The EMBO Journal vol.15 no.9 pp.2031–2049, 1996

Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X

F.Galibert[1,2], D.Alexandraki[3], A.Baur[4], E.Boles[4], N.Chalwatzis[4], J.-C.Chuat[1], F.Coster[5], C.Cziepluch[6], M.De Haan[7], H.Domdey[8], P.Durand[9], K.D.Entian[10], M.Gatius[1], A.Goffeau[5], L.A.Grivell[7], A.Hennemann[10], C.J.Herbert[11], K.Heumann[12], F.Hilger[9], C.P.Hollenberg[13], M.-E.Huang[1], C.Jacq[14], J.-C.Jauniaux[6], C.Katsoulou[3], L.Kirchrath[13], K.Kleine[12], E.Kordes[6], P.Kötter[10], S.Liebl[12], E.J.Louis[15], V.Manus[1], H.W.Mewes[12], T.Miosga[4], B.Obermaier[8,16], J.Perea[14], T.Pohl[17], D.Portetelle[9], A.Pujol[6], B.Purnelle[5], M.Ramezani Rad[13], S.W.Rasmussen[18], M.Rose[10], R.Rossau[19], I.Schaaff-Gerstenschläger[4], P.H.M.Smits[7], T.Scarcez[19], N.Soriano[1], D.Tovan[14], M.Tzermia[3], A.Van Broekhoven[19], M.Vandenbol[9], H.Wedler[20], D.Von Wettstein[18], R.Wambutt[20], M.Zagulski[11,21], A.Zöllner[12] and L.Karpfinger-Hartl[12]

[1]UPR 41 CNRS Recombinaisons Génétiques, Faculté de Médecine, 2 avenue du Professeur Léon Bernard, F-35043 Rennes Cedex, France, [3]Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, PO Box 1527, Heraklion, GR-71110 Crete, Greece, [4]Institut für Mikrobiologie und Genetik, Technische Hochschule Darmstadt, Schnittspahnstrasse 10, D-64287 Darmstadt, Germany, [5]Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2, Bâtiment 20, B-1348 Louvain-La-Neuve, Belgium, [6]Tumorvirologie Abteilung 0610 and Virologie Appliquée à l'Oncologie Unité INSERM U375, Deutsches Krebsforschungszentrum, D-69120 Heidelberg, Germany, [7]University of Amsterdam, Section for Molecular Biology, Kruislaan 318, NL-1098 SM Amsterdam, The Netherlands, [8]Genzentrum, Institut für Biochemie, Würmtalstrasse 221, D-81373 München, Germany, [9]Unité de Microbiologie, Faculté des Sciences Agronomiques de Gembloux, avenue Maréchal Juin 6, B-5030 Gembloux, Belgium, [10]Institut für Mikrobiologie, J.W.Goethe-Universität Frankfurt, Marie-Curie-Strasse 9, Geb. N250, D-60439 Frankfurt/Main, Germany, [11]UPR 2420 CNRS Centre de Génétique Moléculaire, Bâtiment 26, Avenue de la Terrasse, F-91198 Gif-sur-Yvette cedex, France, [12]MIPS am Max-Planck-Institut für Biochemie, D-82152 Martinsried bei München, Germany, [13]Institut für Mikrobiologie der Heinrich-Heine-Universität Düsseldorf, Geb. 26.12, Universitätsstrasse 1, D-40225 Düsseldorf, Germany, [14]URA 1302 CNRS Génétique Moléculaire, Ecole Normale Supérieure, 46 rue d'Ulm, F-75230 Paris Cedex 05, France, [15]Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK, [17]GATC GmbH, Gesellschaft für Analyse Technik und Consulting, Fritz-Arnold-Strasse 23, D-78467 Konstanz, Germany, [18]Carlsberg Laboratory, Department of Physiology, Gamle Carlsberg vej 10, Valby, DK-2500 Copenhagen, Denmark, [19]Innogenetics, Industriepark Zwijnaarde 7, Box 4, B-9052 Ghent, Belgium and [20]AGON GmbH, Gesellschaft für molekularbiologische Technologie mbH, Glienicker Weg 185, D-12489 Berlin, Germany

[16]Present address: MediGene GmbH, Lochhamer Strasse 11, D-82152 Martinsried bei München, Germany

[21]Present address: Institute of Biochemistry and Biophysics, 5a Pawinskiego St., 02-106 Warsaw, Poland

[2]Corresponding author

The complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X (745 442 bp) reveals a total of 379 open reading frames (ORFs), the coding region covering ~75% of the entire sequence. One hundred and eighteen ORFs (31%) correspond to genes previously identified in *S.cerevisiae*. All other ORFs represent novel putative yeast genes, whose function will have to be determined experimentally. However, 57 of the latter subset (another 15% of the total) encode proteins that show significant analogy to proteins of known function from yeast or other organisms. The remaining ORFs, exhibiting no significant similarity to any known sequence, amount to 54% of the total. General features of chromosome X are also reported, with emphasis on the nucleotide frequency distribution in the environment of the ATG and stop codons, the possible coding capacity of at least some of the small ORFs (<100 codons) and the significance of 46 non-canonical or unpaired nucleotides in the stems of some of the 24 *tRNA* genes recognized on this chromosome.

Keywords: chromosome X/gene duplication/open reading frame/*Saccharomyces cerevisiae*/tRNA

Introduction

The traditional methods of genetic analysis involve tracing modified phenotypes back to genotypic alterations. The limit of this approach is an imperceptible modification of the phenotype. The international yeast genome systematic sequencing programme launched in 1989 by the European Communities, aiming at establishing the complete genetic information of bakers' yeast, *Saccharomyces cerevisiae*, has demonstrated the limitations of classical genetics. The pilot sequencing of chromosome III (Oliver *et al.*, 1992) has demonstrated that disruption of a large number of the newly revealed open reading frames (ORFs) does not result in any phenotypic alteration. Subsequent systematic sequencing of seven more chromosomes (Barrell *et al.*, 1994; Dietrich *et al.*, 1994; Dujon *et al.*, 1994; Feldmann *et al.*, 1994; Johnston *et al.*, 1994; Bussey *et al.*, 1995; Murakami *et al.*, 1995) has confirmed that a large proportion of the novel genes cannot be assigned any known function, while on the other hand a large number of proteins unrelated to database entries are being discovered. Last but not least, it stems from numerous cytological studies of chromosome behaviour during the vegetative and meiotic cell cycle that a chromosome is more than its mere genetic content. By making available the complete

© Oxford University Press

Table I. Estimated overall accuracy of chromosome X sequence

	Total bp verified	Number of modified nt[a]			Error rate (‰)
		M	G	T	
Overlap between regions	46 455	11	13	24	0.52
Resequenced regions[b]	~50 000	10	7	17	0.34

[a]M, mismatch; G, gap; T, total mismatches plus gaps.
[b]Occasional overlaps between verification clone sequences were excluded from the calculations.

DNA sequence of a chromosome, parameters not entirely confined to its role as carrier of genetic information may be exposed for analysis. A survey of a new object is thus provided, even though all the topological implications of the results cannot be fully grasped at the present stage and must await at least the completion of the yeast genome enterprise. This paper describes the DNA sequence of chromosome X.

Results

Assembly of the sequence

The sequence was determined from a set of 26 partially overlapping cosmids selected on the basis of an *Eco*RI map based on a cosmid contig of chromosome X (Huang *et al.*, 1994a). These cosmids were distributed within a consortium of 15 contractors. The telomeres were independently isolated and sequenced. While the left-telomere-containing clone was found to overlap with the left terminal cosmid of the chromosome, this was not so at the other end, where no overlap was detected between the right-most cosmid and a right-telomere-containing clone 9.0 kb in size. The missing portion (a few kb) was PCR-amplified from a yeast S288C genomic DNA template using primers designed from sequences flanking the gap. When all bases had been determined by each contractor and each sequencing strategy had been approved by the DNA coordinator, ensuring that the sequence had been independently determined on each strand with sufficient overlap between all the subclones, the sequences were considered as final and entered into the MIPS data library for assembly. Partial sequences of chromosome X have been published independently by some of the authors of this work (Huang *et al.*, 1994b, 1995; Miosga *et al.*, 1994a,b,c, 1995; Purnelle *et al.*, 1994; Vandenbol *et al.*, 1994, 1995; Rasmussen, 1995; Zagulski *et al.*, 1995).

Verification of the sequence

Quality controls were performed concomitantly with sequence assembly. The aim of the project was to keep the error rate as low as possible, with a target $<10^{-4}$. Three procedures were employed to track down errors, including checking sequencing strategy by the coordinator, matching overlapping portions sequenced by independent contractors and finally random resequencing (see Materials and methods for details). The results of the last two procedures are shown in Table I. From these data, the error rate of the yeast chromosome X sequence presented in this paper can be estimated to be 0.4‰, a value of the same order as that reported in similar studies.

General organization of chromosome X

Analysis of the entire nucleotide sequence of chromosome X (745 442 bp) confirms the general features of chromosome organization observed in other systematically sequenced yeast chromosomes. The coding region occupies 74.04% of the sequence, 36.59% and 37.45% on the Watson and Crick strand, respectively.

The average base composition is 38.9% G+C. As expected, the coding regions have a higher than average G+C content (40.2%) than the non-coding (35.6%). The distribution of dinucleotide frequencies over the whole chromosome is the same in the coding and the non-coding regions of either strand. The deviations of the frequencies of complementary dinucleotide pairs tend to occur in the same direction. In contrast to what was reported for chromosomes XI and II, the homopurine pairs do not seem to be in excess in the coding region of either strand (Figure 1). Some compositional periodicity has been noted, at least in the case of chromosomes XI and II, with waves of G+C-rich regions correlating with waves of high gene density. By using the same algorithm, a similar G+C pattern emerges with chromosome X, especially in the right-hand part of the chromosome. This pattern correlates rather well with the gene density plot, as illustrated by the two deep depressions around 200 kb and 470 kb in Figure 2.

Telomeres and centromere

The telomere regions of chromosome X are similar to the other sequenced yeast telomeres. Adjacent to the $C_{1-3}A$ repeat at the left telomere are a Y' element (coordinates 61–6931) and the core X element (7305–7767) shared by most if not all yeast telomeres (Louis *et al.*, 1994; Pryde *et al.*, 1995). However, the X–Y' junction does not contain the usual subtelomeric repeats STR-D, STR-C, STR-B and STR-A, but instead has (6998–7224) part of a copy (Louis and Haber, 1991) of the fourth intron of cytochrome *b* encoded by mitochondrial DNA (Delehodde *et al.*, 1989). A copy of bi4 is also found at the left telomere of chromosome IX (Louis and Haber, 1991; Barrell *et al.*, 1994). In fact, the left ends of chromosomes IX and X share a large, nearly identical block of sequence similarity spanning >21 kb. The right telomere of chromosome X is more conventional, with a core X element (744 593–745 052) and the STR-D, STR-C, STR-B and STR-A elements adjacent to the TG_{1-3} repeats (745 357–end). The core X elements of both ends contain the *ARS1* consensus and the Abf1p binding site found in most core Xs. These elements that are shared by most ends may have functional significance. The right telomere region is analogous to several other sequenced telomeres (II right

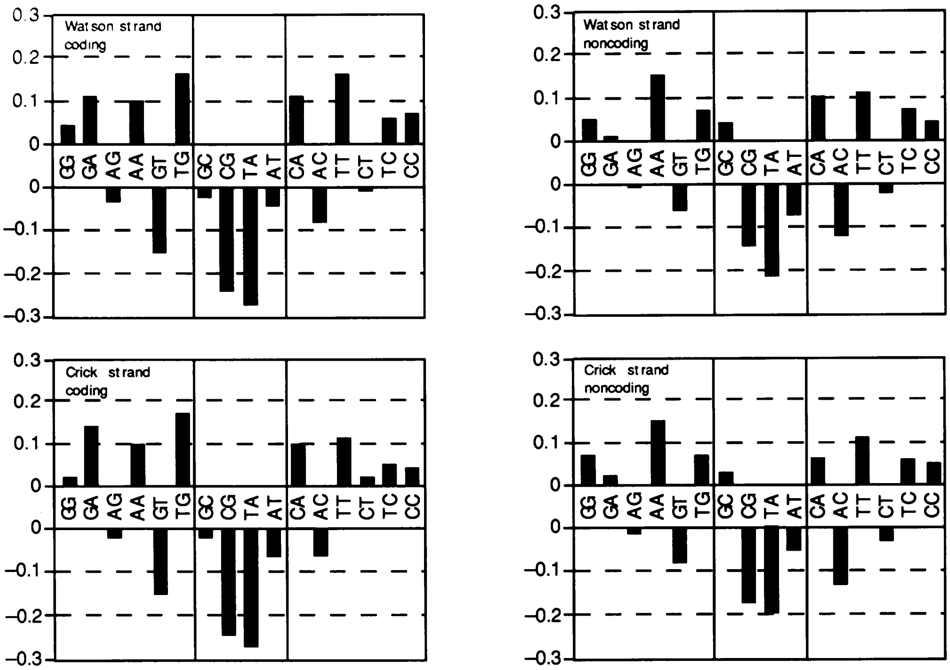


Fig. 1. Distribution of dinucleotide frequencies in the coding and non-coding regions of the two strands of chromosome X. Vertical bars show relative deviations [i.e. (observed–expected)/expected]. Expected frequencies are calculated from mononucleotide frequencies. Complementary pairs are arranged as mirror images. The four self-complementary pairs are placed in the central part.

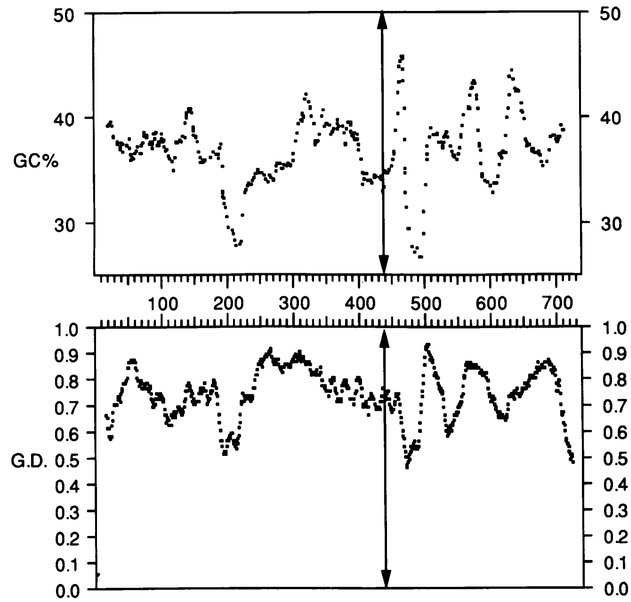


Fig. 2. Compositional variation and gene density distribution along chromosome X. Top: compositional variation calculated as in Dujon *et al.* (1994). Each point represents the average G+C composition calculated from the third base of each codon. Bottom: gene density expressed as the fraction of nucleotides within ORFs in sliding windows of 30 kb. The position of the centromere is indicated by an arrow.

and left, V right and left, VI left, VIII right and left, IX right, XI left) over the last 3–4 kb.

