Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae*

Yasufumi Murakami[1], Masanori Naitou[1], Hiroko Hagiwara[1,2], Takehiko Shibata[2,3], Masashi Ozawa[1], Syun-ichi Sasanuma[1], Motoe Sasanuma[1], Yukari Tsuchiya[1], Eiichi Soeda[5], Kazushige Yokoyama[5], Masaaki Yamazaki[4], Hiroyuki Tashiro[4] & Toshihiko Eki[1]

[1]*Division of Human Genome Research, Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan*
[2]*Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Urawa, Saitama 338, Japan*
[3]*Biodesign Research Group, The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Wako, Saitama 351-01, Japan*
[4]*Bioscience Research Laboratory, Fujiya, Co., 228 Soya, Hatano, Kanagawa 257, Japan*
[5]*Gene Bank, Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan*

Correspondence should be addressed to Y.M.

The complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VI (270 kb) has revealed that it contains 129 predicted or known genes (300 bp or longer). Thirty-seven (28%) of which have been identified previously. Among the 92 novel genes, 39 are highly homologous to previously identified genes. Local sequence motifs were compared to active ARS regions and inactive loci with perfect ARS core sequences to examine the relationship between these motifs and ARS activity. Additional ARS sequences were predominantly observed in 3' flanking sequences of active ARS loci.

The budding yeast *Saccharomyces cerevisiae* is an important model organism for the analysis of basic biological processes of higher eukaryotes. Although the yeast genome is relatively small[1] (16 chromosomes totalling a 16 Mb genome size), its molecular mechanisms for control of cellular growth, DNA replication, transcription, signal transduction and DNA repair are thought to be similar to those of higher eukaryotes. Since the density of the coding region is relatively high and an ordered set of cosmid clones have already been aligned on most of the chromosomes, sequencing of this organism is now being carried out in an international collaboration. Five reports have revealed that sequencing yeast chromosomes is a very efficient procedure for finding novel genes because two-thirds of the yeast genes identified through these projects have not been previously identified[2–6]. Completely sequencing the yeast genome should greatly facilitate our understanding of yeast chromosome organization. Here we describe the DNA sequence of yeast chromosome VI (270 kb) and report an additional 92 novel genes identified through this analysis. The structural features of chromosome VI are discussed, including G+C composition, gene density and distribution of autonomously replicating sequence (ARS) elements.

Among the 16 yeast chromosomes, chromosome VI is the only one which has been subcloned in its entirety into plasmids that were tested for ARS activity[7]. Nine ARS elements were identified previously (ARS8 has been further divided into two elements). Active ARS elements were isolated and their loci mapped. In addition, DNA sequence analysis of several of these ARS elements revealed two common features: the presence of an 11-bp consensus sequence(core sequence or domain A): 5'–(A/T)TTTA(C/T)(A/G)TTT(A/T)–3' and the presence of a domain having a higher A+T content than bulk yeast DNA usually found 3' to the ARS core consensus sequence[7]. The construction and analysis of the effects of point mutations, small deletions, and small

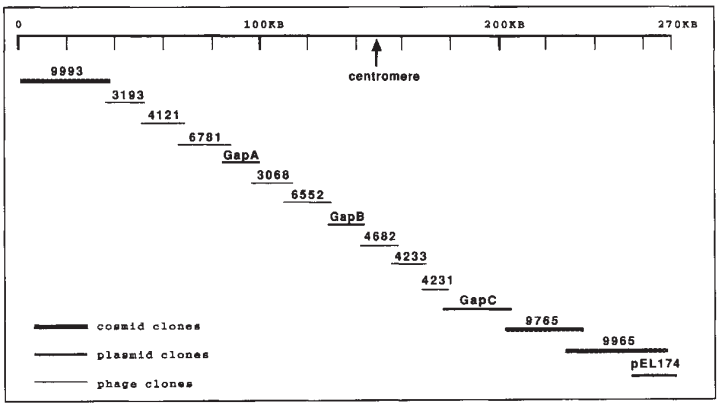

Fig. 1 Schematic representation of the lambda phage, plasmid and cosmid clones sequenced in this project. The nucleotide sequence data in this paper will appear in the GSDB, DDBJ, EMBL and NCBI sequence databases with the accession number D50617.

