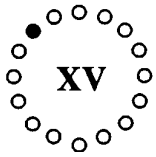
YEAST VOL. **13:** 583–589 (1997)



Yeast Sequencing Reports

Sequence Analysis of a 33·2 kb Segment from the Left Arm of Yeast Chromosome XV Reveals Eight Known Genes and Ten New Open Reading Frames Including Homologues of ABC Transporters, Inositol Phosphatases and Human Expressed Sequence Tags

MARIA TZERMIA†, CHRISTINA KATSOULOU† AND DESPINA ALEXANDRAKI*†‡

†*Foundation for Research and Technology-HELLAS, Institute of Molecular Biology and Biotechnology, Crete, Greece*
‡*Department of Biology, University of Crete, Crete, Greece*

Received 15 August 1996; accepted 25 October 1996

The complete nucleotide sequence of a 33 221 bp segment, contained in cosmid pEOA1044, derived from the left arm of chromosome XV of *Saccharomyces cerevisiae*, appears in public databases between coordinates 177013 and 210234 (http://speedy.mips.biochem.mpg.de/). Computer analysis of that sequence revealed the presence of the previously known genes *IRA2*, *DEC1*, *NUF2*, *HST1*, *RTG1*, *RIB2* and *HAL2*, one previously partially sequenced open reading frame (ORF) of unknown function (SCORFAC) and ten newly identified ORFs. One of the new ORFs is similar to the *Drosophila melanogaster white* gene and other transmembrane ABC transporters, another one has similarities to inositol phosphatases and others are similar to ORFs of unknown function from various organisms, including human Expressed Sequence Tags (ESTs). Potential transmembrane regions, ATP/GTP-binding and WD motifs have also been identified. The existence of yeast ESTs for two of the newly identified ORFs indicates that they are transcribed. © 1997 by John Wiley & Sons, Ltd.

Yeast 13: 583–589, 1997.
No. of Figures: 1. No. of Tables: 1. No. of References: 21.

KEY WORDS — genome sequencing; *Saccharomyces cerevisiae*; chromosome XV; *IRA2*; *DEC1*; *NUF2*; *HST1*; *RTG1*; *RIB2*; *HAL2*; SCORFAC; *white* gene; ABC transporters; inositol phoshatases; membrane proteins; ATP/GTP-binding; human ESTs; WD repeats

INTRODUCTION

We have determined 33 221 bp from pEOA1044, one of the overlapping cosmids containing segments of chromosome XV, constructed by Thierry *et al.* (in preparation). This project was coordinated by Bernard Dujon in the course of the European Union BIOTECH programme of *Saccharomyces cerevisiae* DNA sequencing. The reported DNA segment is derived from the left arm of the chromosome and includes the previously sequenced and experimentally characterized genes *IRA2*, *DEC1*, *NUF2*, *HST1*, *RTG1*, *RIB2* and *HAL2*. Of these, only the *IRA2* gene is included in the published genetic map of

*Correspondence to: Despina Alexandraki, Foundation for Research and Technology-HELLAS, Institute of Molecular Biology and Biotechnology, PO Box 1527, Heraklion 711 10, Crete, Greece.
Contract grant sponsor: Commission of the European Communities under the BIOTECH programme of the Division of Biotechnology
Contract grant sponsor: Greek Ministry of Industry, Energy and Technology

CCC 0749-503X/97/060583–07 $17.50

chromosome XV (Mortimer *et al*., 1989). We have also identified one previously partially sequenced open reading frame (ORF) of unknown function (SCORFAC) and ten new ORFs.

MATERIALS AND METHODS

Strains and vectors

Cosmid pEOA1044 containing partial *Sau*3AI yeast DNA fragments in pWE15 vector was provided from the collection of B. Dujon's group. *Escherichia coli* strain DH5α (*supE*44 Δ*lac* U169 (φ80*lacZ*ΔM15) *hsdR*17 *recA*1 *endA*1 *gyrA*96 *thi*-1 *relA*1) and the vectors pUC19, pUC18, pBluescript and YEp352 were used for all subsequent subcloning and sequencing steps. Vectors were chosen appropriately to facilitate the subsequent generation of ExoIII deletions.