The centromere of chromosome X of strain R95-4A, a derivative of S288C, was isolated by Hieter *et al.* (1985) by selection of yeast DNA fragments capable of suppressing lethality of the *SUP11* gene in high copy number. Comparison of this sequence with that reported in the present paper shows complete identity and enables location of the chromosome X centromere at positions 435 996–436 112. *CEN10* conforms to the consensus structure established for other centromeres.

ORFs and their predicted protein products

By definition, an ORF is considered from its first in-phase ATG codon. Only those ORFs containing at least 99 contiguous sense codons following an ATG, and not entirely contained within a longer ORF in a different reading frame or on the other DNA strand, have been retained for further analysis. The special case of ORFs shorter than 100 codons is described below. A total of 379 ORFs were recorded in the entire chromosome X using this principle (Table II), leaving aside the retroposons, i.e. a density of one ORF/1967 bp. Twelve of these ORFs are interrupted by introns. Table II includes 39 partially overlapping ORFs. Ten are on the same DNA strand, all others being antiparallel overlaps. Informatic and statistical analysis revealed that ORFs both shorter than 150 codons and with a codon adaptation index (CAI) (Sharp and Li, 1987) <0.11 may correspond to randomly occurring ORFs rather than to real genes (Dujon *et al.*, 1994). If these criteria are applied to the ORFs identified in chromosome X, 23 of the 379 ORFs are questionable genes. Thirteen of these belong to the set of partially overlapping ORFs. However, three genes of known function (*YAP17*, *STE18* and *RPL46*) fall into this category as well, making the border between ORF and gene even more elusive. Taking into account the physical position and ATG environment may help tell which ORFs are genes.

Comparison of the nucleotide sequence and of the predicted protein products with public database entries reveals that 118 ORFs (31%) correspond to genes previously identified in *S.cerevisiae*. All other ORFs represent

Table II. List of ORFs longer than 99 sense codons, known genes and other genetic elements of chromosome X

Nomenclature		Size (aa)	Coordinates		Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
Working	Official								
			1	60				left telomere sequence (complement TG_{1-3})	
			61	6931				Y' element	
J0202	YJL225c	1504	469	6130		0.11		probable nucleotide-binding protein, TMM 1+1 (intron from 4582 to 4969)	E
			6998	7224				copy of part of bi4 intron from cytochrome *b* gene (mitochondrial DNA)	
			7305	7767				core X element	
J0208	YJL223c	120	8779	9138		0.65	534 (538)	similar to PAU1 protein (PIR: S48516)	B
J0213	YJL222w	1549	11475	16121		0.16	5326 (7778)	similar to carboxypeptidase Y-sorting protein PEP1 (PIR: S25329), TMM 3+1	B
J0218	YJL221c	589	16770	18536		0.25	2459 (3094)	similar to α-glucosidase MAL35 (PIR: S46183), TMM 1+0	B
J0220	YJL220w	150	18243	18692		0.10		hypothetical protein, TMM 2+1	E
J0222	YJL219w	567	19974	21197		0.17	2913 (2955)	similar to hexose transport protein LGT3 (PIR: 45153), TMM 8+1	B
J0224	YJL218w	196	21973	22560		0.11	453 (943)	similar to galactoside *O*-acetyltransferase (SW: P07464), TMM 1+0	C
J0226	YJL217w	198	23133	23726		0.12		hypothetical protein	F
J0228	YJL216c	581	24344	26086		0.23	2229 (3095)	similar to α-glucosidase (PIR: S45157), TMM 1+0	C
J0231	YJL215c	119	26415	26771		0.10		hypothetical protein, ?	F
J0232	YJL214w	569	26887	28593		0.20	2953 (3021)	probable hexose transport protein HXT6 (PIR: S45159), TMM 11+1	B
J0234	YJL213w	331	32163	33155		0.14		hypothetical protein	F
J0236	YJL212c	799	33853	36249		0.18	1610 (4357)	similar to *S.pombe* ISP4 (PIR: S45161), TMM 10+1	D
J0238	YJL211c	147	36760	37200		0.10		hypothetical protein, ?	F
J0240	YJL210w	271	36919	37731	*CRT1*	0.09		CRT1 protein (PIR: S27422)	A
J0242	YJL209w	654	38005	39966	*CBP1*	0.15		CBP1 protein (PIR: S05829)	A
J0310	YJL208c	329	40197	41183	*NUC1*	0.14		nuclease NUC1 precursor, mitochondrial (PIR: S05888)	A
J0312	YJL207c	2014	41392	47433		0.14		hypothetical protein, TMM 4+1	E
J0316	YJL206c	758	47662	49935		0.15		hypothetical protein, TMM 1+1	E
J0318	YJL205c	187	50632	51192		0.14		hypothetical protein	F
J0320	YJL204c	645	51216	53150		0.16		hypothetical protein	F
J0322	YJL203w	280	53340	54179	*SPP91*	0.14		pre-mRNA splicing factor SPP91 (PIR: S23553)	A
J0323	YJL202c	115	53945	54289		0.12		hypothetical protein, TMM 1+1	E
J0325	YJL201w	599	54378	56174		0.15		hypothetical protein	F
J0327	YJL200c	789	56446	58812		0.22	2130 (3762)	similar to mitochondrial aconitate hydratase (GB: U17709)	C
J0330			59099	59171				$tRNA^{Thr}$	
J0332			59471	59782				δ remnant	
J0334	YJL199c	108	59857	60180		0.09		hypothetical protein, ?	F
J0336	YJL198w	881	60842	63484		0.18	2799 (4318)	similar to YCR037c (PIR: S46633), TMM 13+1	C
J0340	YJL197w	1254	63803	67564		0.14	535 (6137)	probable ubiquitin-carboxyl terminal hydrolase (SW: P35123)	D
J0343	YJL196c	310	67851	68780		0.13	924 (1753)	similar to sterol isomerase SUR4 (PIR: S46638), TMM 5+0	C
J0345	YJL195c	233	69242	69940		0.11		hypothetical protein, TMM 2+0	F
J0347	YJL194w	513	69336	70874	*CDC6*	0.13		cell division control protein CDC6 (PIR: S46640)	A
J0349	YJL193w	402	71364	72569		0.10	447 (2131)	similar to SLY41 protein (PIR: S46641), TMM 6+1	D
J0351	YJL192c	234	72711	73412		0.16		hypothetical protein, TMM 2+0	E
J0353	YJL191w	138i	73785	74606	*CRY2*	0.59		ribosomal protein S14eB (intron from 73795 to 74202) (PIR: S46643)	A
J0355	YJL190c	130	74911	75300	*RPS24*	0.81		ribosomal protein S15aE (PIR: A23082)	A
J0360	YJL189w	51	75931	76469	*RPL46*	0.92		ribosomal protein L39e (intron from 75937 to 76322) (EMBL: X01963)	B
J0403	YKL188c	102	76203	76508		0.15		hypothetical protein	F
J0406	YJL187c	819	76804	79260	*SWE1*	0.13		protein kinase SWE1 (PIR: S40400), TMM 1+0	A
J0409	YJL186w	586	80152	81909		0.16	1039 (3004)	similar to TTP1 protein (PIR: S45870), TMM 2+0	C
J0415	YJL185c	293	82095	82973		0.11		hypothetical protein	F
J0420	YJL184w	123	83445	83813		0.08		hypothetical protein, ?	F
J0425	YJL183w	422	84065	85330		0.18		hypothetical protein, TMM 1+0	E
J0430	YJL182c	105	85435	85749		0.08		hypothetical protein, TMM 1+0, ?	E
J0435	YJL181w	611	85657	87489		0.11	443 (2950)	hypothetical protein, similar to J1575, TMM 1+1	F
J0486	YJL180c	325	87583	88557	*ATP12*	0.12		ATP12 protein precursor (PIR: A39736)	A
J0488	YJL179w	109	88784	89110		0.15		hypothetical protein	F
J0490	YJL178c	196	89282	89869		0.17		hypothetical protein, TMM 1+0	E
J0493	YJL177w	184	90782	91651		0.68	825 (827)	ribosomal protein L17e (intron from 91091 to 91407) (PIR: S38012)	B
J0495	YJL176c	825	92052	94526	*SWI3*	0.15		transcription factor SWI3 (PIR: S26706)	A
J0502	YJL175w	170	94045	94554		0.12		hypothetical protein, TMM 3+0	E
J0504	YJL174w	276	95088	95915	*KRE9*	0.16		secretory pathway protein KRE9 precursor (PIR: S23891), TMM 1+0	A
J0506	YJL173c	122	96160	96525	*RFA3*	0.14		replication factor A chain 3 (PIR: C37281)	A
J0510	YJL172w	411	97729	99456	*CPS1*			Gly-X carboxypeptidase precursor (PIR: S16693)	A
J0512	YJL171c	396	99699	100886		0.22	478 (1923)	hypothetical protein, similar to YBR162C (PIR: S46033), TMM 2+0	D
J0514	YJL170c	183	101145	101693		0.13		hypothetical protein, TMM 2+0	E
J0517	YJL169w	122	102090	102455		0.15		hypothetical protein, TMM 2+0	E
J0520	YJL168c	733	102221	104419		0.14	258 (3593)	similar to trithorax ALL-1 zinc finger motif (PIR: A44264)	D
J0525	YJL167w	282	105005	106060	*FPP1*			farnesyl-pyrophosphate synthetase (SW: A34441), TMM 1+1	A
J0526	YJL166w	94	106425	106706		0.21	QCR8	ubiquinol-cytochrome *c* reductase subunit VIII (PIR: S48138)	
J0531	YJL165c	855	106888	109452	*HAL5*	0.13		HAL5 protein (PIR: S48240)	A
J0541	YJL164c	397	109960	111150	*SRA3*	0.18		protein kinase, cAMP-dependent, catalytic chain 1 (PIR: A27070)	A
J0544	YJL163c	555	111662	113326		0.08		hypothetical protein, TMM 11+1	E
J0549	YJL162c	482	114177	115622		0.14		hypothetical protein	F

Table II. *Continued*

Nomenclature		Size (aa)	Coordinates	Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
Working	Official							
J0550			115932 116003				$tRNA^{Glu}$	
J0552	YJL161w	180	117238 117777		0.09		hypothetical protein, TMM 1+1	E
J0555	YJL160c	180	118280 118819		0.15	326 (751)	similar to PIR1 protein (chr XI) (PIR: S33650)	C
J0558	YJL159w	310	120443 121372		0.47	577 (1162)	similar to PIR2 protein (chr XI) (PIR: S33651)	C
J0561	YJL158c	227	121964 122644		0.59	521 (976)	similar to PIR2 protein (chr XI) (PIR: S33651)	C
J0565	YJL157c	830	123535 126024	*FAR1*	0.13		factor arrest protein FAR1 (SW: S13341)	A
J0570	YJL156c	687	126589 128649		0.13		hypothetical protein, TMM 1+1	E
J0575	YJL155c	452	128985 130340	*FBP26*	0.14		fructose-2,6-bisphosphate 2-phosphatase (PIR: A42569)	A
J0580	YJL154c	944	130801 133632	*VPS35*	0.15		vacuolar protein-sorting protein VPS35 (PIR: S31293)	A
J0610	YJL153c	555	134032 135696	*INO1*	0.18		myo-inositol-1-phosphate synthase (PIR: A30902), TMM 2+1	A
J0628	YJL152w	119	135871 136227		0.07		hypothetical protein, TMM 1+0, ?	E
J0630	YJL151c	133	136072 136470		0.16		hypothetical protein, TMM 2+0	E
J0632	YJL150w	100	136820 137119		0.09		hypothetical protein, TMM 1+0, ?	E
J0634	YJL149w	663	137076 139064		0.16	296 (3276)	hypothetical protein, similar to YD9302.06c (GB: S51858), TMM 1+0	E
J0635			139458 139647	*SNR190*			SnR 190 small nuclear RNA	
J0636			139263 140390	*SNR128*			SnR 128 small nuclear RNA	
J0637	YJL148w	233	140134 140832		0.20		hypothetical protein	F
J0639	YJL147c	382	141119 142264		0.13		hypothetical protein	F
J0642	YJL146w	469	142989 144395		0.11		hypothetical protein, TMM 1+0	E
J0644	YJL145w	294	144857 145738		0.22		hypothetical protein	F
J0646	YJL144w	104	146056 146367		0.07		hypothetical protein, ?	F
J0648	YJL143w	158	146798 147271	*MIM17*	0.18		mitochondrial inner membrane protein MIM17 (PIR: S46257), TMM 1+1	A
J0650	YJL142c	130	147519 147908		0.06		hypothetical protein, TMM 3+1, ?	E
J0652	YJL141c	807	147667 150087	*YAK1*	0.12		protein kinase YAK1 (PIR: A32582), TMM 1+0	A
J0654	YJL140w	221	150658 151320	*RPB4*	0.14		DNA-directed RNA polymerase II chain RPB4 (PIR: A32490)	A
J0657	YJL139c	428	151413 152696	*YUR1*	0.14		YUR1 protein (PIR: S26856), TMM 1+0	A
J0660	YJL138c	395	153204 154388	*TIF2*	0.75		translation initiation factor eIF-4A(GB: X12814)	A
J0663	YJL137c	380	154685 155824		0.14	445 (1978)	hypothetical protein, similar to YKR058w (PIR: S38134)	D
J0664	YJL136c	87	156247 156970		0.60		ribosomal protein S21e (intron from 156487 to 156946)	B
J0666	YJL135w	105	157574 157888		0.14		hypothetical protein	F
J0671	YJL134w	409	157885 159111		0.11	1298 (2332)	hypothetical protein, similar to YKR053c (PIR: S38127), TMM 4+1	E
J0675	YJL133w	314	160316 161257	*MRS3*	0.08		splicing protein MRS3, mitochondrial (PIR: S01267)	A
J0678	YJL132w	750	161611 163860		0.12		hypothetical protein, TMM 1+1	E
J0682	YJL131c	356	163978 165045		0.12		hypothetical protein	F
J0686	YJL130c	2214	165423 172064	*URA2*	0.29		pyrimidine synthesis protein URA2 (PIR: S05767), TMM 1+1	A
J0689	YJL130Ac	115	171926 172929		0.06		hypothetical protein, (intron from 172082 to 172740), ?	F
J0693	YJL129c	1235	173299 177003	*TRK1*	0.14		potassium transport protein, high-affinity (PIR: S05849), TMM 8+1	A
J0699	YJL128c	668	177797 179800	*PBS2*	0.14		polymyxin B resistance protein kinase (PIR: A32714)	A
J0702	YJL127c	640	181999 183918	*SPT10*	0.12		regulatory protein SPT10 (PIR: S47865)	A
J0706	YJL126w	307	184199 185119		0.12	309 (1519)	hypothetical protein, similar to L9638.5 (GB: U19102)	F
J0710	YJL125c	383	185229 186377		0.14		hypothetical protein	F
J0714	YJL124c	172	186828 187343		0.16		hypothetical protein	F
J0718	YJL123c	478	187706 189139		0.15		hypothetical protein	F
J0723	YJL122w	175	189415 189939		0.21		hypothetical protein	F
J0731	YJL121c	238	190076 190789	*RPE1*	0.30		ribulose-5-phosphate 3-epimerase (GB: 83571)	A
J0734	YJL120w	107	190721 191041		0.14		hypothetical protein, TMM 1+1	E
J0738	YJL119c	107	191274 191594		0.13		hypothetical protein, TMM 1+0	E
J0742	YJL118w	219	191338 191994		0.09		hypothetical protein, TMM 1+1	E
J0744	YJL117w	311	192230 193162		0.19		hypothetical protein, TMM 2+0	E
J0748	YJL116c	337	193562 194572		0.25	1091 (1566)	hypothetical protein, similar to YKR042w (PIR: S38114), TMM 1+0	E
J0755	YJL115w	279	195985 196821	*ASF1*	0.14		ASF1 protein (PIR: S30766), TMM 1+1	A
J0760			197011 197083				$tRNA^{Ala}$	
J0765			197193 197242				δ remnant	
J0770			197243 197613				solo τ, LTR of Ty4	
J0775		414	197613 198854	*Ty4A_JL*	0.17		Ty4A_JL protein	
J0780		1803	197613 203022	*Ty4B_JL*	0.15		Ty4B_JL protein	
J0785			203098 203468				solo τ, LTR of Ty4	
J0790			203503 203814				δ remnant	
J0795			203815 204092				δ remnant	
J0799			204431 204502				$tRNA^{Asp}$	
J0802	YJL112w	714	205001 207142		0.12	229 (3303)	probable G-protein, β-transducin type (PIR: B48088)	D
J0804	YJL111w	550	207573 209222		0.19	1754 (2527)	probable chaperonin of the TCP-1 ring complex, similar to mouse CCT7 (PIR: S43058)	C
J0806	YJL110c	551	209621 211273	*GZF3*	0.10	274 (2405)	GATA zinc finger protein 3 (GB: X86353)	B
J0808	YJL109c	1769	211699 217005		0.17		hypothetical protein, TMM 5+1	E
J0811	YJL108c	383	217404 218552		0.17		hypothetical protein, TMM 8+1	E
J0813	YJL107c	387	218552 219712		0.13		hypothetical protein	F
J0817	YJL106w	645	221086 223020	*SME1*	0.15		probable protein kinase SME1 (PIR: S20138), TMM 1+0	A
J0819	YJL105w	560	224751 226430		0.10	586 (2734)	hypothetical protein, similar to YKR029c (PIR: S38101), TMM 1+0	E
J0822	YJL104w	149	227023 227469		0.09		hypothetical protein, ?	F