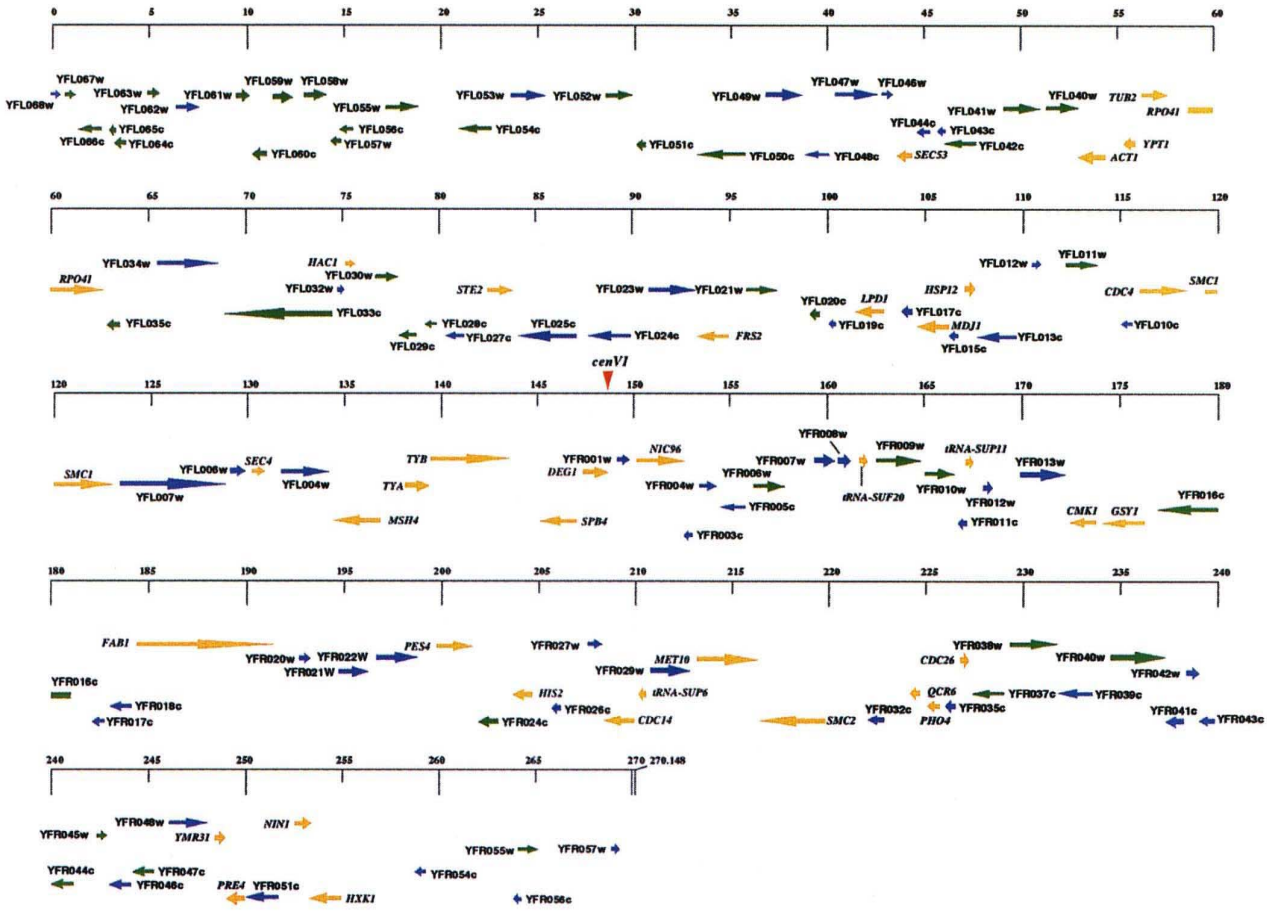

Fig. 2 Distribution of ORFs and ARS elements. All ORFs larger than 300 bp are indicated. Arrows represent ORFs or ARS elements. Orange arrows indicate previously identified genes, green arrows indicate ORFs highly homologous to database entries in SwissProt, PIR, GenBank or EMBL. ORFs of which FASTA optimum scores were larger than 200 were regarded as having a high degree of homology. Blue arrows indicate ORFs with no detectable similarity to sequences in the databases.

substitutions within this consensus sequence have demonstrated that it is essential for ARS function[8]. Although ARS elements do not necessarily have to contain a perfect consensus, all ARS elements should have at least one element which has 10/11 match with the ARS consensus sequence[9]. In chromosome VI, we found 373 elements which fulfilled this criterion. Among these 373 elements, sixteen were 100% identical to the consensus sequence. To investigate the mechanism by which active ARS elements are selected from such numerous consensus loci, we systematically analysed the sequence motifs around the ARS consensus sequence.

Sequencing strategy and physical map

To determine the complete nucleotide sequence of yeast chromosome VI, an ordered set of phage, plasmid and cosmid clones were used (Fig. 1). Most regions were covered by phage and cosmid clones[10]; the gaps between phage contigs were filled with plasmid clones[11]. We used the shotgun method of sequencing and easily obtained two sets of sequence data from one DNA template by sequencing from both sides of the plasmid vector.Gaps found after random sequencing were filled by primer walking on clones which linked two contigs. We used data from both strands and sequenced at least five times to detemine all nucleotide sequences.

The restriction map obtained from the sequence data was compared with a previously reported map[10]. The *Eco*RI and *Hin*dIII restriction sites in both maps matched with the exception of one fragment located approximately 170 kb from the left end (Fig. 2). Reinvestigation of the local restriction map indicated that the value obtained in this sequencing project is the correct size; therefore the 3373-bp fragment should be corrected to 4645 bp.

Gene identification

The distribution of open reading frames (ORFs) on chromosome VI is illustrated in Fig. 2. The sequence contained 129 non-overlapping ORFs greater than 300 bp. The density of ORFs calculated from this number was one per 2.1 kb. This value is consistent with those obtained for yeast chromosomes previously sequenced[2–6]. The average size of the genes identified was 476 codons, and the longest ORF (YFR019w) spanned 6834 bp (2278 codons).

A similarity search between the genes identified in this project and those in the public databases (GenBank,

