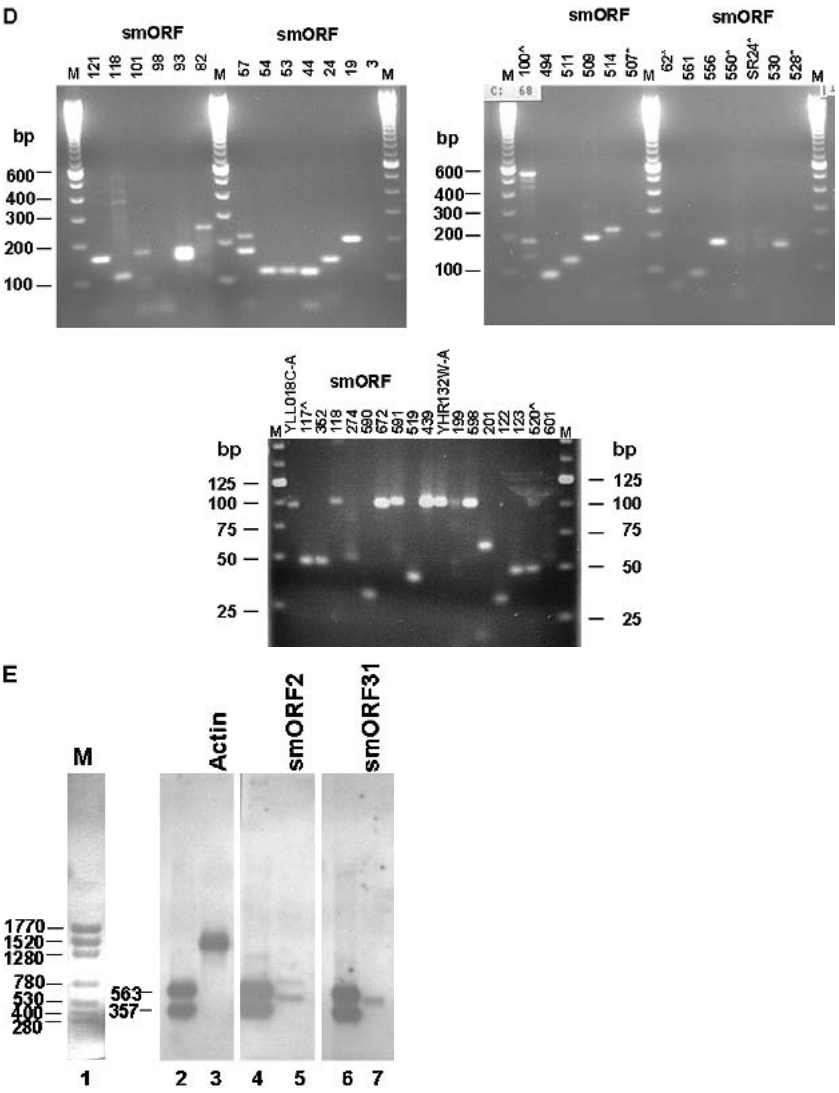

Figure 2 Experimental validation of the *S. cerevisiae* smORFs. (*A*) Primers specific for the yeast *ACT1* gene as well as the three smORFs were used for PCR amplification using no template (lanes *1*,*5*,*9*,*13*), 50 ng genomic DNA (lanes *2*,*6*,*10*,*14*), 500 ng total RNA from cells grown in rich media (lanes *3*,*7*,*11*,*15*), and 500 ng total RNA from cells grown in minimal media (lanes *4*,*8*,*12*,*16*) .(*B*) Primers specific for the yeast *ACT1* gene as well as the three smORFs were used for RT-PCR amplification using no template (lanes *1*,*4*,*7*,*10*), 500 ng total RNA from cells grown in rich media (lanes *2*,*5*,*8*,*11*), and 500 ng total RNA from cells grown in minimal media (lanes *3*,*6*,*9*,*12*). PCR and RT-PCR products were fractionated on a 1% agarose gel with DNA size markers and visualized after ethidium bromide staining. Sizes of DNA fragments in bp are indicated. (*C*) Two-step orientation-specific RT-PCR. Primers whose sequence is complementary to the predicted mRNAs of smORF2, 8, and 31 were used for first-strand cDNA synthesis. After heat inactivation of the reverse transcriptase, PCR amplification was carried out with both smORF-specific primers (lanes *5*,*6*,*11*,*12*,*17*,*18*). As control, the experiment was repeated using primers with the same sequence as the mRNA for first-strand cDNA synthesis (lanes *3*,*4*,*9*,*10*,*15*,*16*). (*D*) Examples of RT-PCR results with various smORFs indicated on top. RT-PCR detection of transcripts from the annotated smORFs YLL018C-A and YHR132W-A are shown. smORFs for which RT-PCR reactions resulted in no products are indicated (*) as are those for which the product is not of the expected size (^). These smORFs were not included in Table 1. Sizes of DNA fragments in bp are indicated. (*E*) Transcript size determination by Northern analysis of the ACT1 (lane *3*), smORF2 (lane *5*), and smORF31 (lane *7*). Unlabeled RNA markers stained with methylene blue (lane *1*) and labeled RNA markers (lanes *2*,*4*,*6*) were fractionated together with yeast poly (A)+ RNA. Sizes are shown in nucleotides.