Table II. *Continued*

Nomenclature		Size (aa)	Coordinates	Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
Working	Official							
J0823			228122 228297	*SNR37*			SnR 37 small nuclear RNA	
J0824	YJL103c	618	228724 230577		0.12	253 (2980)	probable haem dependent regulatory protein, similar to S46116	D
J0826	YJL102w	819	230997 233453	*MEF2*	0.13		translation elongation factor G homologue, MEF2, mitochondrial (PIR: S43748), TMM 1+1	A
J0829			233635 233707				$tRNA^{Arg}$	
J0832	YJL101c	678	234019 236052	*GSH1*	0.14		glutamate-cysteine ligase (PIR: S28648), TMM 2+1	A
J0834	YJL100w	607	236959 238779		0.11		hypothetical protein	F
J0838	YJL099w	746	239110 241347	*CSD3*	0.12		CSD3 protein (GB: U15603)	A
J0840	YJL098w	1058	241778 244951		0.15	1625 (4985)	hypothetical protein, similar to YKR028w (GB: X85021)	F
J0902	YJL097w	217	245287 245937		0.18		hypothetical protein, TMM 6+0	E
J0904	YJL096w	224	245997 246668		0.13		hypothetical protein, TMM 2+0	E
J0906	YJL095w	1478	246950 251383	*BCK1*	0.12		protein kinase BCK1 (PIR: S20117)	A
J0909	YJL094c	873	251519 254137		0.13	264 (4290)	probable transport protein, similar to PIR: A42111, TMM 13+0	E
J0911	YJL093c	691	254435 256507	*TOK1*	0.12		TOK1, outwardly rectifying potassium channel protein, TMM 10+0 F	
J0913	YJL092w	1174	257118 260639	*RADH1*	0.13		helicase RADH (PIR: S46586)	A
J0916	YJL091c	498	260778 262271		0.13		hypothetical protein, TMM 8+1	E
J0918	YJL090c	764	262455 264746		0.14		hypothetical protein	F
J0922	YJL089w	829	265621 268107	*SIP4*	0.14		SIP4 protein, probable regulatory protein (GB: U17643), TMM 2+1	A
J0924	YJL088w	440	268188 269507	*ARG3*	0.16		ornithine carbamoyltransferase (PIR: S00058), TMM 1+1	A
J0927	YJL087c	827	269700 272180	*TRL1*	0.16		tRNA ligase (PIR: A29917), TMM 1+0	A
J0930	YJL086c	122	272176 272541		0.11		hypothetical protein, TMM 1+0	E
J0032	YJL085w	623	272522 274390		0.16		hypothetical protein	F
J0934	YJL084c	1046	274560 277697		0.13	1555 (4683)	hypothetical protein, similar to YKR021W (PIR: S38090)	F
J1002	YJL083w	604	278536 280347		0.09	596 (2822)	hypothetical protein, similar to YKR019c (PIR: S38088)	F
J1007	YJL082w	731	280880 283072		0.17	2652 (3586)	hypothetical protein, similar to YKR018c (PIR: S38087), TMM 1+1	E
J1012	YJL081c	489	283500 284966	*ACT3*	0.13		actin-related protein (PIR: S47608)	A
J1017	YJL080c	1222	285256 288921	*SCP160*	0.33		SCP160 protein, histone-like protein (PIR: S37492)	A
J1022	YJL079c	299	289573 290469		0.30	670 (1268)	hypothetical protein, similar to YKR013W (PIR: S38082), TMM 1+0	C
J1027	YJL078c	881	291034 293676		0.15	597 (3322)	hypothetical protein, similar to YKR013W (PIR: S38082), TMM 2+0	D
J1033	YJL077c	131	294364 294756		0.08		hypothetical protein, TMM 1+1, ?	E
J1038	YJL076w	1189	294940 298506		0.15	345 (4906)	putative protein-binding protein, similar to YKR010c (PIR: S25814)	D
J1044	YJL075c	138	298158 298571		0.11		hypothetical protein, TMM 1+0	E
J1049	YJL074c	1230	298855 302544		0.18	605 (5561)	probable purine nucleotide-binding protein, similar to SMC1 (PIR: S41804), TMM 1+0	D
J1083	YJL073w	692	302735 304810		0.14		hypothetical protein, TMM 1+1	E
J1086	YJL072c	213	304919 305557		0.12		hypothetical protein, TMM 1+0	E
J1091	YJL071w	574	305827 307548		0.12	314 (2803)	similar to acetyl-glutamate synthase (GB: L35484), TMM 1+1	D
J1095	YJL070c	888	307669 310332		0.14	441 (4614)	hypothetical protein, similar to YBR284w (PIR: S47120), TMM 1+1	E
J1098	YJL069c	594	310620 312401		0.17		hypothetical protein	F
J1102	YJL068c	299	312714 313610		0.20	525 (1572)	similar to human esterase D (SW: P10768)	D
J1107	YJL067w	116	313779 314126		0.12		hypothetical protein, TMM 1+1	E
J1111	YJL066c	252	313812 314567		0.16		hypothetical protein	F
J1115	YJL065c	167	314752 315252		0.11		hypothetical protein	F
J1120	YJL064w	131	314870 315262		0.12		hypothetical protein, TMM 1+1	E
J1125	YJL063c	238	315457 316170	*MRPL8*	0.09		ribosomal protein L17, mitochondrial (PIR: S47128)	A
J1132	YJL062w	830	316979 319468		0.12		hypothetical protein, TMM 9+1	E
J1135	YJL061w	713	319711 321849		0.16		hypothetical protein	F
J1138	YJL060w	444	323081 324412		0.21	662 (2193)	probable amino acid transferase, similar to (PIR: S52790)	D
J1139	YJL059w	408	324659 325882		0.12		hypothetical protein, TMM 6+1	E
J1141	YJL058c	543	325940 327568		0.12	1119 (2465)	purine nucleotide binding protein, similar to YBR270c (PIR: S46151), TMM 1+0	C
J1143	YJL057c	667	327816 329816		0.14		hypothetical protein, TMM 1+1	E
J1145	YJL056c	880	330129 332768		0.16	436 (4257)	probable regulatory protein, similar to mouse Kr2 protein (PIR: S00549), leucine zipper D	
J1148	YJL055w	245	333052 333786		0.14		hypothetical protein	F
J1150	YJL054w	478	333960 335393		0.15		hypothetical protein	F
J1152	YJL053w	379	335593 336729	*PEP8*	0.14		PEP8 protein (PIR: S48882)	A
J1154	YJL052w	332	337966 338961	*TDH1*	0.86		glyceraldehyde-3-phosphate dehydrogenase 3 (PIR: A00372), TMM 1+1	A
J1156	YJL051w	822	339482 341947		0.12		hypothetical protein, TMM 3+0	E
J1158	YJL050w	1073	342217 345435		0.20	971 (5214)	viral mRNA translation inhibitors SK12 (GB: D29641)	D
J1162	YJL049w	450	345668 347017		0.16		hypothetical protein	F
J1164	YJL048c	396	347145 348332		0.14	344 (1921)	hypothetical protein, similar to YBR273c (PIR: S46154)	F
J1166	YJL047c	842	349278 351803		0.12		hypothetical protein	F
J1171	YJL046w	451	351955 353307		0.12	302 (2257)	similar to lipoate–protein ligase A *E.coli* (PIR: A54035)	D
J1173			353939 354027				$tRNA^{Tyr}$ (small intron)	
J1177			354233 354555				solo δ	
J1179			354539 354870				solo δ	
J1185			355069 355140				$tRNA^{Arg}$	
J1190			355151 355222				$tRNA^{Asp}$	
J1194	YJL045w	634	355719 357620		0.16	2721 (3048)	similar to succinate dehydrogenase flavoprotein (PIR: S34793)	B
J1202	YJL044c	458	357998 359371	*GYP6*	0.16		GTPase-activating protein GYP6 (PIR: S30061), TMM 1+0	A

Table II. *Continued*

Nomenclature Working	Official	Size (aa)	Coordinates		Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
J1204	YJL043w	257	359825	360595		0.09		hypothetical protein	F
J1206	YJL042w	1398	360944	365137	*MIP1*	0.15		microtubule-associated protein (GB: X84652)	B
J1207	YJL041w	823	365479	368065	*NSP1*	0.16		nucleoskeletal-like protein NSP1 (PIR: S14055) (intron from 365480 to 365597)	B
J1216	YJL039c	1683	368446	373494		0.15		hypothetical protein, TMM 4+1	E
J1221			374119	374190				$tRNA^{Asp}$	
J1226			374201	374272				$tRNA^{Arg}$	
J1230			374539	374630				solo δ	
J1232	YJL038c	219	374813	375469		0.10	405 (1049)	similar to J1234, TMM 3+0	E
J1234	YJL037w	224	376357	377028		0.11	405 (1049)	similar to J1232, TMM 2+1	E
1240			378055	378128				$tRNA^{Val}$	
J1244	YJL036w	423	378520	379788		0.15		hypothetical protein	F
J1246	YJL035c	250	379947	380696		0.12		hypothetical protein	F
J1248	YJL034w	682	381022	383067	*KAR2*	0.44		nuclear fusion protein KAR2 precursor (PIR: A32366), TMM 1+1	A
J1250	YJL033w	770	383532	385841		0.20	530 (3629)	similar to *E.coli* SrmB RNA helicase (SW: P21507)	D
J1252	YJL032w	104	386043	386354		0.15		hypothetical protein	F
J1254	YJL031c	290	386066	386935	*BET4*	0.15		geranylgeranyl transferase α chain (PIR: S48301)	A
J1256	YJL030w	196	387352	387939	*MAD2*	0.12		MAD2 protein (PIR: S48302)	A
J1258	YJL029c	822	388083	390548		0.13	317 (4044)	similar to *C.elegans* T05G5.8 protein (PIR: S41008)	F
J1263			390738	390810				$tRNA^{Met}$	
J1267	YJL028w	111	391006	391338		0.07		hypothetical protein, TMM 2+0, ?	E
J1269	YJL027c	138	391531	391944		0.08		hypothetical protein, ?	F
J1271	YJL026w	399	392099	393295	*RNR2*	0.50		ribonucleoside-diphosphate reductase small chain (PIR: A26916), TMM 1+1	A
J1273	YJL025w	514	393662	395203	*RRN7*	0.13		RRN7 protein (PIR: S50785)	A
J1274	YJL024c	194	395623	396287		0.14	229 (920)	related to mouse clathrin associated protein 19 (intron from 396189 to 396265) (PIR: A40535)	D
J1278			396421	396491				$tRNA^{Gly}$	
J1282	YJL023c	347	397053	398093		0.13		hypothetical protein	F
J1284	YJL022w	102	397804	398109		0.10		hypothetical protein, TMM 1+1, ?	E
J1286	YJL021c	365	398635	399729		0.13		hypothetical protein	F
J1305	YJL020c	771	399789	402101		0.14	206 (3404)	glutamic acid rich protein precursor (*Plasmodium falciparum*) (PIR: A54514)	D
J1310	YJL019w	620	402588	404447		0.12		hypothetical protein, TMM 1+0	E
J1315	YJL018w	104	404321	404632		0.16		hypothetical protein	F
J1320	YJL017w	325	405278	406252		0.13		hypothetical protein	F
J1326	YJL016w	171	406447	406959		0.16		hypothetical protein	F
J1331	YJL015c	124	406834	407205		0.12		hypothetical protein	F
J1336	YJL014w	534	407246	408847	*BIN2*	0.23		chaperonin of the TCP-1 ring complex, TMM 1+1, similar to mouse CCT3 (PIR: S43062)	B
J1341	YJL013c	515	409184	410728		0.13	475 (2454)	similar to protein kinase BUB1 (Yeast chr 7) (GB: LM32027)	D
J1345	YJL012c	648	411143	413086		0.25		hypothetical protein	F
J1349	YJL011c	161	413975	414457		0.12		hypothetical protein	F
J1352			414653	414725				$tRNA^{Lys}$	
J1355			415618	415724				$tRNA^{Trp}$ (small intron)	
J1357	YJL010c	666	417252	419249		0.17		hypothetical protein	F
J1369	YJL009w	108	419542	419865		0.16		hypothetical protein, TMM 1+1	E
J1374	YJL008c	568	419647	421350		0.20	1219 (2622)	probable chaperonin of the TCP-1 ring complex, similar to mouse CCT8 (PIR: S52867)	C
J1379	YJL007c	104	422388	422699		0.13		hypothetical protein, TMM 1+0	E
J1385			422624	422696				$tRNA^{Met}$	
J1390	YJL006c	323	422828	423796		0.11		hypothetical protein, TMM 1+0	E
J1395			424119	424202				$tRNA^{Leu}$	
J1401	YJL005w	2026	424844	430921	*CYR1*	0.12		adenylate cyclase (PIR: A24776)	A
J1402	YJL004c	203	431279	431887		0.09		hypothetical protein, TMM 4+0	E
J1403	YJL003w	118	432331	432684		0.10		hypothetical protein, TMM 1+0, ?	E
J1404	YJL002c	476	432911	434338	*OST1*	0.16		α subunit, oligosaccharyltransferase (GB: Z46719), TMM 2+0	A
J1407	YJL001w	193	435032	435610	*PRE3*	0.17		multicatalytic endopeptidase complex chain PRE3 (PIR: S43669), TMM 1+0	A
			435996	436018	*CDEIII*			centromere	
			436022	436104	*CDEII*			centromere	
			436105	436112	*CDEI*			centromere	
J1409	YJR001w	602	436489	438294		0.12	257 (2951)	similar to *C.elegans*, hypothetical protein (PIR: S42372), TMM 10+1	E
J1411	YJR002w	593	438551	440329		0.17		hypothetical protein	F
J1415	YJR003c	539	440683	442399		0.13		hypothetical protein	F
J1418	YJR004c	650	442598	444547	*AGAL1*	0.13		α-agglutinin (PIR: S22835), TMM 2+0	A
J1422	YJR005w		445609	447708	*YAP80*			clathrin-associated protein complex β chain homolog (PIR: S12934), TMM 1+1	A
J1427	YJR006w	487	448888	450348		0.16		hypothetical protein	F
J1429	YJR007w	304	450706	451617	*SUI2*	0.37		translation initiation factor eIF-2 α chain (PIR: A32108)	A
J1431	YJR008w	338	452116	453129		0.14		hypothetical protein	F
J1433	YJR009c	332	453372	454367	*TDH2*	0.90		glyceraldehyde-3-phosphate dehydrogenase (PIR: S40915)	A
J1436	YJR010w	511	455925	457457	*MET3*	0.29		sulfate adenylyltransferase (PIR: S00906)	A
J1438	YJR011c	261	458330	459112		0.14		hypothetical protein	F
J1440	YJR012c	207	459484	460104		0.12		hypothetical protein, TMM 1+0	E