Table 1 List of genes and features of chromosome VI

Position	ORF ID	Locus	Function or homology	FastA score	Acc. no.	Database
1	Y' element		Y' subtelomeric repeat			
4685	telomere		Telomeric repeat (C(1-3)A)			
4823	X element		X subtelomeric repeat			
836	YFL067w		Period clock protein (fragment)	202	P08399	S
2615	YFL066c		Hypothetical 137.7 kd protein in subtelomeric Y' repeat region	2372	P24089	S
3338	YFL065c		General amino acid permease	587	P24088	S
3846	YFL064c		Hypothetical 31.5 kd protein in *CBP2* 5' region	751	P24088	S
5066	YFL063w		Hypothetical 13.3 kd protein in *URA1* 5' region	383	P36030	S
6426	YFL062w		Hypothetical 45.2 kd protein in *MAL3S* 3' region	179	P25354	S
9545	YFL061w		Cyanamide hydratase (EC 4.2.1.69) (urea hydro-lyase)	303	P22143	S
10969	YFL060c		Hypothetical 21.4 kd protein in *DACA-SERS* intergenic region	233	P37528	S
11363	YFL059w		Hypothetical 31.6 kd protein in *DACA-SERS* intergenic region	803	P37527	S
12929	YFL058w		No message in thiamine protein 1	1191	P36597	S
14763	YFL057c		Hypothetical 40.9 kd protein in HMR 3' region	809	P25612	S
15431	YFL056c		Hypothetical 40.9 kd protein in HMR 3' region	687	P25612	S
17004	YFL055w		General amino acid permease	895	P19145	S
22787	YFL054c		Glycerol uptake facilitator protein	426	P11244	S
28232	YFL052w		Maltose fermentation regulatory protein MAL6R	1904	P10508	S
30540	YFL051c		Hypothetical 122.2 kd protein in *SIR1* 3' region precursor	372	P36170	S
35848	YFL050c		Hypothetical 109.7 kd protein in NUP100-MSN4 intergenic region	299	P35724	S
36803	YFL049w		NPL6 protein	195	P32832	S
42815	YFL046w		Myosin heavy chain A (MHC A)	114	P12844	S
44392	YFL045c	*SEC53*	Phosphomannomutase (EC 5.4.2.8) (PMM)	1291	P07283	S
45560	YFL044c		Dystrophin	116	P11532	S
47745	YFL042c		Hypothetical 149.7 kD Protein in *IRE1-KSP1* Intergenic region	609	P38800	S
49140	YFL041w		Iron transport multicopper oxidase	1303	P38993	S
51351	YFL040w		Glucose transport protein	389	P15729	S
54696	YFL039c[S]	*ACT1*	Actin	1781	P02579	S
55986	YFL038c	*YPT1*	GTP-binding protein (protein YP2)	998	P01123	S
56336	YFL037w	*TUB2*	Tubulin beta chain	2154	P02557	S
58782	YFL036w	*RPO41*	Mitochondrial DNA-directed RNA polymerase (EC 2.7.7.6)	6498	P13433	S
63795	YFL035c[S]		Hypothetical 27.4 kd protein in *PFK26-SGA1* intergenic region	422	P40484	S
74426	YFL033c		Protein kinase CEK1 (EC 2.7.1.-)	598	P38938	S
75178	YFL031w	*HAC1*	*HAC1* gene	2325	D26506	G
76829	YFL030w		Soluble hydrogenase, small subunit (EC 1.12.-.-)	286	P14776	S
79159	YFL029c		Protein kinase CSK1 (EC 2.7.1.-)	208	P36615	S
80211	YFL028c		Lactococcin a transport protein lactococcin (LCNC)	222	Q00564	S
82578	YFL026w	*STE2*	Pheromone alpha factor receptor	2022	P06842	S
87232	YFL025c		NADH-ubiquinone oxidoreductase chain 4 (EC 1.6.5.3)	136	P33511	S
90343	YFL024c		Hypothetical 195.1 kd protein in *DNA43-UBI1* intergenic region	101	P40457	S
90984	YFL023w		Glutamic acid-rich protein	197	P13816	S
95008	YFL022c	*FRS2*	Cytoplasmic phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20)	2454	P15625	S
95964	YFL021w		Nitrogen regulatory protein *GLN3*	303	P18494	S
99593	YFL020c		Hypothetical 13.0 kd protein in *URA1* 5' region	502	P35994	S
103121	YFL018c	*LPD1*	Dihydrolipoamide dehydrogenase precursor (EC 1.8.1.4)	2298	P09624	S
104456	YFL017c		Protease synthase and sporulation negative regulatory protein PAI 1	109	P21340	S
106230	YFL016c	*MDJ1*	MDJ1 protein precursor	2452	P35191	S
107250	YFL014w	*HSP12(GLP1)*	12 heat shock protein shock protein (glucose and lipid-regulated protein)	475	P22943	S
109924	YFL013c		Nucleolin (protein C23)	105	P13383	S
112339	YFL011w		High-affinity glucose transporter HXT2	2369	P23585	S
115737	YFL010c		Hypothetical 98.3 kd protein R10E12.1 in chromosome III	116	P34552	S
116139	YFL009w	*CDC4*	Cell division control protein 4	3646	P07834	S
119424	YFL008w	*SMC1*	Chromosome segregation protein	5660	P32908	S
123474	YFL007w		RNA polymerase (EC 2.7.7.48) (L protein)	101	P33453	S
130328	YFL005w	*SEC4*	Ras-related protein	995	P07560	S
131804	YFL004w		Hypothetical 14.4 kd protein in *RNR1-ILV1* intergenic region	176	P40046	S
137151	YFL003c	*MSH4*	MUTS protein homologue 4	4100	P40965	S
138198	Ty element	*TyA*	Transposon Ty1-17 49.8 kd hypothetical protein	2008	P25383	S
139471	Ty element	*TyB*	Transposon Ty1-17 154.0 kd hypothetical protein	6390	P25384	S
146928	YFL002c	*SPB4*	Putative rRNA helicase	2967	P25808	S
147125	YFL001w	*DEG1*	Depressed growth-rate protein	2195	P31115	S
148503	cenVI	*(CDE I)*				
148512	cenVI	*(CDE II)*				
148597	cenVI	*(CDE III)*				
149104	YFR001w		Myosin heavy chain, clone 203 (fragment)	133	P39922	S
150010	YFR002w	*NIC96*	96 kd nucleoporin-interacting component	3997	P34077	S
156138	YFR006w		X-pro dipeptidase (EC 3.4.13.9) (proline dipeptidase) (prolidase)	607	P12955	S
160528	YFR008w		Centromeric protein E (Cenp-E protein)	109	Q02224	S
162222	tRNA(G)	*SUF20*	Yeast *SUF20*(+) frameshift suppressor gene for tRNA-Gly	2321	X05270	G
162481	YFR009w		Probable ATP-dependent transporter YER036c	918	P40024	S
165059	YFR010w		Queuine trna-ribosyltransferase (EC 2.4.2.29)	325	P40826	S
167429	tRNA(Y)[S]	*SUP11*	Yeast Tyr-tRNA gene (Sup11)	557	J01380	G
173868	YFR014c	*CMK1*	Calcium/calmodulin-dependent protein kinase type I (EC 2.7.1.123)	2071	P27466	S
176382	YFR015c	*GSY1*	Glycogen (starch) synthase, isoform 1 (EC 2.4.1.11)	3516	P23337	S
180734	YFR016c		Neurofilament triplet M protein (160 kD neurofilament protein) (NF-M)	251	P12839	S
184489	YFR019w	*FAB1*	FAB1 protein	10567	P34756	S
199861	YFR023w	*PES4*	PES4 protein (DNA polymerase epsilon supressor 4)	2606	P39684	S
203068	YFR024c[S]		Hypothetical 41.8 kD protein in ARG4 3' region	969	P32793	S
204737	YFR025c	*HIS2*	Histidinol-phosphatase (EC 3.1.3.15)	1701	P38635	S

