YEAST VOL. 12: 787–797 (1996)

Yeast Sequencing Reports

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

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

†*Foundation for Research and Technology-HELLAS, Institute of Molecular Biology and Biotechnology, PO Box 1527, Heraklion 711 10, Crete, Greece*
‡*Department of Biology, University of Crete, Greece*

Received 25 August 1995; accepted 4 December 1995

The complete nucleotide sequence of a 40·7 kb segment about 130 kb from the left end of chromosome X of *Saccharomyces cerevisiae* was determined from two overlapping cosmids. Computer analysis of that sequence revealed the presence of the previously known genes *VPS35*, *INO1*, *SnR128*, *SnR190*, *MP12*, *YAK1*, *RPB4*, *YUR1*, *TIF2*, *MRS3* and *URA2*, three previously sequenced open reading frames (ORFs) of unknown function 5′ of the *INO1*, 5′ of the *MP12* and 3′ of the *URA2* genes and 13 newly identified ORFs. One of the new ORFs is homologous to mammalian glycogenin glycosyltransferases and another has similarities to the human phospholipase D. Some others contain potential transmembrane regions or leucine zipper motifs. The existence of yeast expressed sequence tags for some of the newly identified ORFs indicates that they are transcribed. A cluster of six genes within 10 kb (*YUR1*, *TIF2*, two new ORFs, an *RSP25* homologue and *MRS3*) have homologues arranged similarly within 28·5 kb on the right arm of chromosome XI. The sequence has been deposited in the EMBL data library under Accession Number X87371.

KEY WORDS — genome sequencing; *Saccharomyces cerevisiae*; chromosome X; *VPS35*; *INO1*; *SnR128*; *SnR190*; *MP12*; *YAK1*; *RPB4*; *YUR1*; *TIF2*; *MRS3*; *URA2*; *RSP25* homologue; leucine zippers; membrane proteins; glycogenin glycosyltransferase; human phospholipase D; chromosome XI; gene cluster translocation/recombination

INTRODUCTION

In the course of the European Union BIOTECH project of *Saccharomyces cerevisiae* chromosome X DNA sequencing, we have determined 40,724 base pairs from two overlapping cosmids constructed by Huang *et al.* (1994). This DNA segment is derived from the left arm of the chromosome and includes the previously sequenced and experimentally characterized genes *VPS35*, *INO1*, *SnR128*, *SnR190*, *MP12* (Mim17p), *YAK1*, *RPB4*, *YUR1*, *TIF2* (eIF-4ap), *MRS3* and *URA2*. It also contains three previously sequenced open reading frames (ORFs) of unknown function and 13 newly identified ORFs. Our sequence analysis revealed special features of the identified

*Corresponding author. Mailing address: Foundation for Research and Technology-HELLAS, Institute of Molecular Biology and Biotechnology, PO Box 1527, Heraklion 711 10, Crete, Greece.

CCC 0749-503X/96/080787–11

ORFs and specific translocation/recombination/gene conversion events between a region of six unrelated clustered genes on chromosome X and a region of 11 similarly arranged unrelated genes on chromosome XI.

MATERIALS AND METHODS

Strains and vectors

Cosmids pEJ007 and pEJ031 containing partial *Sau*3AI yeast DNA fragments in pWE15 vector were provided from the collection of F. Galibert's group, described in Huang *et al.* (1994). *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. The vector was chosen appropriately to facilitate the subsequent generation of ExoIII deletions.

Sequencing strategy

We used directed sequencing of ordered subcloned *Eco*RI restriction fragments. Template DNA was prepared by the alkaline lysis method followed by Qiagen column selection. Dideoxynucleotide sequencing with T7 polymerase was performed on both strands of fragments subcloned in both orientations. 'Universal' or 'reverse' fluorescent primers (Pharmacia kit) were used for the original *Eco*RI fragments and for nested ExoIII-S1 nuclease deletions of those fragments. The sequencing reactions were analysed on an ALF sequencer (Pharmacia). Synthetic oligonucleotides (Microchemistry group, IMBB, Crete) corresponding to internal sequences were used to fill in the gaps, using [^{35}S]dATP and the USB sequencing kit. The junctions between the different *Eco*RI fragments have been established by polymerase chain reaction sequencing from [^{32}P]ATP-end-labelled oligonucleotide primers near the ends of the fragments using cosmid DNA as template (Ausubel *et al.*, 1987).

Sequence analysis software

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, Swiss-Prot and NBRF libraries using the GCG package software by us at the Institute of Molecular Biology and Biotechnology (MicroVAX) and by Lydia Hartl at MIPS, Martinsried, Germany.

RESULTS AND DISCUSSION

Sequence analysis

The complete sequence of the 40,724 base pair segment of chromosome X has been deposited in the EMBL data base and appears under the accession number X87371. Our sequence analysis is in agreement with the published genetic map of chromosome X (Mortimer *et al.*, 1989).