is located 1700 bp upstream. This results indicate that the observed transcripts originate in the promoter of the smORFs.

Characterization of smORF2

We now present a comprehensive analysis of smORF2, as homologs of this protein (smORF2p) are found in many organisms from yeast to humans (Fig. 3), and its deletion from *S. cerevisiae* exhibits a tractable phenotype. The human, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *S. pombe* smORF2p homologs are about the same size as the *S. cerevisiae* counterpart. smORF2 was recently annotated by Blandin and coworkers (2000) with the systematic name YBL071W-A.

We extended our study of smORF2p to determine whether we could detect a protein product of the appropriate size. A triple HA tag was fused to the C-terminus of smORF2 by PCR, and the wild-type smORF2 gene was replaced with the tagged version by allele replacement into the chromosome (Ederniz et al. 1997). PCR amplification of the smORF2 $(HA)_3$ gene from genomic DNA, followed by cloning and sequencing, confirmed the identity of the tagged smORF2. Soluble extracts were then prepared and fractionated in 18% polyacrylamide gels containing sodium dodecyl sulfate. The proteins were transferred to a poly (vinylidene) fluoride (PVDF) membrane, and the blot was probed with anti-HA antibodies. The results show a protein band corresponding to a 9-kD protein (Fig. 4, lanes 3,4) in extracts prepared from cells with a tagged smORF2 gene and not in wild-type cells. This result shows that smORF2 is not only transcribed, but also encodes a detectable protein product of the expected size.

We then extended our study of smORF2 to determine whether this gene is essential or whether its deletion results in an observable phenotype. The complete smORF2 gene was deleted in a diploid yeast strain, using homologous recombination. Sporulation and tetrad analysis showed that haploid strains with a *smorf2*Δ were able to grow at 30°C (slow growth), but not at 37°C (Fig. 5). To extend the notion of smORF to humans, we next tested whether the human smORF2 is a functional homolog of the yeast smORF2. The human smORF2 gene, obtained from an EST clone, and the yeast smORF2 were cloned into the pYES vector for expression in yeast under the *GAL1* promoter. Clones were verified by sequencing and transformed into the *smorf2*Δ strain. The resultant transformants were tested for the ability to complement the temperature-sensitive phenotype of the *smorf2*Δ strain and their ability to form colonies at the restrictive temperature. The results show that the