Table 1 List of genes and features of chromosome VI (continued)

Position	ORF ID	Locus	Function or homology	FastA score	Acc. no.	Database
210055	YFR028c	*CDC14*	Probable protein-tyrosine phosphatase (EC 3.1.3.48)	1659	Q00684	S
210694	tRNA(Y)[a]	*SUP6*	tRNA-Tyr(SUP6-o)	361	X07534	G
213299	YFR030w	*MET10*	Sulfite reductase (NADPH) flavoprotein component (EC 1.8.1.2)	4893	P39692	S
220093	YFR031c	*SMC2*	Chromosome segregation protein (DA-box protein SMC2)	5388	P38989	S
222946	YFR032c		Polyadenylate-binding protein (Poly(A) binding protein) (PABP)	145	P31209	S
224756	YFR033c	*QCR6*	Ubiquinal-cytochrome C reductase 17 kD protein (EC 1.10.2.2)	695	P00127	S
225945	YFR034c	*PHO4*	Phoshphate system positive regulatory protein	1338	P07270	S
226949	YFR036w	*CDC26*	Cell division control protein SCD26 (mutance of CDC26)	512	P14724	S
229172	YFR037c		Transcription regulatory protein SWI3	493	P32591	S
229366	YFR038w		Hypothetical 128.5 kD protien in *CCR4-TPD3* intergenic region	822	P31380	S
233531	YFR039c		Hypothetical 38.1 kD protein in *BCR* 5' region	101	P33915	S
234521	YFR040w		Hypothetical 121.4 kD protein in *BCK1* 5' region	886	P40856	S
238243	YFR041c		DNAJ protein	123	P17631	S
241425	YFR044c		Hypothetical TRP-ASP repeats containing protein in *DPB3-MRPL27*	478	P38149	S
242450	YFR045w		Putative mitochondrial carrier YBR291C	224	P38152	S
245153	YFR047c		Nicotinate-nucleotide pyrophosphorylase (carboxylating) (EC 2.4.2.19)	302	P30012	S
248510	YFR049w	*YMR31*	Mitochondrial ribosomal protein YMR31 precursor	602	P19955	S
249853	YFR050c	*PRE4*	Proteosome component Pre4 (EC 3.4.99.46) (Macropain subunit PRE4)	1246	P30657	S
252493	YFR052w	*NIN1*	Nuclear integrity protein 1	1323	P32496	S
255037	YFR053c	*HXK1*	Hexokinase A (EC 2.7.1.1) (Hexokinase PI)	2323	P04806	S
264192	YFR055w		Cystathionine beta-lyase (EC 4.4.1.8) (beta-cystathionase)	706	P06721	S
270012	telomere		telomere(TG1-3)			

Genes which had no homology (FastA score less than 100) were omitted from the table. Column 1, Nucleotide position of the start of each designated element(ATG
Genes which had no homology (FastA score less than 100) were omitted from the table. Column 1, Nucleotide position of the start of each designated element(ATG
for ORFs, the first nucleotide of all other elements). For the LTRs of the *Ty* elements, the beginning of the left LTR and the end of the right LTR are listed. Column 2, Genes are named according to established conventions: Y, yeast; F, chromosome VI; L and R, left or right chromosomal arm, respectively; w and c, gene is encoded on the top or bottom strand, respectively; and superscript 's' genes predicted to be spliced. Genes are numbered from the centromere (CEN) towards each telomere (TEL). Transfer RNA designations also follow convention: t indicates tRNA; the next letter is the one-letter code for the amino acid inserted by the tRNA. Column 3, Genetic names of genes identified previously. Column 4, A description of the function of the genes. Descriptions of proteins most similar to the other genes are also listed. Column 5, The FastA score for the alignment of the encoded protein to its closest homologue. FastA scores greater than 100 are generally considered to indicate significant homology between two proteins. Column 6, Database accession number of the closest homologue. Column 7, Name of the database from which the entry shown in column 6 is derived. S, Swissprot; G, Genbank. Similarity search of this table has been carried out using FastA program which was packaged in the 'Wisconsin GCG Sequence Analysis Package (Ver. 8.01)'. Word size for the search was six and other conditions were the same as the default setting. The FastA algorithm is described in detail by the author of the program[25].The domains analysed each spanned a one kb region upstream and downstream of the ARS core sequence.

EMBL, PIR, and SwissProt) are summarized in Table 1. Genes having FASTA optimum scores higher than 200 were regarded as highly homologous. Among the 129 ORFs identified here, 37 (28%) are identical to previously identified genes. Of the remaining 92 novel ORFs, 39 (30%) were highly similar to known genes in yeast or other organisms. One half of the ORFs (53 out of 129) are predicted to encode proteins that have no similarity to known sequences.

Comparative analysis of the genetic map[12] and the physical map constructed using our sequence data revealed two inversions, one between *cdc4* and *smc1* and one between *sup11* and *suf20* (Fig. 3). Two genes, *pho4* and *cdc26*, were very close on the physical map while they were distant in the genetic map.