Sequencing strategy—sequence analysis software

We have used directed sequencing of ordered subcloned *Eco*RI restriction fragments and polymerase chain reaction cosmid sequencing to determine the junctions between the different *Eco*RI fragments following the methodologies described in Katsoulou *et al*. (1996). The sequencing reactions were analysed on an ALF sequencer (Pharmacia). Synthetic oligonucleotides were purchased from the Microchemistry group of IMBB in Crete and from MWG-BIOTECH GmbH in Ebersberg.

Sequence assembly of each restriction fragment was performed by the 'Fragment Assembly Program' of the GCG sequence analysis software package for VMS. Total sequence assembly and verification was performed as described in Tzermia *et al*. (1994). Restriction and six-phase ORF mapping of the sequences were accomplished by the DNA Strider software (Marck, 1988). Comparisons of the nucleotide and the amino acid sequences were performed to the GenBank, EMBL, SwissProt and NBRF libraries using the GCG package software by us at the IMBB MicroVAX or at appropriate Internet sites (MIPS, EBI, NCBI, Stanford), and by Karl Kleine at MIPS (Martinsried, Germany).

RESULTS AND DISCUSSION

Sequence analysis

The complete nucleotide sequence of a 33 221 bp segment appears in public databases between coordinates 177013 to 210234 of chromosome XV. The six-phase ORF map of this region, performed by the DNA Strider program, revealed 18 ORFs >100 codons. Table 1 includes the cosmid coordinates of all identified coding sequences. Two ORFs that are internal (in opposite orientation) to *IRA2* and *NUF2* coding regions were not included in our analysis. The ORF sizes range from 132 to 1294 codons. Some discrepancies were noticed compared with all previously published sequences, except with the *RTG1* gene, mostly at the untranslated gene regions. We have identified Expressed Sequence Tags (ESTs) in databases (Boguski, 1995) corresponding to some of the known ORFs and to two of the newly identified ORFs, providing proof of their transcription.

In the intergenic regions of the reported sequence we have also identified one ARS consensus sequence (5′WTTTAYRTTTW 3′) and four stretches of 10, 12, 18 and 23 contiguous As (Ts) in addition to several stretches of 7–9 As (Ts).

ORF analysis

Several characteristics of the identified ORFs are provided in Table 1. The optimum and self scores of FastA analysis (Pearson and Lipman, 1988) are given for identical or similar ORF sequences. Scores higher than 200 have been considered as significant, although in some instances lower scores due to similarities in restricted areas of the protein sequences indicated conservation of specific domains. Only the best homology score for each ORF was included in Table 1. The codon adaptation index (CAI; Sharp and Li, 1987; Sharp and Cowe, 1991; determined by Karl Kleine at MIPS) is shown as a parameter indicating the probability for the corresponding ORF to be expressed. ORFs with CAI <0·110 are considered questionable. Potential transmembrane regions have been determined according to Klein *et al*. (1985; by Karl Kleine at MIPS) and visualized by the hydrophobicity profiles (Kyte and Doolittle, 1982) presented by the DNA Strider program. Protein patterns (motifs) have been identified by the ProSite program (Bairoch, 1991) of the GCG package. Additional brief discussion of our major findings is provided below for each ORF individually.

ORF YOL081w corresponds to the previously known GTPase, ras-activating Ira2p protein (Tanaka *et al*., 1990).

ORF YKL080c exhibited similarities to the XPMC2 protein of *Xenopus laevis*, which is involved in mitotic phenomena when expressed in *Schizosaccharomyces pombe* (Su and Maller, 1995). Fewer similarities but at the same regions were also found with the mushroom-inducing putative nucleotide-binding fungal Frt1 protein (S55252). At the same regions, similarities were found to rodent EST sequences (H33693, W10771) and to three yeast hypothetical proteins, YGR276c, YLR107w (which share homologies with the primate GOR protein-D10017) and YGL094c (*PAN2* gene component of polyA ribonuclease; FastA scores 219, 172 and 158 respectively).

ORF YOL079w, which overlaps in the opposite orientation with ORF YKL080c, is probably not transcribed as indicated by its low CAI value.

ORF YOL078w was found to be longer than the previously sequenced hypothetical ORF (SCORFAC). No homologies have yet been found in databases for this ORF.