Table II. *Continued*

Nomenclature: Working	Official	Size (aa)	Coordinates	Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
J1444	YJR013w	305	460363 461277		0.11		hypothetical protein, TMM 5+1	E
J1446	YJR014w	198	461516 462109		0.22		hypothetical protein	F
J1448	YJR015w	510	462408 463937		0.13	1380 (2637)	similar to SNG1 gene (yeast chr 7) (GB: X74920), TMM 5+1	C
J1450	YJR016c	585	464141 465895	*ILV3*	0.38		dihydroxy-acid dehydratase (PIR: S43744)	A
J1452	YJR017c	190	466211 466780	*ESS1*	0.12		ESS1 protein (PIR: S07867)	A
J1454	YJR018w	120	466473 466832		0.08		hypothetical protein, TMM 1+1, ?	E
J1456	YJR019c	349	466922 467968		0.11	222 (1776)	similar to *E.coli* acyl-CoA thioesterase	D
J1458	YJR020w	110	467688 468017		0.11		hypothetical protein, TMM 1+1	E
J1462	YJR021c	292	468310 469266	*MER2*	0.11		meiotic recombination protein MER2 (intron from 468871 to 468950) (PIR: A40271)	A
J1464	YJR022w	128	469414 469797		0.13		hypothetical protein	F
J1470	YJR023c	133	469494 469892		0.09		hypothetical transport protein, TMM 2+1, ?	E
J1545	YJR024c	244	469920 470651		0.12		hypothetical protein	F
J1550	YJR025c	177	470828 471358		0.17	313 (922)	similar to human 3-hydroxyanthranilate 3,4-dioxygenate (PIR: A54070)	D
J1553			472150 472487				δ, LTR of Ty1	
J1555		440	472447 473766		0.14	1990 (2005)	TyA protein	
J1560		1741	472447 477712		0.15	8241 (8276)	TyB protein	
J1563			477738 478071				δ, LTR of Ty1	
J1565		440	478031 479350		0.15	1991 (1997)	TyA protein	
J1570		1741	478031 483296		0.14	8251 (8277)	TyB protein	
J1573			483322 483659				δ, LTR of Ty1	
J1575	YJR030c	745	483649 485883		0.11	443 (3553)	hypothetical protein, similar to J0435	F
J1580	YJR031c	1408	486276 490499		0.13	3171 (6683)	hypothetical protein, similar to YEL022w (PIR: S24168), TMM 6+1	E
J1585	YJR032w	393	490768 491946		0.19	468 (1962)	hypothetical protein, similar to L8167.24 (PIR: S48567)	F
J1590	YJR033c	1357	492068 496138		0.14	3103 (6771)	hypothetical protein	F
J1604	YJR034w	108	496370 496693	*PET191*	0.12		PET191 protein (PIR: S28924)	A
J1606	YJR035w	1085	497042 500296	*RAD26*	0.13		probable helicase RAD26 (SW: P40352), TMM 1+1	A
J1608	YJR036c	892	500403 503078		0.11		hypothetical protein, TMM 1+1	E
J1610	YJR037w	127	502789 503169		0.11		hypothetical protein	F
J1612	YJR038c	120	503400 503759		0.09		hypothetical protein, TMM 2+0, ?	E
J1614	YJR039w	1121	503623 506985		0.13		hypothetical protein, TMM 2+1	E
J1616	YJR040w	779	507433 509769		0.14	788 (3956)	similar to mouse chloride channel protein (GB: D17521), TMM 7+1	D
J1622	YJR041c	1174	509929 513450		0.14		hypothetical protein, TMM 2+1	E
J1624	YJR042w	744	513742 515973		0.13		hypothetical protein, TMM 1+0	E
J1626	YJR043c	350	516151 517200		0.14		hypothetical protein	F
J1631			517500 517571				$tRNA^{Met}$	
J1634			517645 517786				δ remnant	
J1637	YJR044c	140	518453 518872		0.15		hypothetical protein, TMM 4+0	E
J1639	YJR045c	654	519328 521289	*SSC1*	0.52		heat shock protein 70-related protein SSC1 precursor, mitochondrial (PIR: A32493)	A
J1641	YJR046w	604	521735 523546		0.11		hypothetical protein, TMM 1+1	E
J1647			523699 523780				$tRNA^{Ser}$	
J1651	YJR047c	157	524598 525068	*ANB1*	0.70		translation initiation factor eIF-5A.2 (PIR: B40259)	A
J1653	YJR048w	109	526022 526348	*CYC1*	0.37		cytochrome *c* isoform 1	A
J1655	YJR049c	530	526574 528163	*UTR1*	0.13		UTR1 protein (PIR: S46589), TMM 1+1	A
J1657	YJR050w	235	528384 529088	*UTR3*	0.10		UTR3 protein (PIR: S46590)	A
J1659	YJR051w	501	529548 531050	*OSM1*	0.17		OSM1 protein precursor (PIR: S46591), TMM 1+0	A
J1661			531202 531361				δ remnant	
J1663			531515 531585				$tRNA^{Gly}$	
J1665	YJR052w	565	531749 533443	*RAD7*	0.14		RAD7 protein (PIR: A25226)	A
J1667	YJR053w	574	533714 535435		0.15		hypothetical protein	F
J1669	YJR054w	497	535743 537233		0.13	725 (2484)	hypothetical protein, similar to YM9827.05c (GB: Z47816), TMM 4+0	E
J1670			538242 538313				$tRNA^{Arg}$	
J1705	YJR055w	164	538459 538950	*HIT1*	0.13		HIT1 protein (PIR: S30869)	A
1706a			540453 540783				solo δ	
1706b			540786 541114				solo δ	
1707			541195 541266				$tRNA^{Asp}$	
J1710	YJR056c	236	541482 542289		0.10		hypothetical protein	F
J1713			542643 542731				$tRNA^{Tyr}$ (small intron)	
J1715	YJR057w	216	543749 544396	*CDC8*	0.15		dTMP kinase (PIR: A00683)	A
J1720	YJR058c	147	544422 544862	*YAP17*	0.08		clathrin-associated protein 17 (PIR: C40535)	A
J1725	YJR059w	818	545474 547927		0.16	1251 (3786)	similar to serine/threonine specific protein kinase (PIR: S38035), TMM 1+0	D
J1730	YJR060w	351	548446 549498	*CBF1*	0.14		centromere-binding protein CP1 (PIR: A36310)	A
J1736	YJR061w	935	550198 553002		0.13		hypothetical protein, TMM 1+1	E
J1742	YJR062c	457	553166 554536	*NTA1*	0.12		amino-terminal amidase NTA1 (PIR: S47938)	A
J1747	YJR063w	125	554882 555256	*RPA12*	0.20		DNA-directed RNA polymerase I chain A12.2 (PIR: A48107), TMM 1+0	A
J1752	YJR064w	562	555601 557286		0.22	1704 (2637)	probable chaperonin of the TCP-1 ring complex, similar to mouse CCT5 (PIR: S43061), TMM 1+0	C
J1760	YJR065c	449	557499 558845		0.20	1499 (2153)	similar to actin-like protein Act 2 (fission yeast) (PIR: A41790), TMM 1+1	C
J1803	YJR066w	2470	559103 566512	*TOR1*	0.14		TOR1 protein (PIR: S43940), TMM 3+1	A
J1805	YJR067c	141	566709 567131		0.14		hypothetical protein	F

Table II. *Continued*

Nomenclature		Size (aa)	Coordinates	Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
Working	Official							
J1808	YJR068w	353	567330 568388	*RFC2*	0.18		replication factor C chain RFC2 (PIR: S45531)	A
J1811	YJR069c	197	568496 569086		0.20		hypothetical protein	F
J1814	YJR070c	325	569311 570285		0.40		hypothetical protein	F
J1818	YJR071w	122	570092 570457		0.10		hypothetical protein, ?	F
J1821	YJR072c	385	570657 571811		0.17	847 (1816)	similar to *C.elegans* protein C34E10 (GB: U10402)	F
J1824	YJR073c	206	572005 572622	*PEM2*	0.17		methylene-fatty-acyl-phospholipid synthase (PIR: B28443), TMM 3+1	A
J1827	YJR074w	218	572782 573435		0.15		hypothetical protein	F
J1830	YJR075w	396	573668 574855		0.18	209 (2020)	similar to mannosyltransferase (PIR: S22701), TMM 2+0	D
J1833	YJR076c	415	575044 576288	*CDC11*	0.17		cell division control protein CDC11 (PIR: S40911)	A
J1837	YJR077c	311	576945 577877	*MIR1*	0.36		phosphate transport protein, mitochondrial (PIR: S12318), TMM 1+1	A
J1840	YJR078w	453	578547 579905		0.13	514 (2251)	similar to mouse indoleamine 2-3 dioxygenase (PIR: JH0492)	D
J1843	YJR080w	109	579892 580923		0.11		hypothetical protein (intron from 580035 to 580739), TMM 1+0	E
J1847	YJR081c	394	580122 581303		0.14		hypothetical protein	F
J1854	YJR082c	113	581604 581942		0.15		hypothetical protein	F
J1857	YJR083c	309	582298 583224		0.11		hypothetical protein	F
J1860	YJR084w	423	583420 584688		0.10		hypothetical protein	F
J1863	YJR085c	105	584810 585124		0.14		hypothetical protein, TMM 2+0	E
J1866	YJR086w	110	585755 586084	*STE18*	0.10		STE18 protein (PIR: B30102)	A
J1870	YJR087w	116	586087 586434		0.10		hypothetical protein, TMM 2+0, ?	E
J1875	YJR088c	292	586185 587060		0.17		hypothetical protein	F
J1880	YJR089w	954	587405 590266		0.13		hypothetical protein	F
J1885	YJR090c	1151	590562 594014	*GRR1*	0.12		GRR1 protein (PIR: A41529), TMM 1+1	A
J1890	YJR091c	1091	594751 598023		0.15	593 (4842)	hypothetical protein, similar to YP9499.01c (PIR: S54067)	F
J1901	YJR091Ac	200	597437 598035		0.15		ATP/GTG binding site motif A	E
J1905	YJR092w	1320	598809 602768		0.15		hypothetical protein	F
J1911	YJR093c	327	602916 603896	*FIP1*	0.12		component of pre-mRNA polyadenylation factor	B
J1916	YJR094c	360	604265 605344	*IME1*	0.18		meiosis-inducing protein IME1 (PIR: S31137)	A
J1921	YJR095w	322	609466 610431	*ACR1*	0.21		ACR1 protein (PIR: S43280), TMM 2+1	A
J1926	YJR096w	282	610888 611733		0.22	431 (1491)	probable reductase protein, similar to GB: A32950	D
J1931	YJR097w	172	612106 612621		0.13		hypothetical protein	F
J1936	YJR098c	655	612882 614846		0.15		hypothetical protein	F
J1941	YJR099w	236	615266 615973	*YUH1*	0.11		ubiquitin carboxyl-terminal hydrolase YUH1 (GB: S51332), TMM 1+0	A
J1946	YJR100c	327	616044 617024		0.10		hypothetical protein	F
J1950			617609 617709				$tRNA^{Leu}$ (small intron)	
J1952	YJR101w	266	617924 618721		0.11		hypothetical protein	F
J1957	YJR102c	202	618850 619455		0.13		hypothetical protein	F
J1962	YJR103w	564	620444 622135	*URA8*	0.16		CTP synthase URA8 (PIR: S42580), TMM 2+0	A
J1968	YJR104c	154	622242 622703	*SOD1*	0.38		superoxide dismutase (Cu-Zn) (PIR: A36171)	A
J1973	YJR105w	340	623270 624289		0.37		hypothetical protein	F
J1978	YJR106w	725	624527 626701		0.10		hypothetical protein	F
J1983	YJR107w	328	627030 628013		0.13		hypothetical protein, TMM 12+1	E
J1988	YJR108w	123	628403 628771		0.14		hypothetical protein	F
J2002	YJR109c	1118	629279 632632	*CPA2*	0.24		large subunit of arginine specific carbamoyl-phosphate synthase (PIR: A01199)	A
J2007	YJR110w	688	633306 635369	*CPA1*	0.16		small subunit of arginine specific carbamoyl-phosphate synthase (PIR: B33478)	A
J2009	YJR111c	283	635549 636397		0.12		hypothetical protein	F
J2011	YJR112w	201	636721 637323		0.09		hypothetical protein	F
J2020	YJR113c	247	637926 638666		0.10	204 (1185)	similar to ribosomal protein S7 (*Bacillus stearothermophilus*) (PIR: JG0008)	D
J2024	YJR114w	130	638350 638739		0.11		hypothetical protein, TMM 1+0	E
J2027	YJR115w	169	639633 640139		0.10		hypothetical protein	F
J2031	YJR116w	279	640516 641352		0.14		hypothetical protein, TMM 2+1	E
J2032	YJR117w	453	641698 643056		0.27		hypothetical protein, TMM 5+1	E
J2033	YJR118c	203	643184 643792		0.19		hypothetical protein, TMM 3+1	E
J2035	YJR119c	728	644998 646181		0.15	776 (3828)	similar to human retinoblastoma binding protein 2 (GB: S66431)	D
J2039	YJR120w	116	646817 647164		0.07		hypothetical protein, ?	F
J2041	YJR121w	511	647298 648830	*ATP2*	0.42		H^+-transporting ATP synthase β chain precursor (PIR: S27278)	A
J2043	YJR122w	497	649467 650957		0.15		hypothetical protein	F
J2045	YJR123w	125	651592 652266	*RPS5*	0.75		ribosomal protein S5	A
J2046	YJR124c	448	652586 653929		0.14		hypothetical protein, TMM 9+1	E
J2048	YJR125c	408	654431 655654		0.17	283 (1775)	hypothetical protein, similar to L8167.6 yeast protein (PIR: S48557)	F
J2050	YJR126c	811	655948 658388		0.13	521 (3981)	similar to human prostate-specific membrane antigen (SW: Q04609), TMM 1+0	D
J2052	YJR127c	1380	658611 662750	*ZMS1*	0.12		ZMS1 protein (PIR: S43751), TMM 4+1	A
J2059	YJR128w	119	662612 662968		0.06		hypothetical protein, ?	F
J2060			663440 663633	*SNR3*			SnR 3 small nuclear RNA	
J2061	YJR129c	339	663694 664710		0.11		hypothetical protein, TMM 1+0	E
J2063	YJR130c	639	664912 666828		0.13	1778 (3174)	similar to TUB1 3' region (GB: S49644)	C
J2110	YJR131w	549	667335 668981	*MNS1*	0.14		α-mannosidase MNS1 (PIR: A39345), TMM 1+0	A
J2112	YJR132w	1048	669213 672356		0.15		hypothetical protein, TMM 2+1	E
J2118	YJR133w	209	672682 673308		0.28		hypothetical protein	F
J2120	YJR134c	707	673423 675543		0.15	343 (3092)	similar to human TATA element modulatory factor (PIR: A47212)	D