Base composition of chromosome VI

As observed in previously sequenced chromosomes[2–6],

Table 2 Analysis of DNA sequence motifs involved in ARS activity

	ARS core sequence		Number of sequence motif				Relative postion of ARS to ORF[c]
ARS elements	Postion (bp)	Orientation[a]	ARS like (10/11)	SAR[b]	TopoII	ABFI	
YSCARSS 1	256373	c	4	0	0	0	
YSCARSS 2	216458	w (11/12)	2	2	1	0	
YSCARSS 3	199403	w	0	1	2	0	
YSCARSS 4	167731	c (11/12)	1	2	1	0	
YSCARSS 5	135567	c (11/12)	1	3	0	0	
YSCARSS 6	127866	c	0	1	0	0	
YSCARSS 7	68857	w (11/12)	1	2	0	0	
YSCARSS 8	32708, 32971	c+w (11/12)	2	2	0	0	
Inactive perfect match							
1	5492	c	2	1	1	1	
2	27963	w	0	1	0	0	
3	43487	w	0	5	1	0	
4	51029	c	0	2	1	0	
5	80491	w	0	2	2	0	
6	118748	c	1	3	0	0	
7	195135	w	0	3	1	1	
8	229906	c	0	0	1	0	
9	242428	c	0	2	0	1	
10	258900	c	0	0	1	0	
11	269757	w	0	1	4	1	

[a]ARS core sequences located on top (w) or bottom (c) strand. (11/12): one base mismatch to ARS 12-bp consensus sequence. [b]Nuclear scaffold attachment region. [c]Horizontal arrow indicates ORF direction from 5' to 3'. Vertical arrow indicates the postion of ARS.

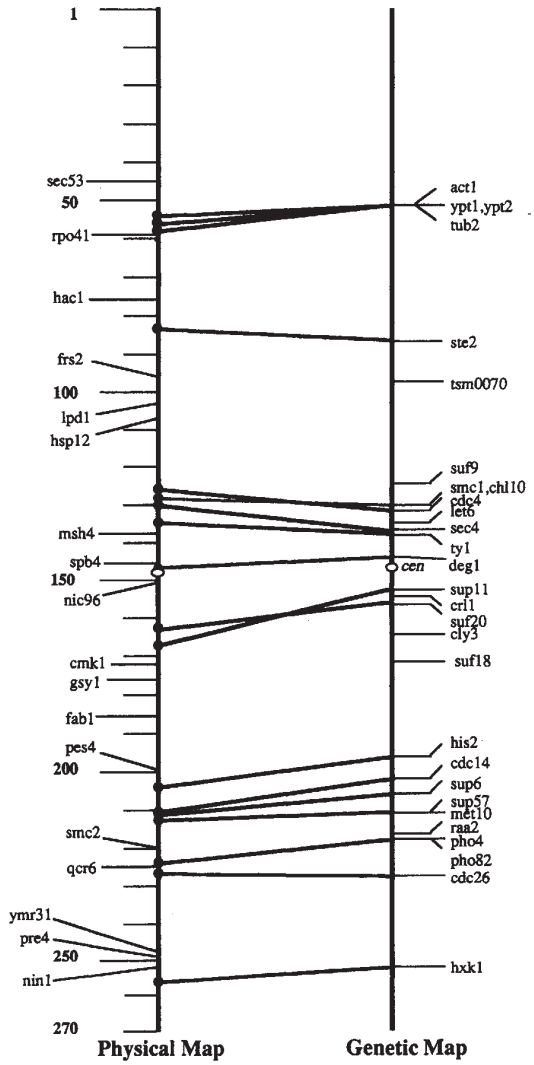

Fig. 3 Genetic and physical maps of chromosome VI. The true locations of the genes mapped previously on the genetic map are indicated by lines connecting them to the scale (in basepairs). Note the two minor discrepancies (two inversions between *cdc4* and *smc1* as well as between *sup11* and *suf20*) in the genetic map. This chromosome is divided into two arms; the region above the centromere is defined as the left arm and the region below the centromere is defined as the right arm[12].

base composition was clearly not uniform along chromosome VI (Fig. 4). The G+C composition of the central domain (108 kb–173 kb) was significantly lower than that at the ends (Fig. 4*a*). Alteration of the window size between 10 kb and 50 kb did not significantly change the pattern of G+C composition (data not shown). In chromosomes II[5] and XI[3] high gene density was observed predominantly in regions where the G+C composition is higher than average; however, no correlation between high G+C content and high gene density was observed in chromosome VI (Fig. 4*d*). On the contrary, the central A+T-rich region exhibited a relatively high gene density. In addition, no such correlation between high G+C content and gene density was observed in the recently sequenced chromosome VIII[4], thus the phenomenon observed in chromosomes II and XI is not universally true of all yeast chromosomes. Of further note, the top strand is preferentially utilized to encode genes in the A+T-rich central domain of chromosome VI (Fig. 4*b*,*c*). The gene density of the right end of the chromosome was significantly lower than that in other regions. The left end of the chromosome contained many small ORFs (Fig. 4*d*). A similar observation has been reported in the recently sequenced chromosome I[6]; the gene density of both ends of chromosome I was significantly lower than that of the central region.

Distribution of ARS elements

Chromosome VI had been previously subcloned, and all the subclones had been assayed for their ability to replicate autonomously (assayed for active ARS elements, which are candidates of the chromosomal replication origin)[7]. Using the complete sequence of chromosome VI, we systematically compared the features of active ARS element sequences and inactive ARS elements that had a complete ARS core consensus sequence. Mutational analysis of the ARS307 consensus sequence[13] and quantitative analysis of the consensus sequence of HO (HO gene: mating-type interconversion endonuclease gene) ARS mutations[14] had indicated that the ARS core consensus sequence should be modified as follows: 5'–(A/T)TTTA(T/C)(A/G)TTT(A/T)(T/C/G)–3'. All active ARS elements mapped to regions with lower than average G+C content with the exception of ARSS1, located near the right telomere (Fig. 5*a*). When one base mismatch with the core sequence was allowed, ARS consensus sequences were found all over the chromosome at a density of one per 800 bp (Fig. 5*b*). To determine the *cis*-factors governing ARS activity, we examined the 3' flanking regions of active and inactive ARS elements.

We chose eleven inactive ARS elements which contained perfect matches with the above core sequence (Table 2). All but one active ARS element (ARSS6) have an additional ARS-like consensus sequence in the 3' flanking region, whereas nine of eleven inactive loci had no additional ARS-like consensus sequences. We also analysed the distribution of nuclear scaffold binding site consensus sequences in the region adjacent to the ARS core sequences[15,16]. However, there was no prominent difference in the distribution pattern of nuclear scaffold binding domains between active ARS elements and inactive loci. Some additional *cis*-elements including transcription factor ABF1 binding sites[8,17,18], and topoisomerase (Topo) II cleavage sites[19] were also found in this region. Interestingly, no active ARS elements have an ABF1 binding site, and there are more Topo II cleavage sites in the 3' flanking region of inactive ARS elements.