ORF YOL077c showed significant similarities with two ORFs of unknown function, one from *Sz. pombe* and one from *Caenorhabditis elegans* (L14331; K12H4.3 protein, 352 amino acids; 40·4% identity in 277 amino acids). In addition, corresponding EST sequences exist from *S. cerevisiae*, as well as similar ESTs from mammalian sources (W00649), including the yy68c11. r1 Soares multiple sclerosis 2NbHMSP *Homo sapiens* cDNA (W1189; 53% identity in 137 amino acids).

The sequence of the YOL076w ORF was identical to the Dec1p (Mdm20p) protein which genetically interacts with the Cin8p.

ORF YOL075c belongs to the family of ABC transporters (Higgins, 1992). The highest FastA scores were found by comparison to the *white* gene product (membrane-associated, ATP-binding protein, involved in transport of pigment precursors in ommochrome and pteridine pathways) from the fruit fly (O'Hare *et al.*, 1984), mouse (Z48745; FastA score 569, 24·9% identity in 575 amino acids) and human (X91249; FastA score 518, 25·5% identity in 517 amino acids). ORF YOL075c contains eight transmembrane regions, two ATP-binding motifs and homologies to membrane transporters from different prokaryotes and eukaryotes including other sequences from yeast. In fact, best overall similarities were found with molecules of similar lengths, visualized also by comparison of the corresponding hydrophobicity profiles (Figure 1), such as the *Candida albicans* Cdr1p protein (Prasad *et al.*, 1995; X77589) conferring multiple resistance to drugs and antifungals (1470 amino acids; 21·5% identity in 1069 amino acids) and the yeast sequences YNR070w (1333 amino acids; 21·9% identity in 474 amino acids) and its homologous drug-resistance transporter Snq2p (1501 amino acids; Servos *et al.*, 1993; 19·1% identity in 572 amino acids). With all of these sequences, small regions of identical residues were found mainly flanking their ABC transporter motif (not shown). Other examples of yeast sequences with similarities to YOL075c ORF are the putative ATP-dependent permeases YCR011c (ADP1; 1049 amino acids; 40·4% identity in 188 amino acids) and YOR011w (1394 amino acids; 22·6% identity in 274 amino acids), as well as Pdr5p (27·34% identity in 267 amino acids). In addition to the existence of a corresponding EST sequence deriving from the yeast gene *YOL075c*, many EST sequences of human, mouse, rice, *C. elegans* and *A. thaliana* origins are translated to ORFs that are significantly similar to YOL075c, indicating a family of highly expressed genes in most species.

The sequences of ORFs YOL073c, YOL072w and YOL070c showed no significant similarities to any known protein sequences or ESTs. ORF YKL073c is probably not transcribed as indicated by its low CAI value.

ORF YOL071w showed significant similarities to an *Sz. pombe* ORF and several human EST sequences of unknown function [W68780, D16894 (24·6% identity in 138 amino acids overlap), N50503, T33308, etc.].

ORF YOL069w corresponds to the known spindle pole body protein Nuf2p (Osborne *et al.*, 1994).

The last 10 kb of the reported sequence contain six ORFS (four of which are known) transcribed at the same orientation. ORF YOL068c corresponds to the Hst1p protein, a member of the *SIR2* gene family. ORF YOL067c corresponds to the Rtg1p protein, a basic helix-loop-helix transcription factor involved in the communication between mitochondria, peroxisomes and nucleus (Liao and Butow, 1993). ORF YOL066c corresponds to the known Rib2p, DRAP deaminase, part of the riboflavin biosynthesis pathway. ORF YOL064c is the known Hal2p (Met22p), a serine/threonine phosphatase, involved in salt tolerance and methionine biosynthesis (Glaeser *et al.*, 1993).

ORF YOL065c, which contains two presumptive transmembrane regions, showed best similarities to two inositol phosphatase

Table 1. Characteristics of open reading frames (ORFs) identified in the 33 221 bp segment of chromosome XV.