Table II. *Continued*

Nomenclature		Size (aa)	Coordinates	Locus	CAI	FastA score	Description (nature of element, function or similarity of product)/Comment	
Working	Official							
J2122	YJR135c	239	675753 676469		0.12		hypothetical protein	F
J2124	YJR136c	421	677135 678397		0.10		hypothetical protein	F
J2126	YJR137c	1442	678651 682976		0.25	1054 (6897)	similar to ferredoxine sulfate reductase (SW: P30008)	D
J2129	YJR138w	1584	684258 689009		0.14		hypothetical protein	F
J2132	YJR139c	359	689139 690215	*HOM6*	0.47		homoserine dehydrogenase (PIR: S33317), TMM 1+1	A
J2161	YJR140c	1648	690444 695387		0.14		hypothetical protein, TMM 1+1	E
J2166	YJR141w	347	695597 696637		0.13		hypothetical protein, TMM1+1	E
J2171	YJR142w	342	696832 697857		0.15		hypothetical protein	F
J2176	YJR143c	762	698020 700305	*PMT4*	0.22		PMT4 protein (PIR: S51284), TMM 8+1	A
J2181	YJR144w	269	700573 701379	*MGM101*	0.16		MGM101 protein (PIR: S34849)	A
J2186	YJR145c	261	701721 702759	*RPS7A*	0.69		ribosomal protein S4ec10 (intron from 702490 to 702745) (PIR: S20054)	A
J2200	YJR146w	117	703576 703926		0.07		hypothetical protein, ?	F
J2204	YJR147w	358	703887 704960		0.12	235 (1782)	similar to heat shock transcription factor 8 (PIR: S25481)	D
J2209	YJR148w	376	705435 706562		0.19	1584 (1900)	similar to TWT1 yeast protein (PIR: S48989)	C
J2213	YJR149w	404	706851 708062		0.14	462 (1937)	similar to 2-nitropropane dioxygenase (PIR: S50891)	D
J2217	YJR150c	298	708505 709398		0.30		hypothetical protein, TMM 2+0	E
J2223	YJR151c	1161	711949 715431		0.23	614 (4382)	similar to human mucin (PIR: A49963), TMM 2+0	D
J2230	YJR152w	543	719357 720985	*DAL5*	0.16		allantoate permease (PIR: A28671), TMM 6+1	A
J2235	YJR153w	361	722506 723588		0.17	907 (1643)	similar to polygalacturonase (PIR: S28771), TMM 1+0	C
J2240	YJR154w	346	725475 726512		0.13		hypothetical protein	F
J2245	YJR155w	288	727036 727959		0.15	1334 (1439)	similar to yeast aryl-alcohol deshydrogenase (PIR: S51335)	B
J2250	YJR156c	340	728268 729287		0.53	1784 (1790)	similar to thiamine-repressed nmt-1 protein (PIR: S48548), TMM 1+0	B
J2255	YJR157w	120	730206 730565		0.13		hypothetical protein, TMM 1+0	F
J2260	YJR158w	567	732131 733831		0.16	1893 (3036)	similar to hexose transport protein HXT7 (PIR: S43186), TMM 9+1	C
J2395	YJR159w	357	735735 736805	*SOR1*	0.22		sorbitol dehydrogenase (GB: L11039)	B
J2400	YJR160c	602	737702 739507		0.13	2585 (4048)	similar to sugar transport protein (SW: P38156), TMM 7+1	C
J2410	YJR161c	383	742542 743690		0.14	1845 (2635)	similar to YB8L (SW: P38363), TMM 3+1	E
			744593 745052				core X element	
			745053 745356				STR-D, C, B and A elements	
J2420	YJR162c	116	744605 744952		0.14	422 (804)	similar to YKW5 (SW: P36030)	F
			745357 745442				right telomere sequence	

Last column: status of the protein deduced from each putative gene. The categories A (fully known) to F (unknown) are defined in the text. The self FastA score of the predicted protein is in parentheses. An accession number in one of the public databases [PIR, Swiss-Prot (SW),GenBank (GB) and EMBL] is indicated. Abbreviations: TMM: transmembrane motif, integral+ peripheral; ?: questionable gene. ORF YJL093c is categorized as F, as it was discovered and sequenced during the systematic sequencing of chromosome X and found to correspond to no known gene. It was subsequently biologically characterized as a potassium channel (Ketchum *et al.*, 1995).

novel putative yeast genes whose function will have to be determined experimentally. However, 57 of these (another 15% of total) encode proteins that show significant similarity to a protein of known function from yeast or other organisms, thus providing some indication as to their function. The 204 (54%) remaining ORFs exhibit no significant similarity to known sequences (FastA score <200). Motif searches have shown that 91 of the latter have some particular protein signature, mostly a structure suggestive of transmembrane domains (Table II).

An approximately equal number of ORFs is observed on each DNA strand. The mean ORF size is 482 codons (1446 bp), the longest (YJR066w) reaching 2470 codons. The mean size of inter-ORF regions, disregarding one in each pair of overlapping ORFs, is 602 bp for terminator–promoter combinations (WW and CC in Figure 3). For divergent promoters (DP) and convergent terminators (CT), the mean size is 725 bp and 311 bp, respectively. This striking difference in inter-ORF size between divergent promoters versus convergent terminators may be indicative of more important sequence requirements in promoter regions for the regulation of gene expression. An exception is the contiguity of the two ORFs YJL108c and YJL107c. The TGA stop codon of the latter overlaps the ATG of the former, so that both codons share TG. This peculiarity was carefully checked by oligo-primed sequencing in either direction on cosmid DNA. The two ORFs in their integrity are translated from a single transcript of ~3 kb (Rasmussen, 1995).

Environment of ATG and stop codons

Compilation of a large number of sequence data surrounding the initiation codon AUG has revealed that these sequences are not random and that higher eukaryotes have in common the consensus sequence GCC(A/G)CC**ATG**G (Kozak, 1987). In the case of the budding yeast, another consensus (A/Y)A(A/Y)A(A/Y)A**ATG**GTCT has been proposed (Hinnebusch and Liebman, 1991).

We examined the 318 chromosome X ORFs longer than 150 codons, in all probability corresponding to real genes, to test this consensus. Table III shows the frequency of the different nucleotides, as determined by tabulating positions –8 to +7 relative to ATG. A χ^2 test was performed at each position to test the non-randomness of this distribution, taking into account the G+C content of the chromosome. At all positions except –5 the distribution was found to be non-random. As these calculations are based on all the ORFs of a chromosome, regardless of their expression level, rather than on a selected subset, the following consensus sequence might be more appropriate: AAANAAAA**ATG**GCTG. The chances of a random distribution at each position is <5%, or even 1%

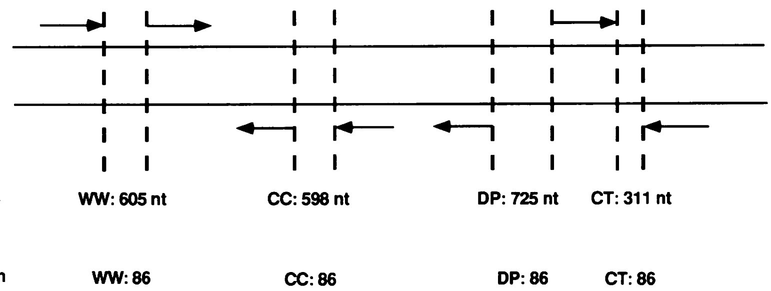


Fig. 3. Mean size and distribution of inter-ORF regions of chromosome X. WW: promoter/terminator combination on Watson strand; CC: promoter/terminator combination on Crick strand; DP: divergent promoters; CT: convergent terminators. The numbers indicate on top line the mean size, on bottom line the distribution of each configuration.

Table III. Initiation and stop codon environment

ATG environment

	−8	−7	−6	−5	−4	−3	−2	−1	ATG	+4	+5	+6	+7
A	0.396	0.393	0.368	0.349	0.399	0.569	0.403	0.456	ATG	0.318	0.283	0.324	0.327
G	0.164	0.160	0.211	0.135	0.148	0.195	0.119	0.145	ATG	0.296	0.129	0.151	0.299
C	0.173	0.192	0.176	0.220	0.189	0.113	0.252	0.173	ATG	0.132	0.362	0.182	0.129
T	0.267	0.255	0.245	0.296	0.264	0.123	0.226	0.223	ATG	0.254	0.343	0.343	0.242
χ^2	7.978	9.616	10.015	7.370	10.060	104.811	30.284	27.741	ATG	20.165	61.227	8.750	22.695

TAG stop codon environment

	−5	−4	−3	−2	−1	TAG	+4	+5	+6	+7	+8	+9
A	0.380	0.268	0.310	0.394	0.296	TAG	0.408	0.282	0.380	0.437	0.366	0.282
G	0.127	0.183	0.253	0.211	0.211	TAG	0.211	0.127	0.293	0.211	0.197	0.141
C	0.183	0.197	0.169	0.085	0.113	TAG	0.113	0.197	0.183	0.056	0.169	0.239
T	0.310	0.352	0.268	0.310	0.380	TAG	0.268	0.394	0.197	0.296	0.268	0.338
χ^2	2.975	2.127	1.173	5.599	5.024	TAG	4.336	2.651	5.580	9.178	1.250	2.522

TAA stop codon environment

	−5	−4	−3	−2	−1	TAA	+4	+5	+6	+7	+8	+9
A	0.368	0.296	0.387	0.452	0.361	TAA	0.297	0.316	0.368	0.355	0.297	0.393
G	0.161	0.226	0.232	0.097	0.142	TAA	0.187	0.136	0.174	0.122	0.161	0.142
C	0.200	0.239	0.129	0.155	0.181	TAA	0.129	0.200	0.148	0.168	0.271	0.155
T	0.271	0.239	0.252	0.296	0.316	TAA	0.387	0.348	0.310	0.355	0.271	0.310
χ^2	2.358	3.484	6.237	17.687	4.314	TAA	4.559	2.173	1.590	3.310	9.646	3.552

TGA stop codon environment

	−5	−4	−3	−2	−1	TGA	+4	+5	+6	+7	+8	+9
A	0.348	0.304	0.402	0.424	0.261	TGA	0.347	0.315	0.304	0.391	0.315	0.272
G	0.174	0.239	0.239	0.087	0.163	TGA	0.185	0.196	0.283	0.196	0.174	0.206
C	0.185	0.120	0.152	0.196	0.163	TGA	0.109	0.109	0.163	0.196	0.152	0.185
T	0.293	0.337	0.207	0.293	0.413	TGA	0.359	0.380	0.250	0.217	0.359	0.337
χ^2	0.626	4.244	4.900	9.008	7.980	TGA	2.966	3.641	7.964	4.773	0.720	1.494

The position relative to start or stop codon is indicated at the top of the column. The numbers in the columns give the relative frequency of each base at each position. χ^2 tests were performed with three degrees of freedom (threshold for an α risk of 5% is 7.815 and for an α risk of 1% is 11.345). Expected frequencies used in χ^2 tests are A = 0.32, T = 0.32, G = 0.17 and C = 0.17 in non-coding regions, A = 0.32, G = 0.20, C = 0.19 and T = 0.28 in coding regions. Tabulation performed on 318 ORFs >150 codons.

at positions −3, −2, −1, +4, +5 and +7. We then addressed the question of the possible existence of a consensus sequence in the environment of the stop codons. Not surprisingly, TAA is the more frequently used stop codon: 155 ORFs longer than 150 codons have it, while 92 have TGA and 71 TAG. When the nucleotide environment between positions −5 and +9 (position +1 being defined by the T of the stop signal) was tabulated, we observed the frequencies reported in Table III. It appears that, in the case of TAA, there is a bias at position −2, which is more frequently than expected occupied by A and less frequently by G, and at position +8, where C is increased. In the case of TAG, at position –2 the frequency of C is depressed, while this nucleotide is nearly always absent from position +7. Finally, in the case of TGA, the distribution deviates from randomness at three positions, –2, –1 and +6.