The six phase ORF map of the sequenced region, performed by the DNA Strider program (Figure 1), revealed 33 ORFs with more than 100 codons. Table 1 includes the coordinates of all identified coding sequences. Seven ORFs that are totally internal to others (in different orientations or frames) were not included in our analysis. The expression of those ORFs is doubtful, until proven experimentally. The same applies for ORFs with less than 100 codons, which are not included in Table 1. The only exception presented in our analysis was the homologue of the *RPS25* (ribosomal protein) gene (87 aa). Three out of the 24 ORFs are partially overlapping with others (in different orientations or frames). The ORF sizes range from 100 to more than 2040 codons. The gene coding portion, including the two known snRNAs, constitutes 75–78% of the sequence (30,717–31,764 bases). This percentage is close to the average of 69–72% found for chromosomes II (Feldmann *et al.*, 1994), III (Oliver *et al.*, 1992), VIII (Johnston *et al.*, 1994) and XI (Dujon *et al.*, 1994).

The ORF sequences determined by us are identical to those of the known proteins Mim17p, Yak1p, Rpb4p, Yur1p, eIF-4ap, Mrs3p, Ura2p, to the *URA2* 3′ region ORF (J0682), to the *INO1* 5′ region ORF (J0630) and to the two small nuclear RNAs (J0635 and J0636). Some discrepancies with the previously published sequences of Vps35p and Ino1p will be discussed below.

The nucleotide data bases contain expressed sequence tags (EST) sequences (K. Weinstock and

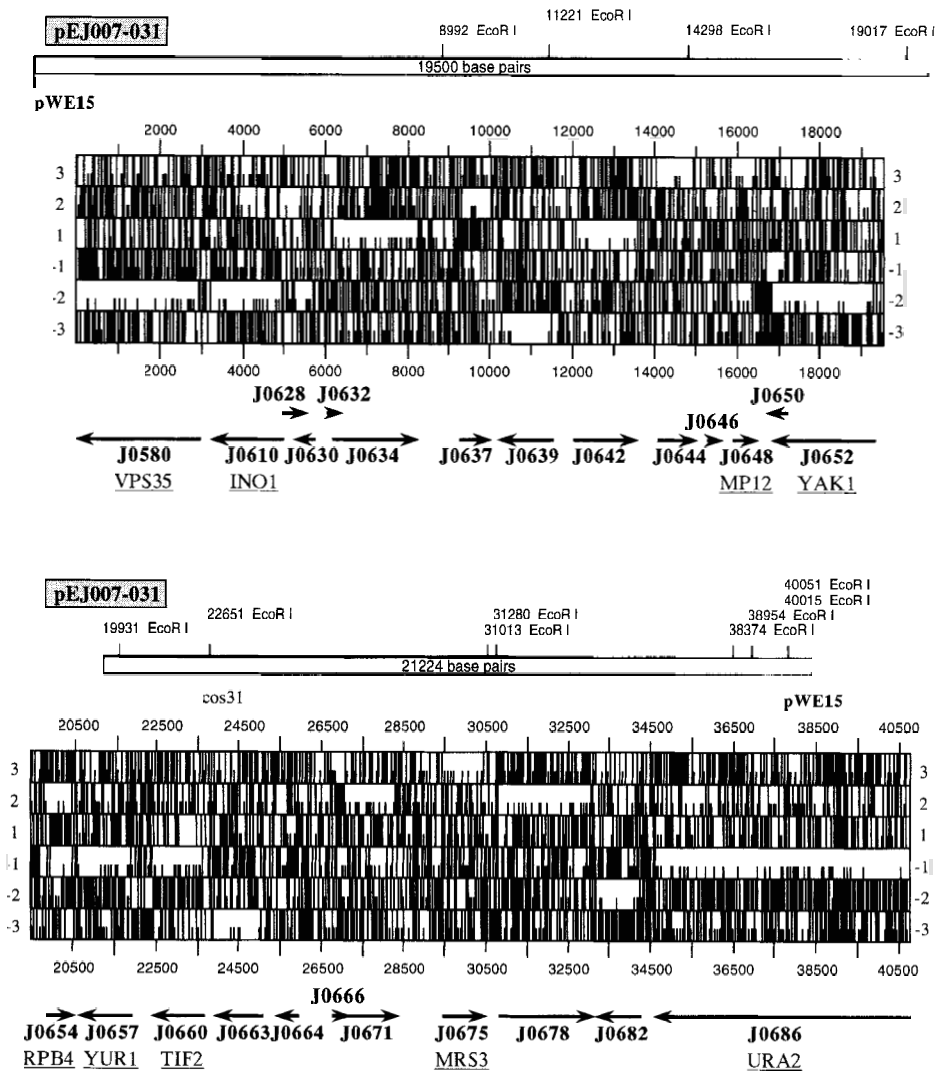

Figure 1. The *Eco*RI restriction map of the 40,724 base pairs of the overlapping cosmids pEJ007 and pEJ031 are presented in two parts. The 5′ and 3′ ends of the insert are *Sau*3AI sites cloned into the *Bam*HI site of pWE15 vector. The six-phase ORF map of the corresponding sequence is also presented below the *Eco*RI restriction map. Small bars indicate initiation codons and full bars indicate stop codons. The location and the direction of ORFs are indicated by arrows. The ORF nomenclature has been assigned by MIPS. Known gene names are given below ORF names.