ORF name	ORF location (bases)	ORF length (aa)	CAI	Identities, homologies, motifs*	Opt. score	Self score
YOL081w (AIB1097)†	2–3292	1097	0·14	*S. c. IRA2* (3079 aa) M33779, A35656 99·8% identity in 1097 aa	5525	14930
(AID102)	392–697			Internal (opposite) of YOL081w		
YOL080c (AIE289)	3547–4413	289	0·16	*X. laevis* XPMC2 protein (421 aa) U10185 34·5% identity in 278 aa	448	1995
YOL079w (AIB132)	4043–4438	132	0·08	*A. mellifera* mitochondrion NADH dehydrogenase 4 (*SGC4*; 447 aa)	148	2579
				L06178, S52968 25·0% identity in 140 aa		
				Transmembrane (aa 1–17)		
				Transmembrane (aa 23–39)		
				Transmembrane (aa 52–68)		
				Transmembrane (aa 79–95)		
				Overlapping (opposite) of YOL080c		
YOL078w (AIC1176)	4668–8195	1176	0·12	*S. c.* hypothetical (XV; 380 aa) M88606, S27437 97·1% identity in 379 aa	1685	1740
ARS (consensus)	6725–6735					
YOL077c (AID291)	8837–9709	291	0·23	*Sz. pombe* cosmid 800 (295 aa) U41410 55·6% identity in 257 aa	810	1417
				Transmembrane (aa 210–226)		
				S. c. EST102442 (T37337)		
YOL076w (AIB796)	10010–12397	796	0·16	*S. c. DEC1* (*MIM20*; 796 aa) U36382 100% identity in 796 aa	4019	4019
YOL075c	12647–16528	1294	0·13	*D. melanogaster white* (687 aa) X51749, FYFFW 27·5% identity in 600 aa	656	3432
				Transmembrane (aa 376–392)		
				Transmembrane (aa 469–485)		
				Transmembrane (aa 496–512)		
				Transmembrane (aa 606–622)		
				Transmembrane (aa 1042–1058)		
				Transmembrane (aa 1125–1142)		
				Transmembrane (aa 1177–1193)		
				Transmembrane (aa 1269–1285)		
				ATP/GTP-binding motif A (aa 62–69)		
				ATP/GTP-binding motif A (aa 627–634)		
				ABC transporters motif (aa 839–853)		
				S. c. EST104485 (T38919)		

YOL073c	(AID322)	16821–17786	322	0·08	Transmembrane (aa 17–33) Transmembrane (aa 200–216)			
YOL072w	(AIC455)	17956–19320	455	0·11				
YOL071w	(AIA162)	19493–19978	162	0·15	*Sz. pombe* SPAC12B10.06c (139 aa) Z70721 44·1% identity in 111 aa	266	738	
YOL070c	(AIF501)	20209–21711	501	0·11				
YOL069w	(AIC451)	21928–23280	451	0·14	*S. c. NUF2* (451 aa) A53910 100% identity in 451 aa	2129	2129	
	(AID100)	22821–23120			Internal (opposite) of YOL069w			
YOL068c	(AIE503)	23357–24865	503	0·13	*S. c. HST1* (503 aa) L47120, S59698 100% identity in 503 aa	2539	2539	
YOL067c	(AE1177)	24974–25504	177	0·12	*S. c. RTG1* (177 aa) M97690, A44344 100% identity in 177 aa	783	783	
YOL066c	(AIE591)	25685–27457	591	0·14	*S. c. RIB2* (591 aa) Z21618, S50972 100% identity in 591 aa	2958	2958	
YOL065c	(AID384)	27720–28871	384	0·11	YIL002c inositol phosphatase homologue (946 aa) 25·8% identity in 326 aa	223	4766	
					Transmembrane (aa 96–112)			
					Transmembrane (aa 367–383)			
YOL064c	(AIF357)	29092–30161	357	0·20	*S. c. HAL2, MET22* (357 aa) X72847, S35318 100% identity in 357 aa	1701	1701	
YOL063c	(AID947)	30381–33221	947	0·13	Transmembrane (aa 192–208) Transmembrane (aa 834–850) β-transducin family WD repeats (aa 253–267)			

*Motifs are described only for the newly identified ORFs.
†Nomenclature originally assigned by MIPS.