Small ORFs (< 100 codons)

The choice of a minimal length of 99 sense codons between the first ATG and the stop signal, which dates back to 1979 (Galibert *et al.*, 1979), probably owes more to the widely used decimal numbering system than to proper insight into biological mechanisms. However, as mentioned above, this size is warranted in the case of yeast (Dujon *et al.*, 1994). In simulation experiments in which chromosome length and nucleotide composition was varied, the chances that ORFs longer than 150 codons will exist and still not correspond to a real gene are negligible. Conversely, the chances that ORFs in the range 100–149 codons will have no biological significance increase in proportion to decreasing size. However, a size of 100 codons is no impassable limit and obviously some ORFs smaller than 100 codons correspond to genes and, for that matter, quite a few proteins shorter than 99 amino acids may not be accounted for by post-translational processing. An example is provided by the small proteolipids PMP1 and PMP2 (40 and 43 amino acids), on chromosomes III and V, respectively (Navarre *et al.*, 1992; 1994). Analysis of the chromosome X sequence has revealed 344 small ORFs 50–98 sense codons in size. Comparison of the deduced proteins with database entries shows that one of these, J0526 (106425–106706), corresponds to the gene encoding subunit VIII of ubiquinol-cytochrome *c* reductase (Hemrika *et al.*, 1993). It is a 94-amino acid protein whose coding gene has been hitherto overlooked. Another instance is YKR057w, which encodes a ribosomal protein of 87 amino acids. Some small ORFs, such as J1567 (479710–479952), J1564 (477910–478074) and J15591 (474126–474368) have perfect or nearly perfect matches with Ty retrotransposon proteins of longer size. These small ORFs most probably result from frameshift mutations, a rather common occurrence in these retroposons. Finally, significant similarity is observed between some small ORFs located in the subtelomeric region, such as J0210 (9452–9852), and similar elements located on other chromosomes (K-B110 on chromosome XI or I.A75 on chromosome IX). The other small ORFs, displaying no significant homology with database entries, cannot simply be discarded, since some probably correspond to real genes. Examples in point are J0523 (105893–106060), J1153 (337859–338143), J2123 (676661–676924) and J1425 (448166–448444), all with CAIs >0.2. Clearly, a screening programme taking into account parameters such as the ATG and stop codon environment and the CAI must be developed to approach the question of their existence as genes.

Sequence duplications

We have analysed the nucleotide sequence of chromosome X for the occurrence of sequences demonstrating high similarity to other genes of chromosome X (intrachromosomal duplications) and to genes in other yeast chromosomes (interchromosomal duplications), both at the nucleotide and the amino acid level (Table IV). Some of the duplicated ORFs have been functionally characterized. These results confirm earlier observations on chromosomes XI (Dujon *et al.*, 1994) and II (Feldmann *et al.*, 1994) of the high level of internal genetic redundancy in the yeast genome. Moreover, in addition to duplication of individual genes, duplication of syntenic segments has also occurred, syntenic in the present context of intraspecies duplications meaning that two or more genes situated closely on the same chromosome have their homologous loci also located close together, with the same respective orientation, on the other chromosome. As a rule, the physical distance and the nucleotide sequence between two ORFs on the same syntenic segment are not conserved. However, some degree of intergenic sequence conservation can be observed in a few cases, as exemplified in Figure 4.

tRNAs and transposons

Twelve *tRNA* genes are found on each strand (Figure 5), a density somewhat higher than that observed in the previously sequenced yeast chromosomes. The 24 *tRNAs* can transfer 13 amino acids in all and include four $tRNA^{Asp}$, all identical with the same GTC anticodon; four $tRNA^{Arg}$, two identical with TCT, one with ACG and one with CCT, the last two with minor sequence differences. Of the three $tRNA^{Met}$, two are identical while the third exhibits slight differences. The two $tRNA^{Tyr}$ have an identical sequence and include the same GTA anticodon.

Upon folding, all the predicted tRNAs fit in readily with the clover-leaf model, regarding stem length as well as loop size. All the canonical bases are observed in all cases but one. The exception is $tRNA^{Met}$ at position 517571, which exhibits an A, instead of T as in the canonical GTΨC sequence. Careful checking of the sequence has shown that this ATC sequence does not result from sequencing errors. However, a cloning artefact at some point in the construction of the cosmid library cannot be ruled out at this stage.

While the clover-leaf model is basically respected, 46 non-canonical or unpaired bases are observable in the stems of this two-dimensional configuration. Thirty-nine correspond to a GT base pairing, three to TT and CA and one to GG. An example of such tRNA folding is presented in Figure 6. These observations cannot be ascribed to sequencing or cloning incidents, since they have been observed by different investigators all working on different cosmids. Furthermore, the reality of such pairings has been established by direct RNA sequencing on mature tRNA and by mutagenesis experiments (Pütz *et al.*, 1993). However, it is also true that in the case of plant mitochondrial tRNAs, some (but not all) mismatched base pairs are so edited as to generate a Watson–Crick pair in the mature tRNA (Maréchal-Drouard *et al.*, 1993). While this phenomenon is not yet documented in nuclear yeast tRNA, the possibility of a similar editing process, whereby some of the 46 mispairings mentioned above would be converted into conventional Watson–Crick pairs, cannot be dismissed without additional sequence data or structural studies at the tRNA level. An alternative hypothesis is that some of the predicted tRNAs actually correspond to inactive pseudogenes.

Four of the tRNA genes encountered in chromosome

Table IV. Related genes from chromosome X

Gene/ORF on chromosome X	Related gene/ORF on other chromosome[a]	Functional description[b]	aa identity %[c]	nt identity %[d]
YJL223c	PAU1(5)	PAU1 protein	96.7 (1–120)/120	96.7 (1–360)/360
YJL210w	LGT3 hexose transport protein	97.9 (1–567)/567	98.4 (883–1701)/1701	
YJL200c	ACO1(12)	similar to aconitin hydratase	55.3 (35–782)/782	50.8 (6–2278)2367
YJL198w	YCR037c (3)	probable transport protein	65.0 (39–879)/881	68.1 (684–2387)/2643
YJL196c	YCR034w (3)'	similar to sterol isomerase SUR4	58.4 (16–310)/310	60.3 (70–891)/930
YJL191w (CRY2)	CRY1 (3)	ribosomal protein S14eB	96.3 (3–138)/138	92.0 (8–414)/414
YJL190c (RPS24)	L8039.6 (12)	ribosomal protein S15ae	99.2 (1–130)/130	89.1 (1–390)/390
YJL164c (SRA3)	TPK3 (11)	cAMP-dependent protein kinase	84.5 (69–397)/397	73.0 (255–486)/1191
YJL139c (YUR1)	KTR2 (11)	YUR1 protein	66.3 (37–424)/426	64.3 (269–1250)/1284
YJL138c (TIF2)	TIF1 (11)	translation initiation factor eIF-2	100 (1–395)/395	99.3 (1–1185)/1185
YJL133w (MRS3)	MRS4 (11)	mitochondrial splicing protein	76.2 (23–312)/314	70.5 (119–875)/942
YJL099w (CSD3)	YKR027w (11)	CSD3 protein	42.3 (1–844)/1058	37.3 (1759–2238)/2238
YJL098w	YKR028w (11)	unknown	45.8 (1–844)/1058	60.0 (164–1442)/3174
YJL084c	YKR021w (11)	unknown	37.6 (4–932)/1046	46.4 (7–1946)/3138
YJL083w	YKR019c (11)	unknown	26.7 (38–604)/604	64.6 (1265–1601)/1812
YJL082w	YKR018c (11)	unknown	66.0 (1–730)/731	53.7 (233–1986)/1986
YJL079c	YKR013w (11)	unknown	47.5 (1–299)/299	61.4 (415–789)/897
YJL078c	YKR013w (11)	unknown	67.3 (15–161)/881	39.0 (1295–1711)/2643
YJL076w	YKR010c (11)	unknown	16.1 (1–772)/1189	33.7 (2103–3317)/3567
YJL045w	SDH1 (11)	succinate dehydrogenase flavoprotein	83.5 (1–634)/634	78.6 (620–1766)/1902
YJL034w (KAR2)	SSA1 (1)	nuclear fusion protein KAR2 precursor	63.5 (50–663)/682	67.0 (156–1962)/2046
YJL034w (SSC1)	YEL030w (5)	heat shock protein	82.6 (17–642)/654	75.8 (205–1889)/1962
YJR047c (ANB1)	YEL034w (5)	translation initiation factor	90.4 (2–157)/157	91.4 (1–465)/471
YJR048w (CYC1)	YEL039c (5)	cytochromic isoform 1	85.8 2–107)/109	81.9 (113–323)/327
YJR049c (UTR1)	YEL041w (5)	UTR1 protein	57.0 (104–509)/530	63.8 (419–1392)/1590
YJR051w (OSM1)	YEL047c (5)	involved in osmotic redulation	63.5 (36–499)/501	63.7 (218–1469)/1503
YJR066w (TOR1)	TOR2 (11)	phosphatidyl-inositol kinase	68.0 (62–2470)/2470	67.2 (2786–7410)/7410
YJR103w (URA8)	URA7 (2)	CTP synthase	79.0 (1–562)/564	71.7 (146–1631)/1692
YJR155w	N0300 (14)	similar to aryl-alcohol dehydrogenase	89.9 (1–288)/288	87.7 (1–389)/864
YJR156c	N0295 (14)	similar to thiamine-repressed nmt-1	98.8 (1–340)/340	98.4 (568–1011)/1020
YJL221c	YJL216c	similar to α-glucosidase MAL35 (S46183)	66.3 (11–587)/589	62.8 (199–1767)/1767
YJL219w	YJL214w	similar to hexose transport protein LGT3	65.2 (33–567)/567	66.3 (226–1685)/1701
YJL079c	YJL078c	unknown	66.7 (152–298)/299	66.2 (551–861)/897
YJL052w (TDH1)	YJR009c (TDH2)	glyceraldehyde-3-phosphate dehydrogenase	65.0 (1–331)/331	92.4 (1–996)/996
YJL038c	YJL037w	unknown	36.3 (5–218)/219	34.0 (295–640)/657

[a]Where known, chromosomal location is indicated in parenthesis.
[b]Function of genes on chromosome X, when available, or else function of their homologues on other chromosomes.
[c]Numbers indicate % of aa identity, boundaries of aa comparison (in brackets) and size of the ORF on chromosome X (number after dash).
[d]Same as above, but in nt.

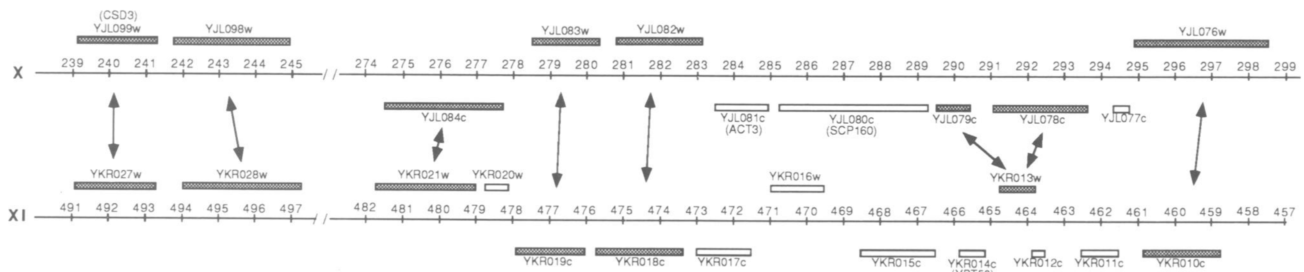


Fig. 4. Physical comparison of the location of genes and syntenic segments on chromosome X with that of their counterparts on other chromosomes. The precise position of the genes was deduced from the present sequence and re-drawn to scale (coordinates are in kb). Elements above and below the scale belong to the Watson and the Crick strands, respectively. Shaded boxes represent the ORFs with a counterpart on the other chromosome. On the whole, physical distance (and the structures located therein) between any two ORFs on the same syntenic segment is not respected on chromosomes other than X. Exceptions are the consecutive ORFs YJL099w (*CSD3*) and YJL098w on chromosome X and their homologues YKR027w and YKR028w on chromosome XI, the consecutive ORFs YJL083w and YJL082w on chromosome X and their homologues YKR019c and YKR018c.

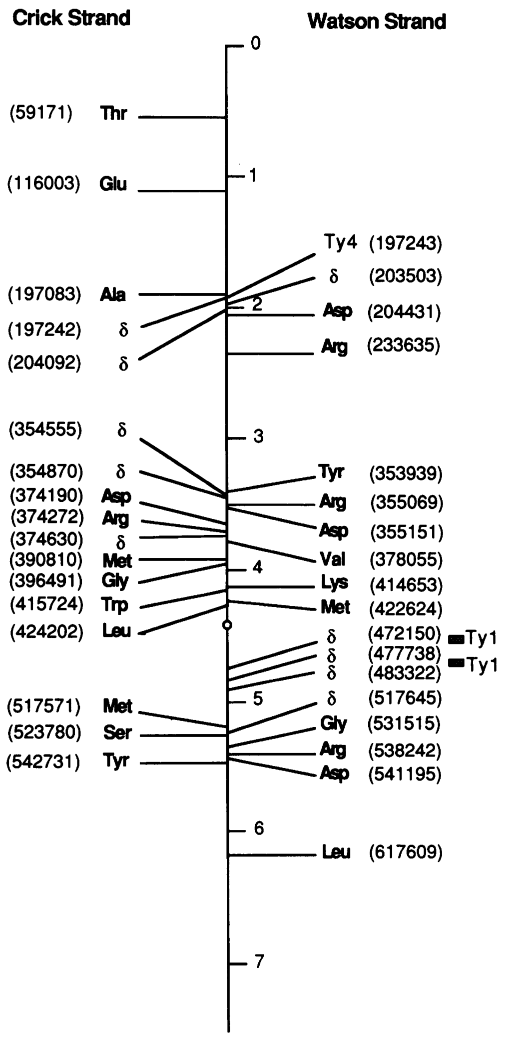


Fig. 5. Position of tRNA genes, Ty sequences and LTRs on chromosome X. The positions were drawn to scale relative to the complete sequence. Elements on the Watson and Crick strands are displayed on the right- and left-hand side, respectively. Only the 5′ coordinate is given.

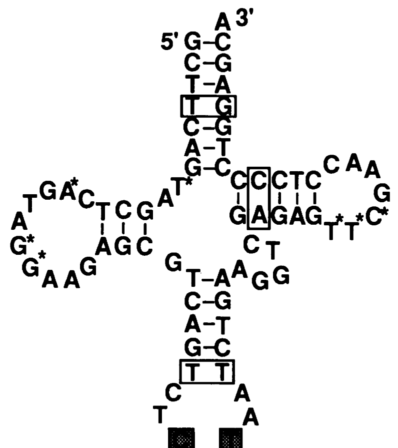


Fig. 6. A clover-leaf structure of yeast tRNA^{Met} on chromosome X (422 624–422 696). All canonical bases are indicated by asterisks. Mismatched base pairs in the stems are boxed. The shadowed nucleotides are the anticodon.

X display an intron 3′ to the anticodon sequence, as previously observed. These include two *tRNA*^{Tyr} with an intron of 14 nt, one of the two *tRNA*^{Leu} with a 19-nt intron and the unique *tRNA* ^{Trp} with an intron of ~29 nt. Its exact size is difficult to assess because base pairing is possible between several short sequences in the anticodon stem, creating an extra arm of variable length.

The entire chromosome X sequence was scanned in parallel for the presence of complete Ty elements or solo remnants or LTR thereof. As shown in Figure 5, several of these have been found. One complete Ty4 is present at position 197243–203468 and two complete Ty1 at position 472150–483659. The two elements are in tandem and share a central δ element. In addition, several solo LTRs are observed. As reported, with the exception of Ty1 these elements are located in the vicinity of *tRNA* sequences. However, this association seems to be rather loose and, besides, it involves partners located on either strand relative to one another.