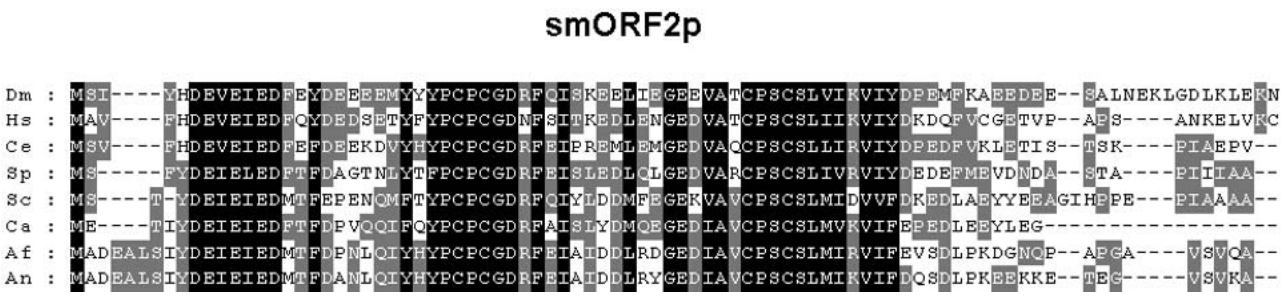

Figure 3 Multiple sequence alignments for smORF2p. smORF2p has highly conserved homologs in other fungi and in mammalian species. Abbreviations: Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Ce, *Caenorhabditis elegans*; Sc, *Saccharomyces cerevisiae*; Ca, *Candida albicans*; Af, *Aspergillus fumigatus*; An, *Aspergillus nidulans*; Sp, *Schizosaccharomyces pombe*; Bt, *Bos taurus*; Mm, *Mus musculus*. Residues that are identical or similar in all protein homologs are shaded in black, and those identical or similar in two or more, but not all, proteins in the alignment are shaded in gray. Homology shading was done with GeneDoc (Nicholas et al. 1997).

cloned human smORF2 as well as the yeast smORF2 can complement the temperature-sensitive phenotype of the *smorf2*Δ strain (Fig. 5). These results indicate that the human smORF2 is a functional ortholog of the yeast smORF2. Interestingly, the human smORF2 maps to two loci in the human genome, one in chromosome 3 where the gene contains two introns and codes for a predicted mRNA identical to the EST, and to a locus in chromosome 20 without introns but with nine predicted amino acid substitutions. These data indicate that small ORFs are present and expressed in humans, and they underscore the importance of looking for small genes in the genomes of higher eukaryotes.

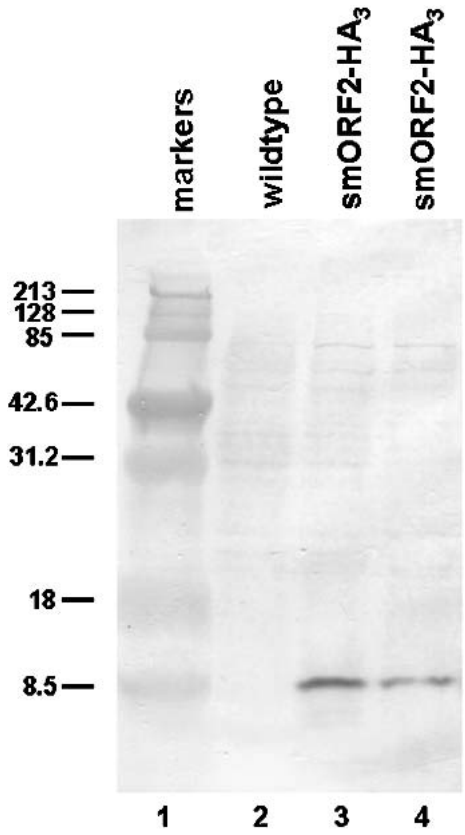

Figure 4 smORF2 is expressed in yeast. A triple HA tag was fused to the C-terminal end of smORF2 using PCR, and the wild-type smORF2 gene was replaced by the tagged smORF2 gene by allele replacement into the chromosome. Soluble extracts were prepared and analyzed in a Western blot probed with monoclonal antibodies that recognize the HA epitope. Extracts from wild-type cells (lane *2*) and extracts from two separate isolates carrying the HA-tagged smORF2 (lanes *3*,*4*).

DISCUSSION

We have validated a method for gene identification in sequenced genomes and used it to identify new genes in *S. cerevisiae*. With this method one should be able to find new coding ORFs in *S. cerevisiae* by simply searching potential budding yeast ORF products against sequences from other fungal and nonfungal species. Even though we did not verify the expression of every predicted smORF, we found strong evidence for close to 100 new genes in the *S. cerevisiae* genome. The limited study reported here can be expanded to include smORFs that partially overlap with annotated ORFs and smORFs that are completely located within previously annotated ORFs as described recently by Kumar et al. (2002). This systematic genome comparison approach to identify ORFs will accelerate and refine genome annotation and gene identification and will impact future experimental design. The identification of conserved protein products across a wide range of species can provide us with the opportunity to use *S. cerevisiae* and other fungi to study the function of their counterparts in humans. In addition, our approach can be applied to other sequenced genomes including human in order to identify coding ORFs not readily detected by conventional methods. We anticipate finding additional smORFs which do not yet have a homolog as the amount of available sequence data increases. Conversely, we will miss some smORFs that are species-specific and therefore have no homologs in other species. We refer to these species-specific smORFs as orphan smORFs. This study and others involved in gene discovery will change the landscape of genome annotation and therefore the approach in experimental design.