J. C. Venter, unpublished) corresponding to several of the known or the newly identified ORFs and the snRNAS in the analysed DNA segment. Although all of these EST sequences contain nucleotide mismatches when compared to our sequences, they provide proof for their transcription. We have not found any ESTs corresponding to the internal (opposite orientation) ORFs.

ORF analysis

Several characteristics of the identified ORFs are provided in Table 1. The optimum and self scores (for better evaluation of each score's significance) of FastA analysis (Pearson and Lipman, 1988) revealed identities or homologies with known proteins and indicated disrepancies with

Table 1. Characteristics of open reading frames (ORFs) and other coding sequences identified in the 40,724 bp segment of chromosome X.

ORF name	ORF location (bases)	ORF length (aa)	Strand*	CAI	Fop	Identities, homologies, motifs†	Opt. score	Self score	Reference of gene or EST
J0580	2–2812	937	c	0·15	0·45	*S. c. VPS35* (937 aa) 99·1% identity in 937 aa	4552	4620	Paravicini *et al.* (1992) EMBL:S42186
J0610	3212–4876	555	c	0·18	0·52	*S. c. INO1* (553 aa) myo-inositol-1-phosphate synthase 87·4% identity in 538 aa	2153	2668	Dean-Johnson and Henry (1989) EMBL:J04453, L23520
J0611	3995–4348	188	w			Internal (opposite) of J0610			
J0628	5051–5407	119	w	0·07	0·35	Transmembrane (aa 86–102)			
J0630	5252–5650	133	c	0·16	0·49	*S. c. INO1* 5′ region (133 aa) 100% identity in 133 aa ESTs: Sc440, Sc903, Sc441 Asp rich (11·3%) Transmembrane (aa 21–37) Transmembrane (aa 47–63)	650	650	Dean-Johnson and Henry (1989) EMBL:J04453 EMBL:T36440, T36903, T36441
J0632	6000–6299	100	w	0·09	0·35	Transmembrane (aa 26–42)			
J0634	6256–8244	663	w	0·16	0·47	Leucine zipper (aa 72–93) Transmembrane (aa 376–392)			
J0635	8443–8570					*S. c.* snR128 100% identity in 128 bp			Zagorski *et al.* (1988) EMBL:M21124
J0636	8638–8827					*S. c.* snR190 100% identity in 190 bp			Zagorski *et al.* (1988) EMBL:M21124
J0637	9314–10012	233	w	0·20	0·46	ESTs: Sc17545 Lys rich (19·3%), Asp rich (11·2%)			EMBL:T17545
J0639	10299–11444	382	c	0·13	0·42	Leucine zipper (aa 13–34) Leu rich (15·4%)			
J0640	10300–10617	106	w			Internal (opposite) of J0639			
J0642	12169–13575	469	w	0·11	0·40	ESTs: Sc006, Sc733 Transmembrane (aa 277–293)			EMBL:T37006, T36733
J0644	14037–14918	294	w	0·22	0·54	Lys rich (12·2%)			
J0646	15236–15547	104	w	0·07	0·33	*S. c. MPI2* 5′ region (104 aa) 99% identity in 104 aa Asn rich (12·5%)	450	456	Maarse *et al.* (1994) EMBL:X77796
J0648	15978–16451	158	w	0·18	0·49	*S. c. MPI2,* Mim17p (158 aa) mitochondrial inner membrane 100% identity in 158aa	819	819	Maarse *et al.* (1994) EMBL:X77796
J0650	16699–17088	130	c	0·06	0·31	Leu rich (16·9%), Ile rich (11·5%) Transmembrane (aa 27–43) Transmembrane (aa 50–66) Transmembrane (aa 109–125)			