Comparison of the physical and genetic maps of the chromosome X

The genetic map of chromosome X includes 60 genes or markers, of which 48 were mapped in a linear array and 12 remained unmapped (Mortimer *et al.*, 1995). Figure 7 shows a comparison of this map with the physical map deduced from the complete nucleotide sequence. Contrary to what has been reported for chromosome XI (Dujon *et al.*, 1994), no gross translocation or inversion was observed here. On the whole, the intergenic distance on the genetic map is roughly proportional to the physical distance, indicative of a relatively uniform recombination frequency over chromosome X. However, closer examination reveals some interesting discrepancies. First, genetic mapping has assigned the previously sequenced *CYR1* gene (alias *CDC35*, *HSR1*, *SRA4* and *TSM0185*), encoding adenylyl cyclase, to a site indistinguishable from that of *sui2*. This assignment is clearly incorrect, as the sequence data shows that this gene is in fact located on the left arm of the chromosome, close to the centromere. Second, marked differences are observed in map distances, the ratio between genetic and physical map distances ranging from 0.02 cM per kb for the *TDH2/met3* marker pair, to 0.84 and 4.74 cM per kb for the *met3/ilv3* and *ilv3/ess1* pairs, respectively. The relatively high frequency of recombination observed in these latter intervals strongly suggests the existence of preferred sites for the initiation of meiotic recombination, similar to those found in the *arg4* region on chromosome VIII (Nicolas *et al.*, 1989; Sun *et al.*, 1989) and the *MAT/thr4* region on chromosome III (Jacquet *et al.*, 1991). It is interesting to note that these intervals of high recombination frequencies in chromosome X appear to coincide with the sharp peak in the G+C content in the right arm of the chromosome (Figure 2).

In all, 31 of the mapped and one, *tRNA*^{Ser}, of the unmapped could be unambiguously assigned to an ORF or a *tRNA* gene on the basis of sequence comparison. A total of 28 loci cannot at present be attributed to specific ORFs on the physical map of chromosome X.

Discussion

The various elements of the chromosome X sequence referred to above are depicted in Figure 8. The present

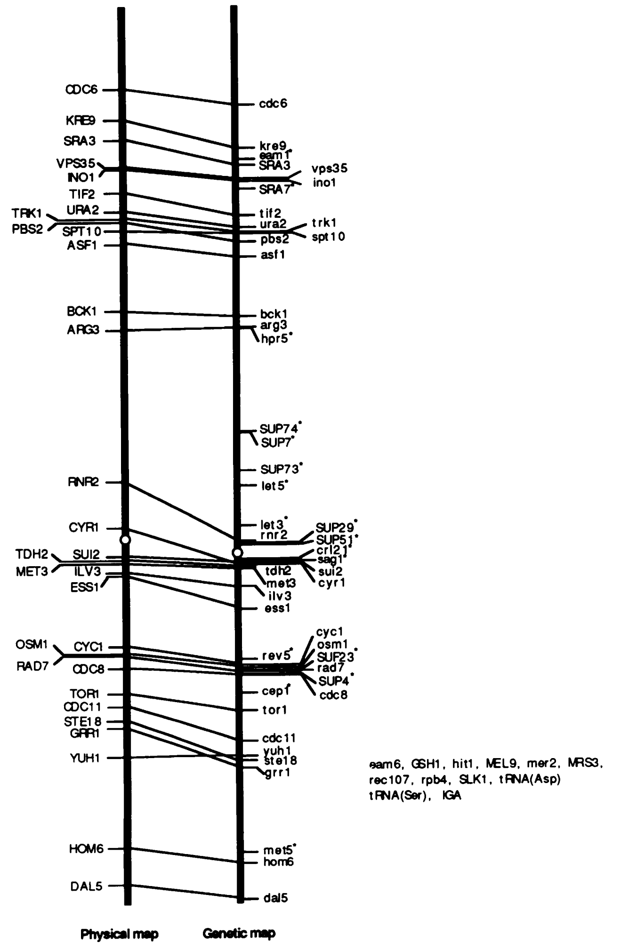


Fig. 7. Comparison of the genetic and physical maps of yeast chromosome X. The genetic map is re-drawn from Mortimer (Mortimer *et al.*, 1995). The unmapped genes or markers are listed on the right. The physical map deduced from this work has been drawn to scale. The circle indicates the position of the centromere. Genes or markers for which no corresponding ORF has been identified on the physical map are indicated by an asterisk.

report brings the number of completely sequenced chromosomes from the yeast *S.cerevisiae* to nine, chromosome X ranking second in this series by virtue of its size. Thus, nearly 40% of the *S.cerevisiae* genome sequence is now accessible to analysis, availability of the whole sequence being anticipated for 1997. The sequence of chromosome X has been established in S288C, a *S.cerevisiae* strain chosen by all members of the European Union sequencing consortium led by André Goffeau. While the study of this sequence reveals no features that are specific for chromosome X, it corroborates several observations made with the previously sequenced chromosomes.

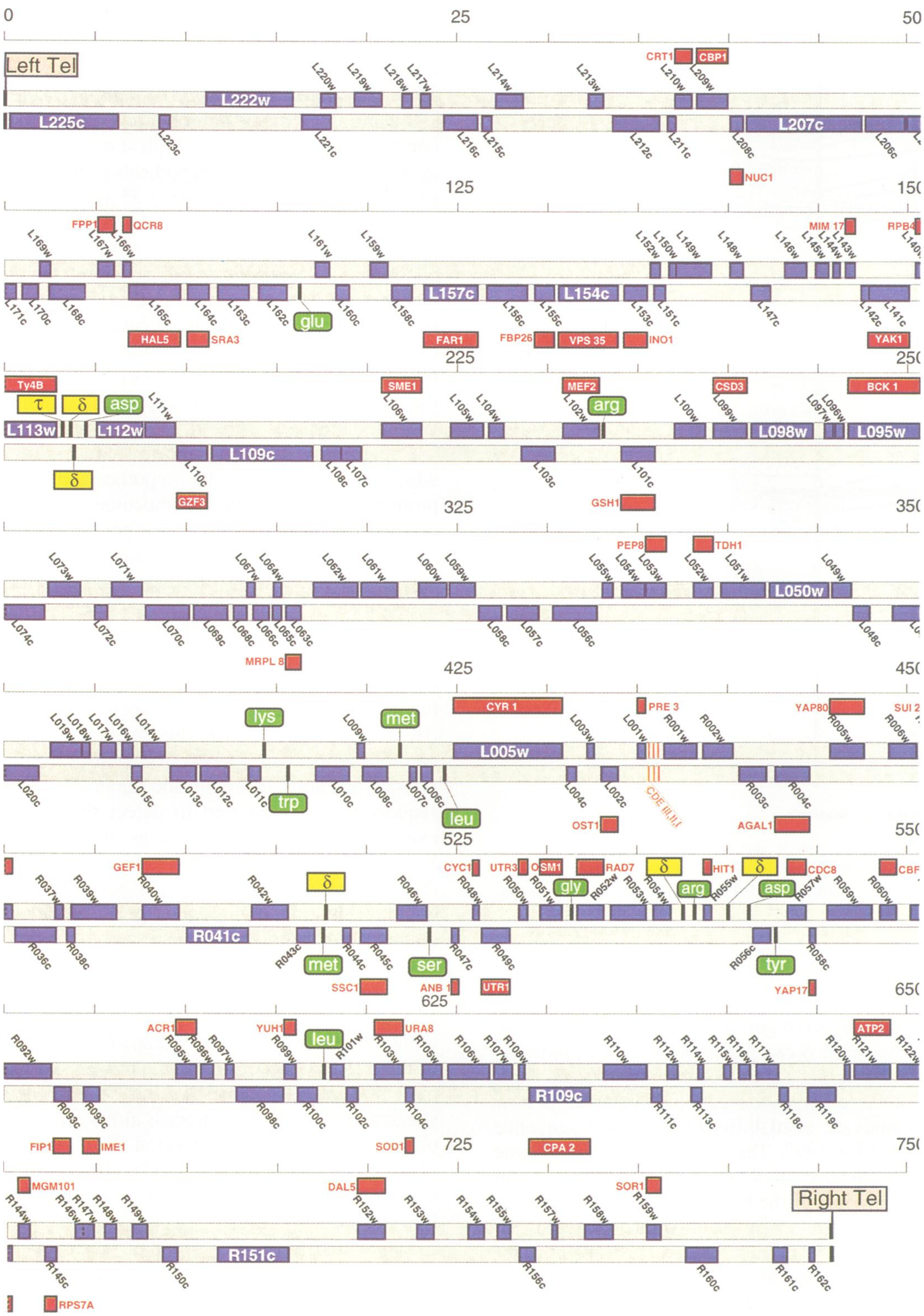
Taking into account only those ORFs whose characterictics, such as size, CAI and disposition leave no doubt as to their existence as real genes, a minimal density of one gene per 2000 nt can be estimated. All these genes are regularly spaced along the chromosome, with no predilection for either strand. Following translation and comparison of the deduced amino acid sequence with database entries, the products of these ORFs can be categorized as follows: (i) 102 proteins previously identified in *S.cerevisiae* and encoded by genes already assigned to chromosome X; (ii) 16 proteins with strong similarity, or even near identity, to known *S.cerevisiae* proteins, but whose coding gene has not previously been shown to reside on chromosome X; (iii) 22 proteins with a FastA score much greater than 200—equal to at least half the self-score, i.e. the score obtained when the protein is compared with itself. Such high scores can be considered as warranting a realistic hypothesis regarding the function of ORFs in this category; (iv) 35 proteins with a FastA score >200, though lower than half the self-score. A function can also be envisaged in this case, but with more caution; (v) 92 proteins with no significant FastA score but displaying a particular motif signature; (vi) 112 proteins with no match at all in database entries. This last category remains numerically important, since it includes nearly 30% of the ORFs, a proportion that fully vindicates the systematic sequencing approach of the *S.cerevisiae* genome launched in 1989.

Regarding ORFs in categories (iii) and (iv) above, for which a function can be hypothesized, several of the proteins discovered in chromosome X are worth mentioning. For instance, three new genes encoding different subunits of the cytosolic chaperone complex (*CCT5*, *CCT7* and *CCT8*) have been discovered on chromosome X in addition to *CCT3*. This brings the number of fully sequenced *CCT* genes in *S.cerevisiae* to eight. Together with the versatility of yeast versus mouse genetics, availability of these sequence data will undoubtedly promote fine molecular analysis of this important chaperone system. Another remark concerns the discovery of a Cl^- channel gene (Huang *et al.*, 1994c) on chromosome X. In this respect, it is both surprising and remarkable that systematic sequencing was required to detect the first Cl^- channel ever described in a species as thoroughly studied as *S.cerevisiae*. Here again, availability of the gene and of disruption mutants thereof will permit identification by complementation homologous genes in other species of interest, in particular in plants.

Chromosome X stands out because of the number of *tRNA* genes (24) it accommodates, capable of transferring 13 different amino acids. However, what is even more remarkable and has so far escaped notice is that folding of these tRNAs according to the clover-leaf model reveals quite a few mismatches in the several stems. This is suggestive of an editing process aiming at correcting some of these mismatches, as reported for various tRNAs from plants (Maréchal-Drouard, 1993). Of course, validation or dismissal of this hypothesis must await analysis at the RNA level.

Duplicated genes are found in chromosome X, as in other *S.cerevisiae* chromosomes. These include both intra- and interchromosomal duplications. Furthermore, actual syntenic regions can be recognized in the latter case. The implications are 2-fold, pertaining (i) to the study of the evolution of the yeast genome and (ii) to function analysis, as it is known that disruption of a single gene frequently does not result in any phenotypic alteration. By the same token, a clue to the function of a gene might in some instances be provided by disruption of all the genes belonging to a given family.

To conclude, it must be stressed that this brief account of the sequence analysis of chromosome X cannot cover all the information embedded in the nucleotide sequence



and that many biological analyses will be needed to exploit this mine of information in the years to come.

Materials and methods

Chromosome X DNA

Total yeast DNA was obtained from FY1679, a diploid strain issued from the cross between strains FY23 (*MAT***a**, *ura3-52*, *trp1Δ63*, *leu2Δ1*, *GAL2*) and FY73 (*MATα*, *ura3-52*, *his 3Δ200*, *GAL2*). FY23 and FY73 are derived from strain S288C and are isogenic with it except for the markers indicated (Winston *et al.*, 1995). The construction of an ordered cosmid library and of an *Eco*RI restriction map have been previously published (Huang *et al.*, 1994a). Overlapping cosmids covering the chromosome X contig were distributed within a consortium of 15 laboratories. The telomeres and subtelomeric regions were cloned in vector pEL61, as described by Louis and Borts (1995).

Determination, assembly and analysis of the sequence

Sequencing strategies and methods varied among the 15 collaborating laboratories (Table V). Sequence assembly in the single contracting laboratories was performed by a variety of software program packages.

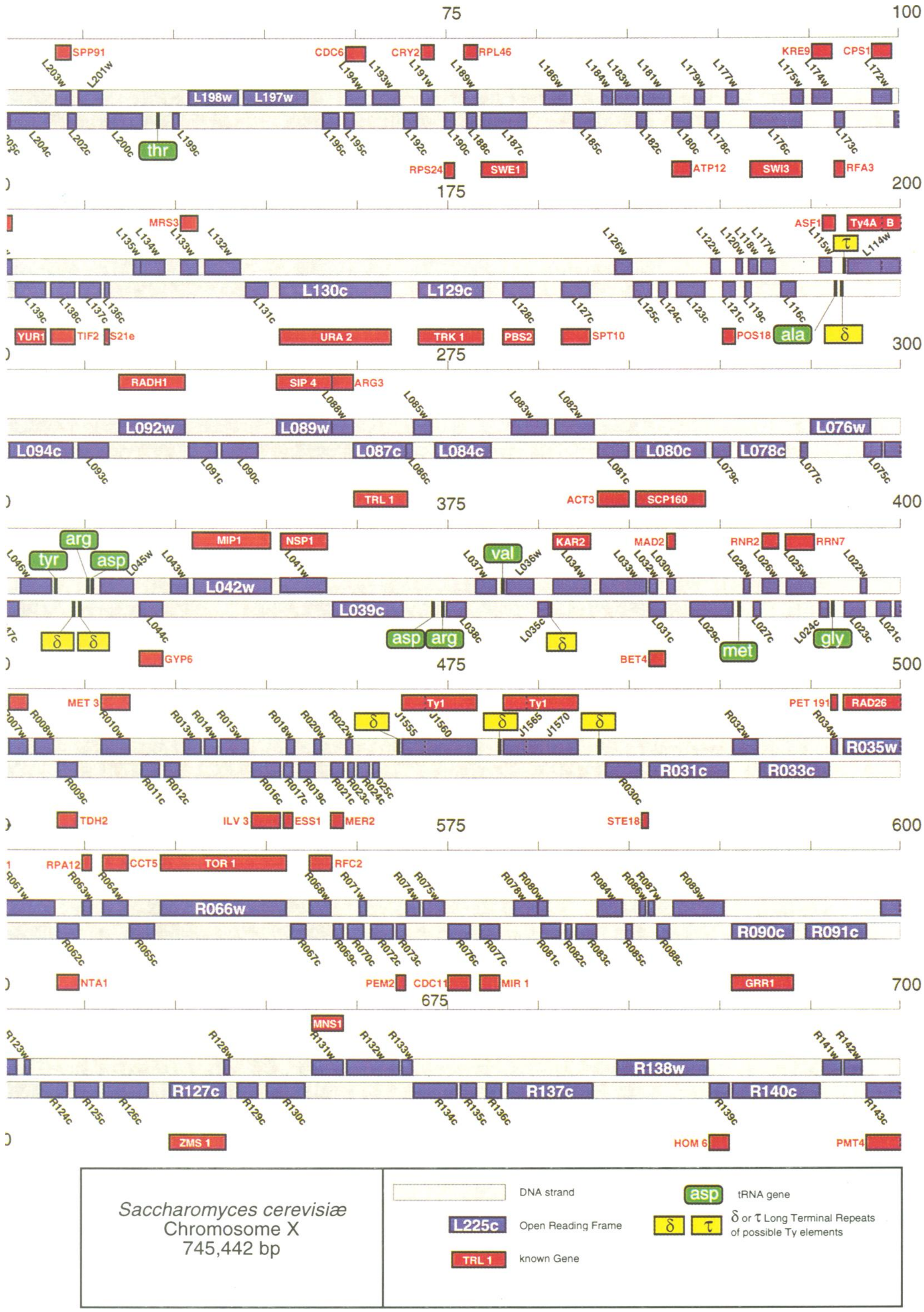


Fig. 8. Chromosome X map deduced from the complete sequence. The chromosome and its constitutive elements are drawn to scale. The top bar represents the Watson strand oriented 5′ to 3′ from left to right, the bottom bar the Crick strand. The conserved elements of the centromere are designated as CDE I, II and III. ORFs on the left and right arm are designated by the letters L and R, respectively, before their number (numbering is in increasing order from the centromere). Full designations, in accordance with the official ORF nomenclature, are obtained by adding again the letters Y (for yeast) and J (for chromosome X) at the beginning, and w (Watson) or c (Crick) at the end.