METHODS

RT-PCR Analysis

Genomic DNA was prepared from strain W303 (Thomas and Rothstein 1989) using the YeaStar Genomic DNA kit (Zymo Research). Primers pairs were chosen to amplify 250–300-bp regions within the coding ORFs of the yeast *ACT1* gene (5′-TGTCACCAACTGGGACGATA-3′; 5′-AACCAGCGTAAATTGGAACG-3′), smORF2 (5′-TGACGAAATCGAAATCGAAG-3′; 5′-GATGCCTGCCTCTTCGTAGT-3′), smORF8 (5′-TG

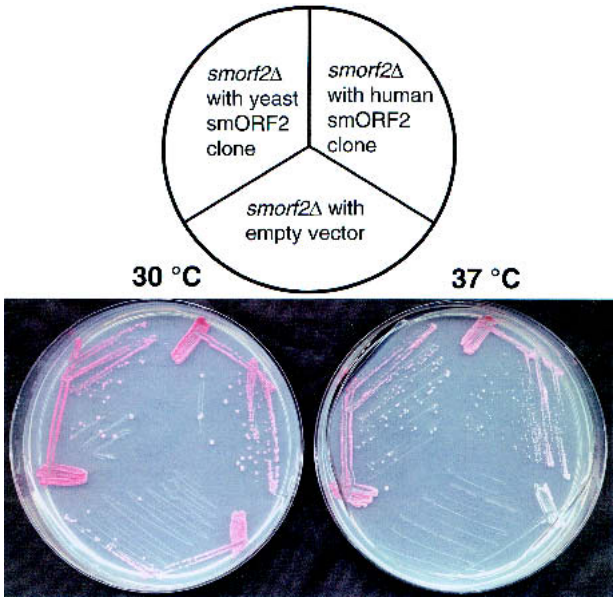

Figure 5 Human smORF2 complementation of the temperature-sensitive phenotype of the *smorf2*Δ strain. A yeast strain with a deleted smORF2 (*smorf2*Δ) was transformed with plasmids carrying the wild-type yeast smORF2, human smORF2 under the control of the *GAL1* promoter, or empty vector. Transformants were obtained at 30°C, and individual colonies were streaked and then incubated at 30°C and 37°C.

CCTAAGAGATTAAGTGGGTT-3′; 5′-CGTCAGTTCAGGGT GTGAAA-3′), smORF31 (5′-TGTCTGCATTATTTAATTTTC GTTC-3′; 5′-AGCTGTTAAATTGACTGATGGC-3′).

RNA was isolated from 5×10^7 yeast cells (strain W303) growing exponentially in YEPD or synthetic complete synthetic minimal media using the RNeasy Mini kit from QIAGEN including a DNase (Roche) digestion step. RT-PCR reactions were done with the OneStep RT-PCR Kit from QIAGEN as recommended by the manufacturer. RT-PCR products were fractionated on a 1% agarose gel and visualized after ethidium bromide staining. For sequencing, the RT-PCR products were isolated from an agarose gel and then cloned into pCR2.1-TOPO (Invitrogen). To test expression of the 117 smORFs, primers were chosen within the coding sequence to amplify fragments of 25 bp or longer.

Strand–Specific RT–PCR Analysis

First-strand ynthesis was conducted using yeast RNA, primers as indicated, and SuperScript II reverse transcriptase (Life Technologies) as recommended by the manufacturer. PCR amplification was conducted using PCR SuperMix (Life Technologies), smORF-specific primers, and 8.5% of the first-strand reaction.