J0652	16847–19267	807	c	0·12	0·44	*S. c. YAK1* (807 aa) protein kinase 100% identity in 807 aa	3873	3873	Garrett and Broach (1989) GB:X16056
J0654	19838–20500	221	w	0·14	0·43	*S. c. RPB4* (221 aa) RNA pol II subunit 100% identity in 221 aa	980	980	Woychik and Young (1989) EMBL:M27253, X58099
J0655	19960–20271	104	c			Internal (opposite) of J0654			
J0657	20593–21876	428	c	0·14	0·42	*S. c. YUR1* (428 aa) 100% identity in 428 aa	2346	2346	Foreman *et al.* (1991) EMBL:X58099
J0660	22384–23568	395	c	0·75	0·87	*S. c. TIF2*, eIF-4a (395 aa) translation initiation factor 100% identity in 395 aa	1876	1876	Linder and Slonimski (1988) EMBL:X12814
J0663	23865–25004	380	c	0·14	0·44	*S. c. YKR058w* (618 aa) hypothetical protein 41·8% identity in 376 aa Leucine zipper (aa 103–124)	861	2136	van Vliet-Reedijk and Planta (1994) EMBL:Z28283, X56444 Corrected sequence (R. J. Planta, personal communication)
J0664	25430–25666	87	c			*S. c. RPS25,* Rs21p (87 aa) ribosomal protein	406	407	van Vliet-Reedijk and Planta (1994) EMBL:Z28282, P05760
	26127–26150					98·9% identity in 87 aa ESTs: Sc091, Sc900, Sc221			EMBL:T37091, T36900, T37221
J0666	26754–27068	105	w	0·14	0·44	Leu rich (14·3%), Lys rich (12·4%)			
J0671	27065–28291	409	w	0·11	0·41	*S. c. YKR053c* (404 aa) hypothetical protein 56·4% identity in 381 aa Transmembrane (aa 187–203) Transmembrane (aa 212–228) Transmembrane (aa 314–330) Transmembrane (aa 393–409)	1295	2260	Vissers *et al.* (1994) EMBL:Z28278
J0675	29496–30437	314	w	0·08	0·37	*S. c. MRS3* (314 aa) mitochondrial RNA splicing 100% identity in 314 aa	1541	1541	Wiesenberger *et al.* (1991) EMBL:X56445, X06239
J0678	30791–33040	750	w	0·12	0·42	Human phospholipase D (817 aa) 33·6% identity in 116 aa Transmembrane (aa 3–19)	187	4272	Tsang *et al.* (1992) EMBL:L11701
J0679	32611–32943	111	c			Internal (opposite) of J0678			
J0682	33158–34225	356	c	0·12	0·40	*S. c. URA2* 3′ region (124 aa) 100% identity in 124 aa ESTs: Sc165	585	585	Souciet *et al.* (1989) EMBL:M27174 EMBL:T37165
J0686	34603–40722	>2040	c	0·28	0·59	*S. c. URA2* (2214 aa) carbamyl phosphate synthetase 100% identity in 2040 aa	9651	10502	Souciet *et al.* (1989) EMBL:M27174
J0687	35004–35306	101	c			Internal of J0686			
J0688	35342–35809	156	w			Internal (opposite) of J0686			
J0689	38031–38339	103	c			Internal of J0686			

*w and c indicate the direction of translation (Watson and Crick strands).
†Motifs are described only for the newly identified ORFs.

previously published sequences. Scores higher than 200 have been considered as significant, although in some instances lower scores due to homologies in restricted areas of the protein sequences indicated conservation of specific domains. Only the best homology found is included in Table 1. Homologies of questionable significance are discussed below for each individual ORF. The codon adaptation index (CAI; Sharp and Li, 1987) and the frequency of optimal codons (Fop; Sharp and Cowe, 1991) (determined by Lydia Hartl at MIPS) are given as parameters indicating the probability for the corresponding ORFs to be expressed. ORFs with CAI <0·110 are considered questionable. That CAI range corresponds to Fop numbers <0·4. Potential transmembrane regions have been determined according to Klein *et al.* (1985) (by Lydia Hartl 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. Below we discuss our major findings on each identified ORF in the same order as they are presented in Table 1.

ORF J0580 differs from the published sequence of the vacuolar protein sorting Vps35p by the insertion of seven amino acids close to the amino terminus of the protein. This difference does not affect the hydrophobicity profile of the protein (data not shown) and probably reflects strain polymorphism.

The ORF J0610 sequence exhibits many differences with the published sequences of the L-myo-inositol-1-phosphate synthase Ino1p. There are two published sequences with accession numbers J04453 and L23520 (strains AB320 and A322 respectively). Our sequence is closer to that in file L23520. The main differences are included at the amino terminus, within the first 120 residues. In that region, our sequence is more similar to the published *Candida albicans* Ino1p sequence (accession number L22737; Klig *et al.*, 1994). The same holds true when comparing the hydrophobicity profiles of the three different amino acid sequences (data not shown). Moreover, our nucleotide sequence is identical to the EST sequence Sc467 (accession number T36467) in the discussed region. These facts indicate discrepancies due to previous sequencing errors in addition to strain polymorphisms.

ORFs J0628 and J0632, which overlap with ORFs J0630 and J0634 respectively, probably do not correspond to expressed sequences as indicated by their unfavourable codon indices. Each sequence contains a potential transmembrane region as described in Table 1.

In contrast, ORF J0630 has favourable codon indexes and corresponds to several EST sequences (Table 1). It contains two potential transmembrane regions and it shows some restricted similarity to a hypothetical protein on yeast chromosome IV (MIPS, unpublished data; CDYS:RRA140, FastA scores 127/745) and to the hypothetical protein ZK632.10 of *Caenorhabtitis elegans* (accession number P34655, 32·6% identity in 89 aa overlap, FastA scores 101/490).

The sequence of ORF J0634 that has no known homologue, contains a leucine zipper motif, a potential transmembrane region and a polyaspartic acid stretch (loop-forming acidic region; residues 533–539). It is distinct from ORF J0632 and not the result of a frameshift, as indicated by the pronounced differences in their codon index values.

ORF J0637 is probably expressed since it corresponds to an EST sequence. It is a very hydrophilic molecule as indicated by its hydrophobicity profile (not shown). It contains a carboxy terminus highly rich in lysine, aspartic acid and glutamic acid residues (alternate positively and negatively charged residues). This region contributes to the apparent similarity of the ORF with the hypothetical yeast protein L8167.9 from chromosome XII (accession number U14913, S48550), which was the best homology detected considering the FastA scores (135/2351), the length (25·3% identity in 150 aa overlap) and the location of the similarities in the two molecules. Several other yeast hypothetical proteins appeared in the homology search with lower scores, such as YKL202w, YKL201c and YKL023w.