The relationships between the position of ARS core sequences and the distribution of ORFs were also investigated. All but one active ARS element(ARSS5) were mapped in non-coding regions while five of eleven inactive loci were mapped in coding regions. This suggests that transcriptional regulation may play some role in the activation of ARS core sequences.

Discussion

Analysis of the sequence of chromosome VI revealed a unique distribution pattern of ORFs; there was preferential utilization of one strand in the central region of the chromosome. Further analysis of the possible function of

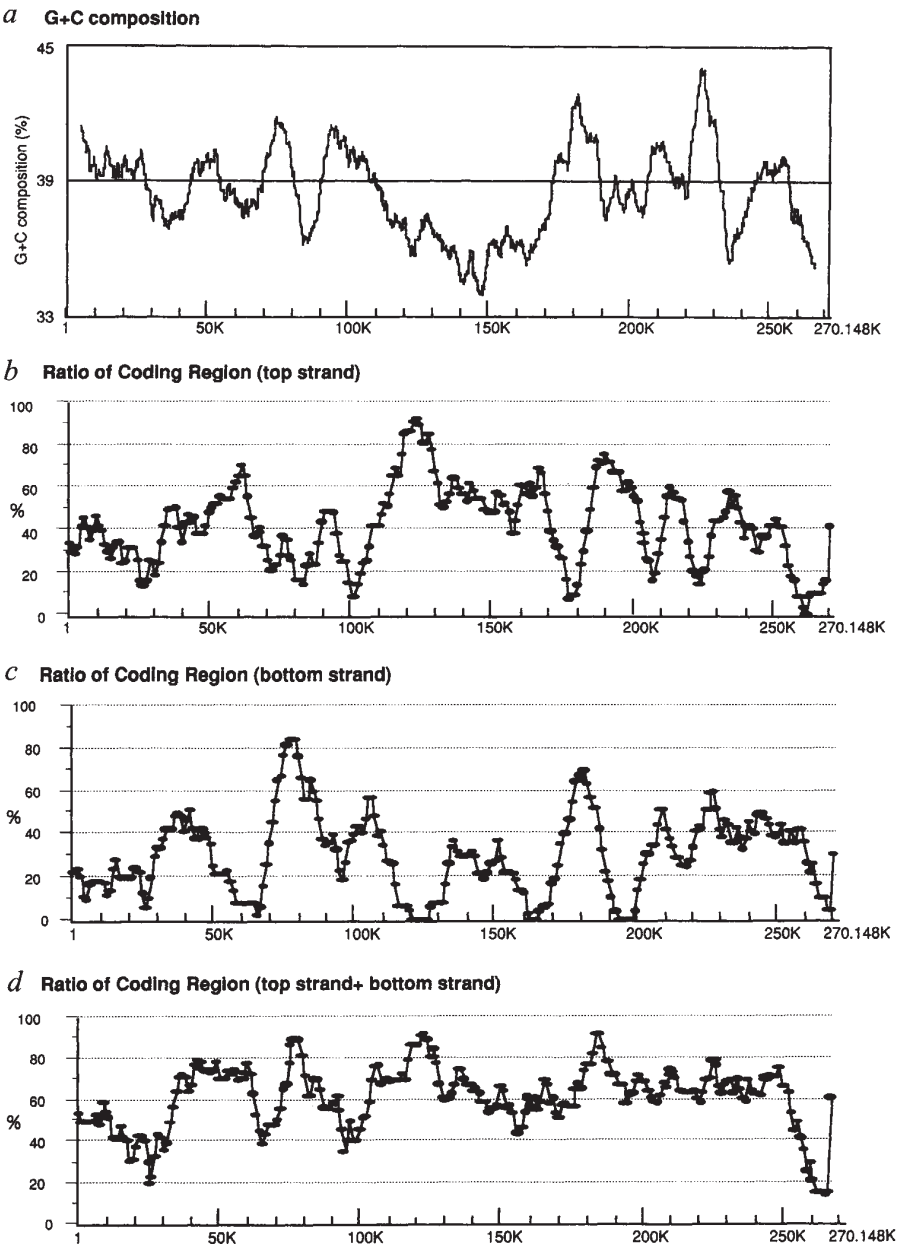

Fig. 4 Plot of coding density and G+C composition over the length of chromosome VI. *a*, Overall G+C composition was calculated over 10 kb windows spaced every 100 bp. The horizontal line marks the average G+C composition (38.50%). Utilization of sequence for protein coding: *b*, top strand (a strand which has 5' terminus at the left telomere); *c*, bottom strand (a strand which has 5' terminus at the right telomere); *d*, both strands. The ratios of coding to non-coding sequence, calculated with 10 kb windows spaced every 100 bp, are plotted.

these aligned genes and analysis of sequences upstream of these unidirectional genes may reveal the occurrence of polycistronic control which has been reported recently in a nematode genome study[20]. Such analysis is now under way in our laboratory.

In chromosome VI, we did not observe a similar relationship between gene density and G+C composition to that seen previously in chromosomes II[5] and XI[3]. Also, the apparent organization of chromosomes II and XI into regularly spaced intervals of G+C-rich and G+C-poor segments was not observed in chromosome VI. The results from analysis of both chromosomes VIII and VI suggest that the generality of these phenomena are unlikely[4]. A recent report on the chromosome I[6] showed that both ends of the chromosome were gene-poor and contained many non-functional gene fragments. In chromosome I the region 10 kb to 25 kb from the right end is duplicated. Interestingly, the right end of the chromosome VI contains few ORFs and the left end has a large number of short ORFs. The region 5–8 kb from the left end of chromosome VI is almost identical to a part of chromosome II and the region between 14–15 kb from the left end of chromosome VI had a common sequence with chromosome III. These results suggest that the end regions of such small chromosomes exhibit a high frequency of recombination events. A previous report[6], in conjunction with this observation, suggests that the role of the end region of such small yeast chromosomes is to increase the length of the chromosomes to ensure mitotic and meiotic stability of the chromosome. Functional analysis of the short ORFs on