For Northern analysis, 0.8 µg yeast poly (A)+ RNA together with DIG-labeled and unlabeled RNA markers (Invitrogen) were fractionated in 1.5% formaldehyde-agarose gel as recommended (Ambion), transferred to Nylon membrane, and probed with single-stranded antisense probes against *ACT1* (Fig. 2E, lanes 2,3), smORF2 (lanes 4,5), and smORF31 (lanes 6,7) labeled with DIG-UTP and detected as recommended (Roche).

Epitope Tagging

The modified smORF2 $(HA)_3$ was constructed by two-way PCR. First, a PCR amplification was made using a primer corresponding to 400 bp upstream of smORF2 (5′-AGAAAGCCCTCAAGCTTTCCCAGCG) and a second primer containing the C-terminus of smORF2 fused to the HA tag (5′-GGAGCCTGATCCAGCGTAGTCTGGGACGTCG TATGGGTAGCCAGCGTAGTCTGGGACGTCGT ATGGGTAGCCAGCGTAATCCGGAACATCATAC GGGTATCCTACGGCAGCAGCGGCAATAGGCTCAGG-3′). A second amplification was carried out with a forward primer containing the tag (5′-GTAGGATACCCGTATGATGTTCCG GATTACGCTGGCTACCCATACGACGTCCCAG ACTACGCTGGCTACCCATACGACGTCCCAGA CTACGCTGGATCAGGCTCCTAAAGATGAGAG GCTAGATCGAG-3′) and a primer located downstream of smORF2 (5′-TGTCGCTTTTTCTCCTCGATGAAGC CAAGCGCCGAACCAATTGATATCATCGGCACG-3′). The tagged smORF2 gene was introduced into the smORF2 locus by allele replacement (Ederniz et al. 1997). Allele replacement was first checked by PCR and then verified by PCR amplification of the tagged gene, cloning into pCR2.1-TOPO (Invitrogen) and sequencing.

S100 extracts were prepared from diploid W303 yeast cells grown in 25 mL of rich medium (YPD) to mid-log phase as described (Brown et al. 1996). Immunoblots were probed with a 1:1000 dilution of the 16B12 monoclonal antibody (Berkeley Antibody) against HA_3-tagged proteins.

Gene Disruption

smORF2 was disrupted in the diploid W303. Cells were transformed with a PCR fragment containing the *HIS3* marker flanked by 400 bp of smORF2 sequences. The *HIS3* sequences replaced amino acids 1 to 82 of smORF2. Histidine prototrophs were selected, and PCR was used to verify correct genomic integration. Sporulation and tetrad analysis were as described (Guthrie and Fink 1991). The human smORF2 coding sequence was amplified from I.M.A.G.E. clone 1047404 (Research Genetics). The yeast smORF2 was amplified from genomic DNA. PCR fragments were cloned into pYES2.1/V5-His-TOPO (Invitrogen) and transformed into yeast as described (Guthrie and Fink 1991).

ACKNOWLEDGMENTS

We thank Kim Fechtel for valuable suggestions and for critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

NOTE ADDED IN PROOF

While this paper was in being reviewed, Fichtner and Schaffrath reported the isolation of the *KTI11* gene by complementation of zymocin-resistant yeast mutants. Kti11p is smORF2. (Fichtner, L. and Schaffrath R. 2002. KTI11 and KTI13, Saccharomyces cerevisiae genes controlling sensitivity to G1 arrest induced by Kluyveromyces lactis zymocin. *Mol. Microbiol.* **44:** 865–875.)

REFERENCES

Andrade, M.A., Daruvar, A., Casari, G., Schneider, R., Termier, M., and Sander, C. 1997. Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. *Yeast* **13:** 1363–1374.

Basrai, M.A., Velculescu, V.E., Kinzler, K.W., and Hieter, P. 1999. *NORF5/HUG1* is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19:** 7041–7049.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett.* **487:** 31–36.