The main information characterizing the J0639 ORF sequence is a leucine zipper motif at its amino terminus. ORF J0642 is potentially expressed, since it corresponds to several EST sequences, and contains a potential transmembrane region. Only borderline similarities were found between the J0644 ORF sequence and a hypothetical protein on yeast chromosome XIV (MIPS, unpublished data; CDYS:VEE351, FastA scores 129/1825). Similarly low homologies, spanning the middle portion of both compared protein sequences, were found with the hypothetical protein T23G5.2 of *C. elegans* (accession number S28303, 34·6% identity in 81 aa overlap, FastA

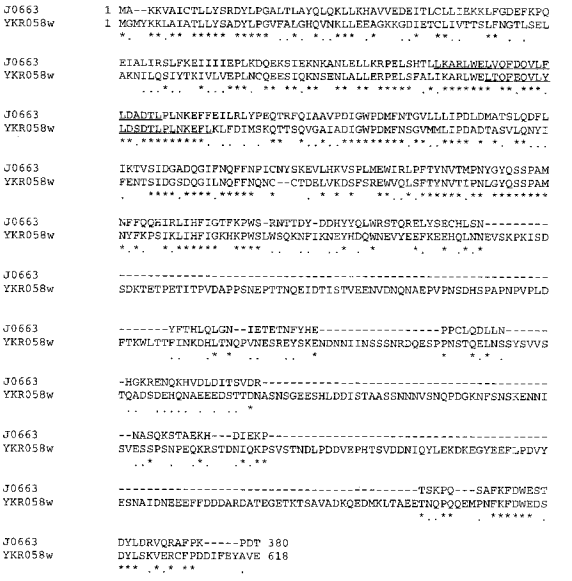

Figure 2. CLUSTAL alignment (Higgins and Sharp, 1988) of the J0663 ORF sequence (380 aa) with the hypothetical protein YKR058w (618 aa). Asterisks indicate residue identities and dots indicate conservative substitutions. A potential leucine zipper sequence motif is underlined in each sequence.

scores 128/2419), with the yeast Sec14p phosphatidylinositol-phosphatidylcholine transfer protein (accession number S43745, 28·3% identity in 127 aa overlap, FastA scores 125/2558) and with two EST sequences of *Arabidopsis thaliana* (accession numbers T76862, T76582).

ORF J0646 is known to be located on the 5′ site of the *MP12* gene. Its sequence has no similarity to any known sequences. It contains a cluster of seven asparagine residues (residues 72–78). ORF J0650, which overlaps with the 3′ end of the Yak1p, contains mostly hydrophobic residues with three potential transmembrane spans. ORFs J0646 and J0650 have low CAI and Fop indices.

ORF J0663 (380 aa) appeared homologous to the yeast hypothetical protein YKR058w (480 aa) as shown in Figure 2. ORF J0663 is missing an internal region present in YKR058w but the two yeast sequences are related, as indicated by the clusters of identical amino acids throughout their common regions extending from their amino to their carboxy termini. The first 260 residues of both sequences contain conserved domains characteristic of the mammalian glycogenin glycosyltransferases (Viskupic *et al.*, 1992; a rabbit sequence, accession number A45094, 34·4% identity in 189 aa overlap, FastA scores 270/1652; a

Figure 3. Multiple alignment of the two yeast hypothetical ORF sequences J0663 and YKR058w with the two mammalian glycogenin glycosyltransferases A45094 (rabbit) and S45141 (human) using the CLUSTAL program.

human sequence, accession number S45141, 33% identity in 179 aa overlap, FastA scores 243/1401; and a bovine sequence, accession number L01792, 35·5% identity in 172 aa overlap, FastA score 247); Figure 3. All three mammalian sequences are very similar to each other (>80% identity). The J0663 ORF product is more similar to the glycogenin glycosyltransferases, both in sequence and in length, than its yeast counterpart YKR058w, probably being a closer relative (not shown).

ORF J0664 was included in Table 1, although it is 87 aa long, because it was found to be almost identical to the *YKR057w* (*RPS25*) gene encoding the ribosomal protein Rs21p. Residue 51 in J0664 is an isoleucine instead of a valine. The same unique substitution was found in the EST sequence Sc091. The protein sequence of ORF J0664 results from the splicing of two exons, similar to that of Rs21p. The corresponding intron sizes are 461 and 321 bases.

ORF J0666, which overlaps with ORF J0671, has no distinct characteristic but its codon parameters are favourable for expression. It is probably distinct from ORF J0671 since the starting site of ORF J0671 correlates well with that of the highly similar hypothetical protein YKR053c.