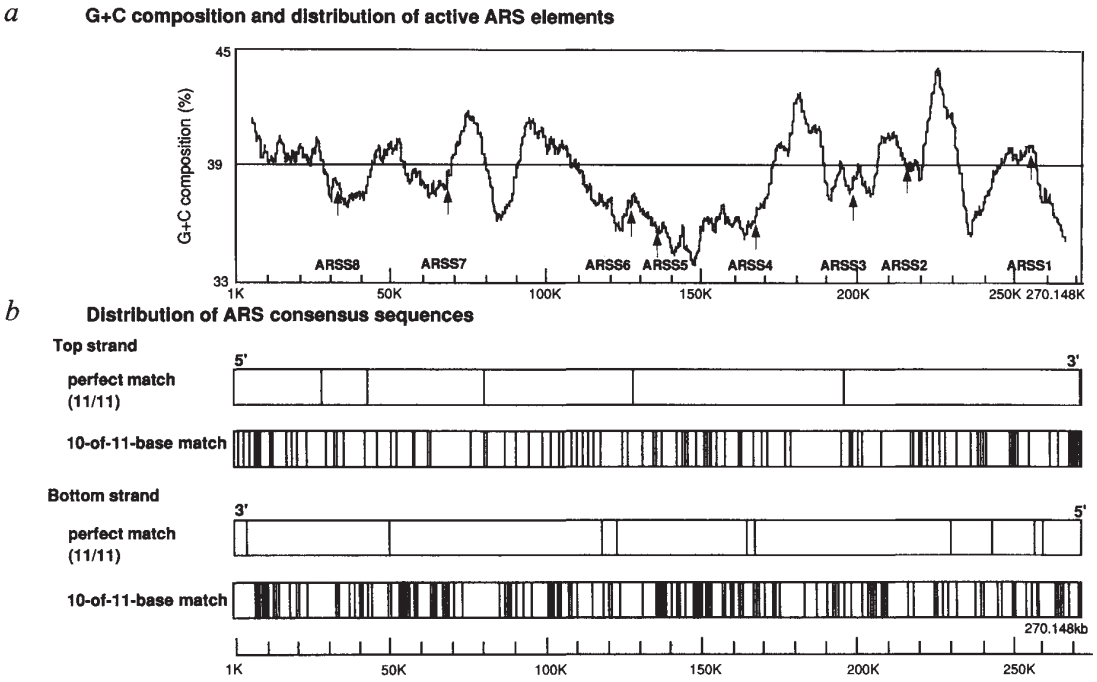

Fig. 5 Analysis of distribution of ARS core sequence and core-like sequence. *a*, G+C composition and distribution of previously reported active ARS elements. Each arrow indicates the position of an active ARS element. *b*, Distribution of ARS core sequence and core-like sequence (10 out of 11 bp match).

the left of chromosome VI may confirm this possibility.

To investigate the mechanism of how a small number of ARS elements are selected for activation from a large number of candidate loci, we systematically analysed the sequence motifs of the flanking regions around ARS core sequences. The results indicated that all but one ARS element have ARS-like consensus sequences in the 3' flanking region, whereas nine of eleven inactive loci had no additional 10-of-11-base matches to the consensus sequence. It is interesting to note that ARSS6, the active ARS consensus sequence that does not have an ARS-like consensus sequence, was reported previously to have markedly low ARS activities[7].

A comparison of G+C content with various window sizes showed no marked differences in G+C content around active ARS elements and inactive loci. This result argues against the importance of the existence of an A+T-rich domain for ARS activity, we are now carefully analysing the distribution of domains for unwinding adjacent to the core sequence.

We also analysed the relationship between the positions of active ARS elements and the distribution of ORFs. Interestingly, all but one active ARS element (ARSS5) mapped to non-coding regions while five of eleven inactive loci mapped to areas with coding sequences. Although this result suggests that transcriptional regulation may play a role in the activation of ARS core sequences, more detailed analysis regarding the distribution of transcriptional elements around ARS elements is required to reach a definite conclusion.

The most important impact of this study is the identification of numerous novel genes as has been found in previous analyses of other chromosomes. Since the gene density is relatively high in the yeast genome, sequencing chromosomal DNA leads to the immediate identification of novel genes. As the international collaboration to sequence the whole yeast genome proceeds, more and more interesting features of the yeast genome structure will be revealed. Upon completion of this international effort, all yeast genes will be identified, which will have far-reaching effects on the studies of biological processes in higher eukayotes.

Methods

Strains and vectors. *Escherichia coli* strain DH5α (*sup*E44 D*lac*U169(f80*lac*ZDM15) *hsd*R17 *rec*A1 *end*A1 *gyr*A96 *thi*-1 *rel*A1) was used for all subcloning and sequencing steps. Lambda phage and cosmid clones shown (Fig. 1) were isolated and mapped by Olson *et al.*[10]. *GapA*, *GapB* and *GapC* clones were isolated by Iwasaki *et al.*[11], and were kindly supplied by Drs. M. Olson (Washington University) and H. Yoshikawa (Nara Institute of Science and Technology). A plasmid clone containing the right telomere of chromosome VI, pEL174, was kindly supplied by E. Louis (John Radcliffe Hospital).

Preparation of the shotgun library. We isolated the phage clone inserts and recloned them into the Charomid 9-28 vector[21] at the *Sma*I site. To sequence plasmid clones, inserts were isolated and processed for shotgun library preparation. Cosmid clones were directly sonicated. Charomid DNA, plasmid DNA and cosmid DNA were purified by the alkaline lysis method followed by CsCl centrifugation in the presence of ethidium bromide[21]. Several hundred micrograms of purified DNA was obtained from an overnight culture (500 ml) grown in CircleGrow media (Funakoshi Co.). The insert DNA was purified by agarose gel electrophoresis and then subjected to sonication (Model SH7250, Seiko Co.). Sonicated DNA (over 1 kb in size) was fractionated by agarose gel

electrophoresis, purified, and treated with *Bal*31 nuclease (type S, Takara Co.) and Klenow enzyme (Toyobo Co.) to produce repaired blunt ends according to the manufacturer's protocol. The blunt-ended DNA was then ligated into *Sma*I-digested, dephosphorylated pUC19 DNA at an insert to vector ratio of 10:1. The ligation mixture was then transfected into competent DH5α cells prepared as described previously[22]. Several thousand recombinant clones were regularly obtained from 10 μg of purified fragment. The ratio of clones which had yeast DNA inserts was as high as 90% in a typical preparation.