Brown, C.E., Tarun Jr., S.Z., Boeck, R., and Sachs, A. 1996. PAN3 encodes a subunit of the Pab1p-dependent poly(A) nuclease in *Saccharomyces cerevisiae*. *Mol. Cell. Biol* **16:** 5744–5753.
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.
Erdeniz, N., Mortensen, U.H., and Rothstein, R. 1997. Cloning-free PCR-based allele replacement methods. *Genome Res*. **7:** 1174–1183.
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:** 546–567.
Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A., Aytekin-Kurban, G., Bekiranov, S., Fajardo, J.E., Eswar, N., Sanchez, R., Sali, A., et al. 2001. Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. *Nat. Genet*. **27:** 337–340.
Graber, J.H., McAllister, G.D., and Smith, T.F. 2002. Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3′-processing sites. *Nucleic Acids Res*. **30:** 1851–1858.
Guthrie, C., Fink, G.R., Simon, M.I., and Abelson, J.N., Eds. 1991. *Guide to yeast genetics and molecular biology*. *Methods Enzymol*. Vol. 194. Academic Press, New York, NY.
Hector, R.E., Nykamp, K.R., Dheur, S., Anderson, J.T., Non, P.J., Urbinati, C.R., Wilson, S.M., Minvielle-Sebastia, L., and Swanson, M.S. 2002. Dual requirement for yeast hnRNP Nab2p in mRNA poly(A) tail length control and nuclear export. *EMBO J*. **21:** 1800–1810.
Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol*. **296:** 1205–1214.
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.
Kumar, A., Harrison, P.M., Cheung, K.H., Lan, N., Echols, N., Bertone, P., Miller, P., Gerstein, M.B., and Snyder, M. 2002. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol*. **20:** 58–63.
Kupfer, D.M., Reece, C.A., Clifton, S.W., Roe, B.A., and Prade, R.A. 1997. Multicellular ascomycetous fungal genomes contain more than 8000 genes. *Fungal Genet. Biol*. **21:** 364–372.
Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Laszkiewicz, A., Dudek, M.R., and Cebrat, S. 2002. How many protein-coding genes are there in the Saccharomyces cerevisiae genome? *Yeast* **19:** 619–629.
Mewes, H.W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., et al. 1997. Overview of the yeast genome. *Nature* **387:** 7–65.
Nicholas, K.B., Nicholas Jr., H.B., and Deerfield II, D.W. 1997. GeneDoc: Analysis and visualization of genetic variation. *EMBnet News* **4:** 14–17. http://www.no.embnet.org/embnet.news/vol4_2/contents.html
Olivas, W.M., Muhlrad, D., and Parker, R. 1997. Analysis of the yeast genome: Identification of new noncoding and small ORF-containing RNAs. *Nucleic Acids Res*. **25:** 4619–4625.
Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-I, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., et al. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet*. **27:** 332–336.
Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L., et al. 1999. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402:** 413–418.
Thomas, B.J. and Rothstein, R. 1989. The genetic control of direct-repeat recombination in *Saccharomyces*: The effect of rad52 and rad1 on mitotic recombination at *GAL10*, a transcriptionally regulated gene. *Genetics* **123:** 725–738.
Tzung, K.W., Williams, R.M., Scherer, S., Federspiel, N., Jones, T., Hansen, N., Bivolarevic, V., Huizar, L., Komp, C., Surzycki,. R., et al. 2001. Genomic evidence for a complete sexual cycle in *Candida albicans*. *Proc. Natl. Acad. Sci*. **98:** 3249–3253.
Velculescu, V.E., Zhang, L., Zhou,. W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E. Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Winzeler, E.A. and Davis, R.W. 1997. Functional analysis of the yeast genome. *Curr. Opin. Genet. Dev*. **7:** 771–776.
Zeng, Q., Morales, A., and Cottarel, G. 2001. Fungi and humans: Closer than you think. *Trends Genet*. **17:** 682–684.

WEB SITE REFERENCES

http://www.LabOnWeb.com; *Aspergillus fumigatus* genomic sequences are available on the Web.
http://www.genetics.wisc.edu/; *Escherichia coli* Genome Center. October 13, 1998, revision date: *E. coli* Genome Center, University of Wisconsin, Madison.
http://genome-www.stanford.edu/Saccharomyces/; Saccharomyces Genome Database as of December 5 1997.
http://genome-www.stanford.edu/Saccharomyces/; Saccharomyces Genome Database as of October 2001.
http://www.molbio.oc.ac.uk; Codon adaptation index (CAI).

Received March 1, 2002; accepted in revised form November 7, 2002.