The telomeres were cloned in Oxford. The left telomere was sequenced in one of 15 laboratories. The right telomere and the PCR fragment filling the gap were sequenced in Berlin. Completed contigs submitted to MIPS were stored in a data library and assembled using the GCG software package 7.2 for the VAX (Devereux *et al.*, 1984) The nature and position of genetic elements have been deduced from the sequence using the following principles: (i) all possible intron splice site/branch-point pairs were detected using specially defined patterns (Fondrat *et al.*, 1994); (ii) ORFs occurring in all possible frames were listed. ORFs containing at least 99 contiguous sense codons following an ATG and

Table V. Methods used by each of the collaborating laboratories

Whole cosmid Shotgun	Restricted fragments		
	Shotgun	TN*1000*	Nested deletions
Louvain (M)	Gembloux (M)	Darmstadt (M)	München (A)
Heidelberg (M)	Amsterdam (A)	Frankfurt (A)	Copenhagen (A)
Konstanz (M)			Düsseldorf (A)
Paris (A)			Ghent (A)
Gif (A)			Herakleion (M)
Rennes (A)			

M, manual methods; A, automated methods.

those containing 50–98 codons were retained for further analysis, in both cases provided they were not entirely contained within a longer ORF on either DNA strand. Searches for similarity of the deduced protein sequences to entries in the databanks were performed by FastA (Pearson and Lipman, 1988) in the Protein Sequence Database of PIR International (release 44) and other databases. Protein signatures were detected using the PROSITE dictionary (release 11.1) (Bairoch, 1989). ORFs were assigned probable functions when the alignments from FastA searches showed significant similarity and/or protein signatures were apparent, whereas FastA scores <200 were considered insufficient to confidently assign function. The complete sequence was also searched for *tRNA* genes ('trnascan') (Fichant and Burks, 1991), centromere and telomere consensus elements and for δ, σ or τ elements by comparison with a data set of such elements previously characterized in yeast. Compositional analyses of the chromosome were performed using the X11 program package (C.Marck, unpublished results). For calculations of CAI and GC content of ORFs, the algorithm CODONS (Lloyd and Sharp, 1992) was used.

Sequence verifications and quality controls

All sequences submitted by collaborating laboratories to the Martinsried Institute for Protein Sequences (MIPS) data library were subjected to quality controls. The procedure was comprised of three major steps. First, the strategy of each contractor was checked by the coordinator to pinpoint possible weak points and request the sequencers to review their electrophoretograms to assess the quality of their reads in these less documented regions. Second, once cosmid sequences had been entered in the database, the match between the overlaps was held to provide an assessment of the respective quality of the neighbouring partial sequences. Third, each of the cosmids that had been distributed to the contractors for sequencing was shotgunned, size-selected to ~300–500 bp and cloned in plasmid vector, the size of the inserts ensuring that sequencing with the universal forward and reverse primers would provide a 300–400 double-stranded sequence. The subclones from each cosmid were sent with coded names to a different sequencer. The double-stranded part of each sequence was then sent to MIPS and compared with the initial sequence. The number of verification sequences per cosmid clone (averaging 15–30) varied according to the quality of the initial sequencing as deduced from alignment within the overlaps. Any discrepancy detected between overlapping partial sequences or between the sequence initially submitted and the verification sequence was addressed as follows. A stretch of 20 bp including the discrepancy, but not centering on it, was pointed out to each party for reviewing and re-submission to MIPS, whether modified or not. This procedure was sufficient to remove most discrepancies, as one party usually provided a revised sequence matching the other's. Resistant cases were dealt with by requesting both parties to send the electrophoretograms corresponding to the conflicting sequences to the coordinator, who made a decision and requested resequencing if necessary.

The sequence data reported are available through http: //mips.biochem.mpg.de/yeast

Acknowledgements

We wish to thank B.Dujon for fruitful discussion and for help with the gene density and G+C composition plots, and G.Le Provost for secretarial assistance. The laboratory consortium operating under contracts with the European Commission was initiated and organized by A.Goffeau. This study is part of the second phase of the European Yeast Genome Sequencing Project carried out under the administrative coordination of A.Vassarotti (DG-XII) and the Université Catholique de Louvain, and under the scientific responsibility of F.Galibert as DNA coordinator. This work was supported by the European Commission under the BRIDGE and Biotech programmes, the Groupe de Recherche et d'Etudes sur le Génome (GREG) and the Centre National de la Recherche Scientifique (CNRS) (FR), the Wellcome Trust (UK), the Région Wallonne and the Fonds National de la Recherche Scientifique (BE), the Bundesminister für Forschung und Technologie (DE) and the Ministry of Industry and Technology (GR).

References

Bairoch,A. (1989) EMBL Biocomputing Technical Document 4. EMBL, Heidelberg, Germany.

Barrell,B.G. *et al.* (1994) Sequence of *S. cerevisiae* chromosome IX, http: //www.sanger.ac.uk/~ yeastpub/svw/sequencing.html

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

Delehodde,A., Goguel,V., Becam,A.M., Creusot,F., Perea,J., Banroques,J. and Jacq,C. (1989) Site-specific DNA endonuclease and RNA maturase activities of two homologous intron-encoded proteins from yeast mitochondria. *Cell*, **56**, 431–441.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Dietrich,F.S. *et al.* (1994) Sequence of *S. cerevisiae* chromosome V, http: //speedy.mips.biochem.mpg.de/mips/yeast/chr5

Dujon,B. *et al.* (1994) Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.

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

Fichant,G.A. and Burks,C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220**, 659–671.

Fondrat,C. and Kalogeropoulos,A. (1994) Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III. *Curr. Genet.*, **25**, 396–406.

Galibert,F., Mandart,E., Fitoussi,F., Tiollais,P. and Charnay,P. (1979) Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. *Nature*, **281**, 646–650.

Hemrika,W., Berden,J.A. and Grivell,L.A. (1993) A region of the C-terminal part of the 11-kDa subunit of ubiquinol-cytochrome-c oxidoreductase of the yeast *Saccharomyces cerevisiae* contributes to the structure of the Q(out) reaction domain. *Eur. J. Biochem.*, **215**, 601–609.

Hieter,P., Pridmore,D., Hegemann,J.H., Thomas,M., Davis,R.W. and Philippsen,P. (1985) Functional selection and analysis of yeast centromeric DNA. *Cell*, **42**, 913–921.

Hinnebusch,A.G. and Liebman,S.W. (1991) Protein synthesis and translational control in *Saccharomyces cerevisiae*. In Broach,J.R. *et al.* (eds), *The Molecular Biology of the Yeast Saccharomyces*. Cold Spring Harbor Laboratory Press, Plainview, NY, pp. 627–735.

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

Huang,M.E., Manus,V., Chuat,J.C. and Galibert,F. (1994b) Revised nucleotide sequence of the COR region of yeast *S. cerevisiae* chromosome X. *Yeast*, **10**, 811–818.

Huang,M.E., Chuat,J.C. and Galibert,F. (1994c) A voltage-gated chloride channel in the yeast *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **242**, 595–598.

Huang,M.E., Chuat,J.C. and Galibert,F. (1995) Analysis of a 42.5 kb DNA sequence of chromosome X reveals three tRNA genes and 14 new open reading frames including a gene most probably belonging to the family of ubiquitin-protein ligase. *Yeast*, **11**, 775–781.

Jacquet,M., Buhler,J-M., Iborra,F., Francingues-Gaillard,M.C. and Soustelle,C. (1991) The *MAT* locus revisited within a 9.8 kb fragment of chromosome III containing *BUD5* and two new open reading frames. *Yeast*, **7**, 881–888.

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

Ketchum,K.A., Joiner,W.J., Sellers,A.J., Kaczmarek,L.K. and Goldstein,S.A.N. (1995) A new family of outwardly rectifying potassium channel proteins with two pore domains in tandem. *Nature*, **376**, 690–695.

Kozak,M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.

Lloyd,A.T. and Sharp,P.M. (1992) CODONS: A microcomputer program for codon usage analysis. *J. Hered.*, **83**, 239–240.

Louis,E.J. and Borts,R.H. (1995) A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning. *Genetics*, **139**, 125–136.

Louis,E.J. and Haber,J.E. (1991) Evolutionarily recent transfer of a group I mitochondrial intron to telomere regions in *Saccharomyces cerevisiae*. *Curr. Genet.*, **20**, 411–415.

Louis,E.J., Naumova,E.S., Lee,A., Naumov,G. and Haber,J.E. (1994) The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics*, **136**, 789–802.

Maréchal-Drouard,L., Ramamonjisoa,D., Cosset,A., Weil,J.H. and Dietrich,A. (1993) Editing correct mispairing in the accepter stem of bean and potato mitochondrial phenylalanine transfer RNAs. *Nucleic Acids Res.*, **21**, 4909–4914.

Miosga,T., Witzel,A. and Zimmermann,F.K. (1994a) Sequence and function analysis of a 9.46 kb fragment of *Saccharomyces cerevisiae* chromosome X. *Yeast*, **10**, 965–973.

Miosga,T., Boles,E., Schaaff-Gerstenschläger,I., Schmitt,S. and Zimmermann,F.K. (1994b) Sequence and function analysis of a 9.74 kb fragment of *Saccharomyces cerevisiae* chromosome X including the BCK1 gene. *Yeast*, **10**, 1481–1488.

Miosga,T., Schaaff-Gerstenschläger,I., Chalwatzis,N., Baur,A., Boles,E., Fournier,C., Schmitt,S., Velten,C., Wilhelm,N. and Zimmermann,F.K. (1995) Sequence analysis of a 33.1 kb fragment from the left arm of *Saccharomyces cerevisiae* chromosome X, including putative proteins with leucine zippers, a fungal Zn(II)2-Cys6 binuclear cluster domain and a putative alpha2-SCB-binding site. *Yeast*, **11**, 681–689.

Mortimer,R.K., Cherry,J.M., Dietrich,F.S., Riles,L., Olson,M.S. and Botstein,D. (1995) Genetic map of *Saccharomyces cerevisiae*. Edition 17. http: //genome.www.stanford.edu/sacchdb/edition 12.html

Murakami,Y. *et al.* (1995) Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae*. *Nature Genet.*, **10**, 261–268.

Navarre,C., Ghislain,M., Leterme,S., Ferroud,C., Dufour,J.P. and Goffeau,A. (1992) Purification and complete sequence of a small proteolipid associated with the plasma membrane H(+)-ATPase of *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **267**, 6425–6428.

Navarre,C., Catty,P., Leterme,S., Dietrich,F. and Goffeau,A. (1994) Two distinct genes encode small isoproteolipids affecting plasma membrane H(+)-ATPase activity of *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **269**, 21262–21268.

Nicolas,A., Treco,D., Schultes,N.P. and Szostak,J.W. (1989) An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature*, **333**, 87–90.

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

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

Pryde,F.E., 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 abd functional significance of sub-telomeric repeat sequences. *Yeast*, **11**, 371–382.

Purnelle,B., Coster,F. and Goffeau,A. (1994) The sequence of a 36 kb segment on the left arm of yeast chromosome X identifies 24 open reading frames including *NUC1*, *PRP21(SPP91)*, *CDC6*, *CRY2*, the gene for S24, a homologue to the aconitase gene *ACO1* and two homologues to chromosome III genes. *Yeast*, **10**, 1235–1249.

Pütz,J., Puglisi,J.D., Florentz,C. and Giegé,R. (1993) Addictive, cooperative and anti-cooperative effects between identity nucleotides of a tRNA. *EMBO J.*, **12**, 2949–2951.

Rasmussen,S.W. (1995) A region of yeast chromosome X includes the *SME1*, *MEF2*, *GSH1* and *CSD3* genes, a TCP-1 related gene, an open reading frame similar to the *DAL80* gene, and a $tRNA^{Arg}$. *Yeast*, **11**, 873–883.

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

Sun,H., Treco,D., Scultes,N.P. and Szostak,J.W. (1989) Double stranded breaks at an initiation site for meiotic gene conversions. *Nature*, **333**, 87–90.

Vandenbol,M., Durand,P., Bolle,P.-A., Dion,C., Portetelle,D. and Hilger,F. (1994) Sequence analysis of a 40.2 kb DNA fragment located near the left telomere of yeast chromosome X. *Yeast*, **10**, 1657–1662.

Vandenbol,M., Durand,P., Dion,C., Portetelle,D. and Hilger,F. (1995) Sequence of a 17.1 kb DNA fragment from chromosome X of *Saccharomyces cerevisiae* includes the mitochondrial ribosomal protein L8. *Yeast*, **11**, 57–60.

Winston,F., Dollard,C. and Ricuperdo-Hovasse,S.L. (1995) Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast*, **11**, 53–55.

Zagulski,M., Babinska,B., Gromadka,R., Migdalski,A., Sulicka,J. and Herbert,C.J. (1995) The sequence of 24.3 kb from chromosome X reveals 5 complete open reading frames all of which correspond to new genes, and a tandem insertion of a Ty1 transposon. *Yeast*, **11**, 1179–1186.

Received on November 3, 1995; revised on January 5, 1996