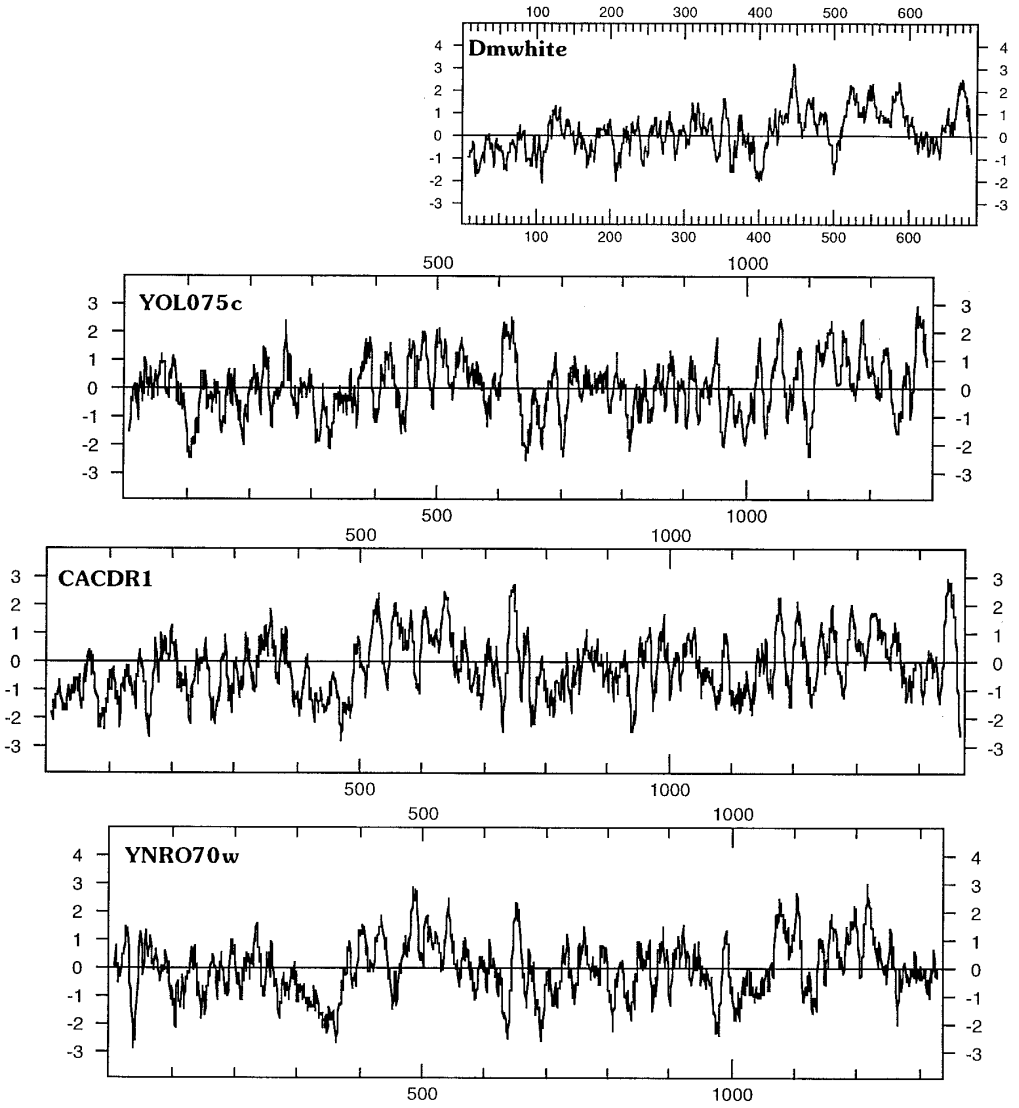


Figure 1. Hydrophobicity profiles of the *Drosophila melanogaster white* gene product (Dmwhite), the *Candida albicans CDR1* gene product (CACDR1) and the yeast YOL075c and YNR070w ORF sequences.

homologues from yeast: YIL002c, bearing one transmembrane region and YNL106c (1183 amino acids), bearing two transmembrane regions (27·2% identity in 283 amino acids overlap). All three ORFs exhibit significant regional similarities to the human protein OCRL (968 amino acids), highly homologous to the inositol polyphosphate-5-phosphatases, responsible for Lowe's oculo-cerebrorenal syndrome (M88162) and to the mouse SH2-containing inositol-5-phosphatase (1190 amino acids; U51742). It could be of some structural or functional significance that the Hal2p phosphatase expressed from the neighbouring *YOL064c* gene also contains an inositol monophosphatase family motif.

ORF YOL063c did not show any homologies to known proteins but contains a Trp-Asp repeat signature: MHQDF LLACGDNGIVYIWEI NKVIK (Duronio *et al.*, 1992).