The similarities of the J0671 and YKR053c ORF sequences are shown in Figure 4. ORF J0671 contains four potential transmembrane regions and its hydrophobicity profile is similar to that of the YKR053c product. Both ORFs are probably membrane proteins, differing mostly at the 30 amino-terminal residues, as expected for leader sequences that usually diverge in amino acid sequence, preserving mostly the appropriate charge (Von Heijne, 1983). ORF J0671 showed in

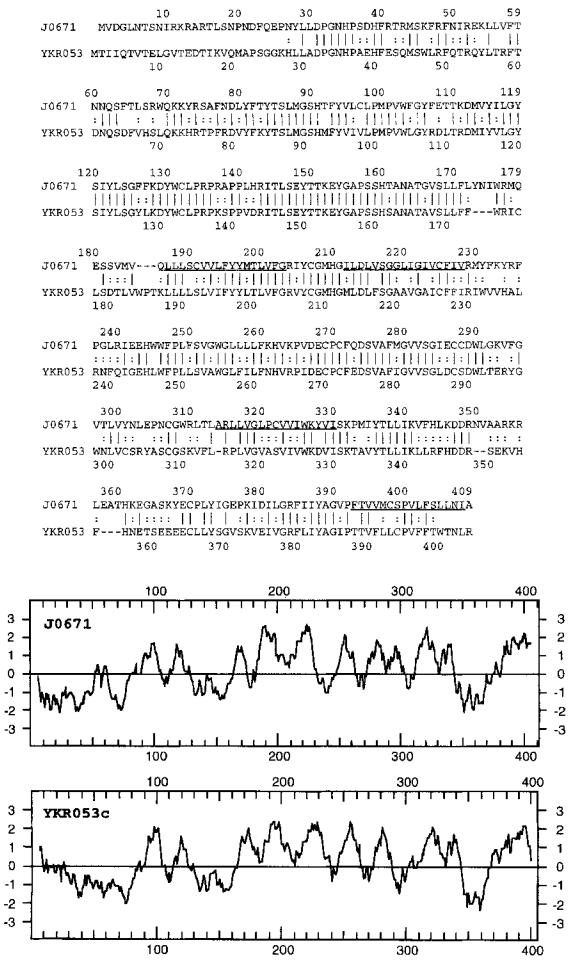

Figure 4. FastA alignment of the hypothetical proteins J0571 and YKR053c (top). The inner boundaries of four potential transmembrane spans are underlined in ORF J0671. The hydrophobicity profiles of the J0571 and YKR053c ORF products (bottom).

addition homologies in restricted areas of the *E. coli* hypothetical protein o489 (S47716, 23·5% identity in 166 aa overlap, FastA scores 141/2558) and the human latent membrane protein LMP1 (S21660, 25·8% identity in 93 aa overlap, FastA scores 123/2085).

The sequence of the J0678 ORF product has similarities in scattered restricted domains (mainly in the amino-terminal half) with the human phospholipase D precursor, as shown in Figure 5. The same regions are also conserved in the bovine phospholipase D sequence (not shown). ORF J0678 contains one potential transmembrane hydrophobic stretch (residues 3–19) at the amino

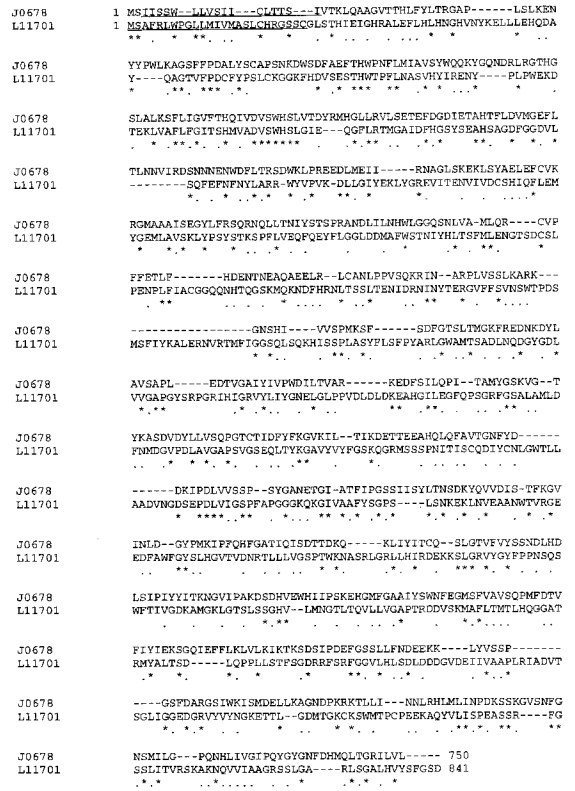

Figure 5. CLUSTAL alignment of the J0678 ORF sequence (750 aa) with the human phospholipase D precursor L11701 (841 aa). Asterisks indicate residue identities and dots indicate conservative substitutions. Potential signal peptides are underlined.

terminus corresponding to the hydrophobic phospholipase D signal peptide. The overall hydrophobicity profiles of the yeast and mammalian protein products are similar (not shown).

The product of ORF J0682, previously identified at the 3′ end of the *URA2* gene, has the profile of a hydrophilic protein. It is expressed since it corresponds to EST sequences.

A cluster of six genes is conserved between chromosomes X and XI

We have presented three newly sequenced ORFs, J0663, J0664 and J0671, which showed significant homologies to three ORFs closely arranged on chromosome XI (Dujon *et al.*, 1994). Moreover, the *TIF2* gene, identified next to ORF J0663, was previously known to be identical in amino acid sequence to the *TIF1* gene located next to the above-mentioned homologues on

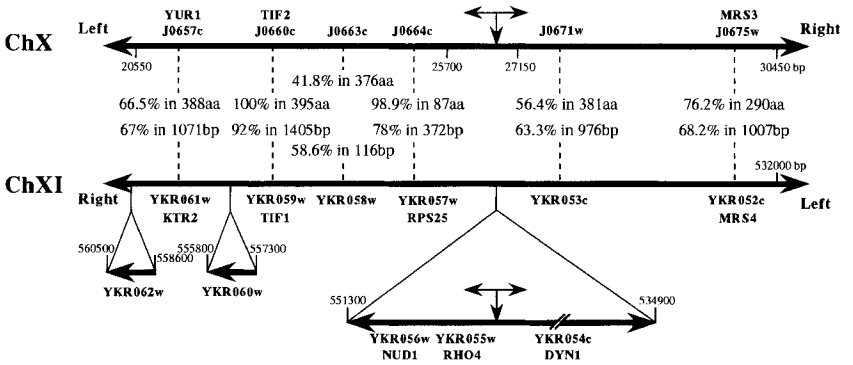

Figure 6. Diagram of the similarities between the gene cluster on the left arm of chromosome X and the gene cluster on the right arm of chromosome XI. The sequence number coordinates correspond to the cosmid sequence for chromosome X and to the entire sequence for chromosome XI. The percentages of identities in amino acid and nucleotide sequence overlaps are shown for pairs of homologous genes. The inserts below the chromosome XI line indicate the ORFs that have no homologues on the chromosome X cosmid region. The vertical arrows pointing to areas of no homologies indicate the borders of the divergently expressed gene clusters on each chromosome. ORF YKR062w has only been included in the diagram as the last one in the group with similar orientation of transcription.

chromosome XI. In addition, we have noticed that the region of homology between the two chromosomes extends to previously known genes with similar functions (*YUR1/KTR2*, *MRS3/MRS4*). As shown in Figure 6, a region of about 10 kb (included in the reported cosmid sequence), approximately 151–161 kb from the left end of chromosome X, is related to a region of 28·5 kb approximately 106–134·5 kb from the right end of chromosome XI. Six contiguous ORFs on chromosome X are arranged into two groups of divergently transcribed genes. The corresponding six genes on chromosome XI are similarly arranged, but one group additionally contains four ORFs with no homologues on chromosome X and the other group contains the very long gene *DYN1* also with no counterpart on chromosome X. These intermingled unrelated genes have the same direction of expression with the others in the group to which they belong. Finally, it should be mentioned that there is not any obvious common feature or function between the genes that are transcribed from the same DNA strand.

The emerging picture from this analysis is one of a translocation event having occurred between opposite arms of the two chromosomes which preserved the orientations relative to the ends. The non-homologous regions may have been deleted or added by further recombination events. The observed percentages of nucleotide and amino acid identities between the homologous regions imply that different genes have evolved with different rates and/or mechanisms. Additional recombination or gene conversion events have possibly resulted in the observed fluctuation of alternate medium and very high conservation. For example, it is striking that the J0663/YKR058 ORF pair is significantly similar only in amino acid sequence and the immediately flanking regions on each side, *TIF1/TIF2* and J0664/YKR057 pairs, are extremely conserved in both amino acid and nucleotide sequences. The extent of each gene pair's divergence is probably indicative of the functional diversification of the two homologues.

Gene redundancy is a fact even for organisms with small genomes, such as yeast. The analysis of accumulating sequence data from various yeast chromosomes has revealed a considerable number of similar genes with various degrees of homologies scattered on different chromosomes (Bussey *et al.*, 1995; Dujon *et al.*, 1994; Feldmann *et al.*, 1994; Johnston *et al.*, 1994). In fact, there is an increasing probability that the currently sequenced ORFs of known or unknown function will have previously sequenced homologues. In some instances, unclear divergence or convergence events have resulted in experimentally proven functionally related genes that are only homologous at the amino acid sequence level embedded in totally unrelated DNA (Georgatsou

and Alexandraki, 1994). On the other hand, multiple copies of highly conserved functionally redundant genes have been found on the subtelomeric regions of yeast chromosomes (Pryde *et al.*, 1995; Dujon *et al.*, 1994; Alexandraki and Tzermia, 1994). In this report we present the case of redundancy of genes as a group located in analogous internal regions on two different chromosomes. Detailed analysis of several similar examples of gene cluster evolution may clarify the mechanisms responsible both for gene redundancy and for the preservation/diversification of entire regions on chromosomes.

ACKNOWLEDGEMENTS

We thank Francis Galibert and Jean-Claude Chuat for coordinating the chromosome X sequencing project. We thank Lydia Hartl and all MIPS staff for help with the sequence analysis. We also thank Yannis Papanikolaou for helpful discussions and 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

Alexandraki, D. and Tzermia, M. (1994). Sequencing of a 13·2 kb segment next to the left telomere of yeast chromosome XI revealed five open reading frames and recent recombination events with the right arms of chromosomes III and V. *Yeast* **10,** S81–S91.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (Eds) (1987). *Current Protocols in Molecular Biology.* Greene Publishing Associates and Wiley-Interscience, New York.