ACKNOWLEDGEMENTS

We thank Bernard Dujon for the coordination of the chromosome XV sequencing project and Karl

Kleine and all MIPS staff for help with the sequence analysis. We are grateful to Hartmut Voss and Wilhelm Ansorge for hospitality at EMBL and continuous help on the ALF sequencing. We thank Georgia Houlaki for help with the preparation of the figures. This work was supported by the Commission of the European Communities under the BIOTECH programme of the Division of Biotechnology and by the Greek Ministry of Industry, Energy and Technology.

REFERENCES

Bairoch, A. (1991). A dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **16,** 2241–2245.

Boguski, M. S. (1995). The turning point in genome research. *Trends in Biochem. Sci.* **20,** 295–296.

Duronio, R. J., Gordon, J. I. and Boguski, M. S. (1992). Comparative analysis of the beta transducin family with identification of several new members including *PWP1*, a nonessential gene of *Saccharomyces cerevisiae* that is divergently transcribed from *NMT1*. *Proteins* **13,** 41–56.

Glaeser, H. U., Thomas, D., Gaxiola, R., Montrichard, F., Surdin-Kerjan, Y. and Serrano, R. (1993). Salt tolerance and methionine biosynthesis in *Saccharomyces cerevisiae* involve a putative phosphatase gene. *EMBO J.* **12,** 3105–3110.

Higgins, C. F. (1992). ABC transporters: from microorganisms to man. *Ann. Rev. Cell Biol.* **8,** 67–113.

Katsoulou, C., Tzermia, M., Tavernarakis, N. and Alexandraki, D. (1996). Sequence analysis of a 40·7 kb segment from the left arm of yeast chromosome X reveals 14 known genes and 13 new open reading frames including homologues of genes clustered on the right arm of chromosome XI. *Yeast* **12,** 787–797.

Klein, P., Kanehisa, M. and DeLisi, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815,** 468–476.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157,** 105–132.

Liao, X. and Butow, R. A. (1993). *RTG1* and *RTG2*: two yeast genes required for a novel path of communication from mitochondria to the nucleus. *Cell* **72,** 61–71.

Marck, C. (1988). 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucl. Acids Res.* **16,** 1829–1836.

Mortimer, R. K., Schild, D., Contopoulou, C. R. and Kans, J. A. (1989). Genetic map of *Saccharomyces cerevisiae*, Edition 10. *Yeast* **5,** 321–403.

O'Hare, K., Murphy, C., Levis, R. and Rubin, G. M. (1984). DNA sequence of the white locus of *Drosophila melanogaster*. *J. Mol. Biol.* **180,** 437–455.

Osborne, M. A., Schlenstedt, G., Jinks, T. and Silver, P. A. (1994). Nuf2, a spindle pole body-associated protein required for nuclear division in yeast. *J. Cell Biol.* **125,** 853–866.

Pearson, V. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.

Prasad, R., de Wergifosse, P., Goffeau, A. and Balzi, E. (1995). Molecular cloning and characterization of a novel gene of *Candida albicans, CDR1*, conferring multiple resistance to drugs and antifungals. *Curr. Genet.* **27,** 320–329.

Servos, J., Haase, E. and Brendel, M. (1993). Gene *SNQ2* of *Saccharomyces cerevisiae*, which confers resistance to 4-nitroquinoline-N-oxide and other chemicals, encodes a 169 kDa protein homologous to ATP-dependent permeases. *Mol. Gen. Genet.* **236,** 214–218.

Sharp, P. M. and Cowe, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7,** 657–678.

Sharp, P. M. and Li, W.-H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.* **15,** 1281–1295.

Su, J. Y. and Maller, J. L. (1995). Cloning and expression of a *Xenopus* gene that prevents mitotic catastrophe in fission yeast. *Mol. Gen. Genet.* **246,** 387–396.

Tanaka, K., Nakafuku, M., Tamanoi, F., Kaziro, Y., Matsumoto, K. and Toh-e, A. (1990). *IRA2*, a second gene of *Saccharomyces cerevisiae* that encodes a protein with a domain homologous to mammalian ras GTPase-activating protein. *Mol. Cell. Biol.* **10,** 4303–4313.

Tzermia, M., Horaitis, O. and Alexandraki, D. (1994). The complete sequencing of a 24·6 kb segment of yeast chromosome XI identified the known loci *URA1*, *SAC1* and *TRP3*, and revealed six new open reading frames including homologues to the threonine dehydratases, membrane transporters, hydantoinases and the phospholipase A_2-activating protein. *Yeast* **10,** 663–679.