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

Bussey, H., Kaback, D. B., Zhong, W., *et al.* (1995). The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA*, **92,** 3809–3813.

Dean-Johnson, M. and Henry, S. A. (1989). Biosynthesis of inositol in yeast. Primary structure of myo-inositol-1-phosphate synthase and functional analysis of its structural gene, the *INO1* locus. *J. Biol. Chem.* **264,** 1274–1283.

Dujon, B., *et al.* (1994). The complete DNA sequence of chromosome XI of *Saccharomyces cerevisiae* (666 kb). *Nature* **369,** 371–378.

Feldmann, H., *et al.* (1994). Complete DNA sequence of yeast chromosome II. *EMBO J.* **13,** 5795–5809.

Garrett, S. and Broach, J. (1989). Loss of Ras activity in *Saccharomyces cerevisiae* is suppressed by disruptions of a new kinase gene, *YAK1*, whose product may act downstream of the cAMP-dependent protein kinase. *Genes Dev.* **3,** 1336–1348.

Georgatsou, E. and Alexandraki, D. (1994). Two distinctly regulated genes are required for ferric reduction, the first step of iron uptake in *Saccharomyces cerevisiae. Molec. Cell. Biol.* **14,** 3065–3073.

Higgins, D. G. and Sharp, P. M. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73,** 237–244.

Huang, M.-E., Chuat, J.-C., Thierry, A., Dujon, B. and Galibert, F. (1994). Construction of a cosmid contig and of an *Eco*RI restriction map of yeast chromosome X. *DNA Sequence* **4,** 293–300.

Johnston, M., *et al.* (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science* **265,** 2077–2082.

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

Klig, L. S., Zobel, P. A., Devry, C. G. and Losberger, C. (1994). Comparison of *INO1* gene sequence and products in *Candida albicans* and *Saccharomyces cerevisiae. Yeast* **10,** 789–800.

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

Linder, P. and Slonimski, P. P. (1988). Sequence of the genes *TIF1* and *TIF2* from *Saccharomyces cerevisiae* coding for a translation initiation factor. *Nucl. Acids Res.* **16,** 10359.

Maarse, A. C., Blom, J., Keil, P., Pfanner, N. and Meijer, M. (1994). Identification of the essential yeast protein MIM17, an integral mitochondrial inner membrane protein involved in protein import. *FEBS Lett.* **349,** 215–221.

Marck, C. (1988). 'DNA Strider': a 'C' programme 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.

Oliver, S. G., *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46.

Paravicini, G., Horazdovsky, B. F. and Emr, S. D. (1992). Alternative pathways for the sorting of soluble vacuolar proteins in yeast: a *vps35* null mutant missorts and secretes only a subset of vacuolar hydrolases. *Mol. Biol. Cell* **3,** 415–427.

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

Pryde, F. A., Huckle, T. C. and Louis, E. J. (1995). Sequence analysis of the right end of chromosome XV in *Saccharomyces cerevisiae*: an insight into the structural and functional significance of sub-telomeric repeat sequences. *Yeast* **11,** 371–382.

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.

Souciet, J. L., Nagy, M., Le Gouar, M., Lacroute, F. and Potier, S. (1989). Organisation of the yeast *URA2* gene: identification of a defective dihydroorotase-like domain in the multifunctional carbamoylphosphate synthetase-aspartate transcarbamylase complex. *Gene* **79,** 59–70.

Tsang, T. C., Fung, W. J. C., Levine, J., Metz, C. N., Davitz, M. A., Burns, D. K., Huang, K. S. and Kochan, J. P. (1992). Isolation and expression of two human glycosylphosphatidylinositol phospholipase D (GPI-PLD) cDNAs. *FASEB J.* **6,** 1922.

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.

Viskupic, E., Cao, Y., Zhang, W., Cheng, C., Depaoli-Roach, A. A. and Roach, P. J. (1992). Rabbit skeletal muscle glycogenin. Molecular cloning and production of fully functional protein in *Escherichia coli*. *J. Biol. Chem.* **267,** 25759–25763.

Von Heijne, G. (1983). Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* **133,** 17–21.

Wiesenberger, G., Link, T. A., von Ahsen, U., Waldherr, M. and Schweyen, R. J. (1991). *MRS3* and *MRS4*, two suppressors of mtRNA splicing defects in yeast, are new members of the mitochondrial carrier family. *J. Mol. Biol.* **217,** 23–37.

Woychik, N. A. and Young, R. A. (1989). RNA polymerase II subunit RPB4 is essential for high- and low-temperature yeast cell growth. *Mol. Cell. Biol.* **9,** 2854–2859.

Zagorski, J., Tollervey, D. and Fournier, M. J. (1988). Characterisation of an SNR gene locus in *Saccharomyces cerevisiae* that specifies both dispensable and essential small nuclear RNAs. *Mol. Cell. Biol.* **8,** 3282–3290.