Purification of the sequencing template. Plasmid DNA was purified using an automated plasmid DNA purifier (Model PI-100, Kurabo, Kurashiki), which uses the alkaline lysis method[23]. Full grown cultures of *E. coli* cells harbouring pUC19 plasmids containing yeast DNA inserts were transferred to serial tubes (five tubes each containing 2.5 ml of culture were connected in one serial tube). In this study, the procedure was modified slightly (we added one extra ethanol precipitation step). It took 14 h to finish purification of 160 samples. On average approximately 30 μg of purified DNA were obtained from a 5 ml overnight culture under our conditions. This was enough for up to ten sequencing reactions.

Sequencing. Sequencing reactions were performed using 1 μg of double-stranded plasmid DNA. DNA sequences were determined using an AmpliTaq polymerase dye primer sequencing kit and were analysed using an ABI 373A autosequencer[24] (Perkin Elmer).

Primers were synthesized on a DNA synthesizer (ABI 394 DNA synthesizer) and purified with OPC columns according to the manufacturer's protocol. Dye terminator cycle sequence reactions were carried out to fill the gaps of the contigs which had been assembled using 'Shotgun' software (Mitsui Knowledge Inc., Tokyo). To determine the order of the contigs, sequence reactions were carried out in each direction with the universal primer and reverse primer of the sequencing kits.

Sequence data assembly. Raw sequence data were assembled using 'Shotgun' software. The sequence data was transferred to a SUN Workstation (Sun 4/10) through a computer network. Determination of the final sequence was performed by comparing the chart obtained from the ABI 373A sequencer and the results of the assembled sequence data. All bases were covered by more than five fragments. The entire chromosome was fully covered by sequence data in both directions. For the regions where conflicts were observed after assembly of the sequence data, the charts from the sequencer were carefully compared and the samples were re-analysed or analysed with synthetic primers from alternative sequencing start sites. To verify the sequence data, comparisons of the sequence data of overlapping clones were carried out. There were only four base discrepancies out of 20 kb of data. These conflicts were resolved by re-sequencing. Although we estimate our sequence data has an overall accuracy of 99.98%, we will continue to revise the sequence of chromosome VI through DDBJ.

Analysis of the sequence data and database submissions. The final sequence was initially analysed using Genetyx software (Software Development Co., Tokyo) on a Macintosh computer. ORFs of 300 bp or longer were analysed for similarity with sequences in the GenBank EMBL, PIR and SwissProt databases.

The sequence data reported has been submitted to DDBJ/EMBL/GenBank data library under accession numbers D50617 (270 Kb full sequence), D44603 (clone 9993), D44594 (clone 3193), D44598 (clone 4121), D44595 (clone 6781), D44601 (clone gapA), D44596 (clone 3068), D31600 (clone 6552), D44604 (clone gapB), D44600 (clone 4682), D44599 (clone 4233 and 4231), D44606 (clone gapC), D44602 (clone 9765) and D44597 (clone 9965), respectively.

Acknowledgements

We thank M. V. Olson (Washington University), H. Yoshikawa (Nara Institute of Science and Technology) and E. Louis (John Radcliffe Hospital) for their help and contribution to this work. The telomere clone was developed under the support of the EU fund as well as a research grant from The Welcome Trust. We thank Y. Ikawa (Tokyo Medical and Dental University), F. Imamoto (Kyoto Pharmaceutical University), and D. Schlessinger (Washington University); K. Shinozaki and F. Hanaoka (RIKEN); A. Ono and K. Watanabe (Fujiya), N. Takahashi, S. Ishii (RIKEN), M. Muramatsu (Saitama Medical School) and T. Sakakura (Mie University). This project was supported by grants to the Life Science Research Project of RIKEN and the Science and Technology Agency of Japan as well as the McDonnell Foundation.

Received 27 January; accepted 2 May 1995.

1. Olson, M.V. Genome structure and organization in *Saccharomyces cerevisiae*. in *The molecular and cellular biology of the yeast Saccharomyces: genome dynamics, protein synthesis and energetics* (eds. Broach, J.R., Jones, E.W., and J.R. Pringle, J.R.) 1–39 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor,1991).
2. Oliver, S.G. *et al.* The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46 (1992).
3. Dujon, B. *et al.* Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371–378 (1994).
4. Johnston, M. *et al.* Complete nucleotide sequence of *Saccharomyces cerevisiae* Chromosome VIII. *Science* **265**, 2077–2082 (1994).
5. Feldmann, H. *et al.* Complete DNA sequence of yeast chromosome II. *EMBO J.* **13**, 5795–5809.
6. Bussey, H. *et al.* The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*. *Proc. natn. Acad. Sci. U.S.A.* **92**, 3809–3813 (1995).
7. Shirahige, K., Iwasaki, T., Rashid, M.B., Ogasawara, N. & Yoshikawa, H. Localization and characterization of autonomously replicating sequence from chromosome VI of *Saccharomyces cerevisiae*. *Molec. cell. Biol.* **13**, 5043–5056 (1993).
8. Campbell, J.L. & Newlon, C.S. Chromosomal DNA replication in *Saccharomyces cerevisiae*. in *The molecular and cellular biology of the yeast Saccharomyces: genome dynamics, protein synthesis and energetics* (eds Broach, J.R., Jones, E.W. & J.R. Pringle, J.R.) 41–146 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor,1991).
9. Deshpande, A.M. & Newlon, C.S. The ARS consensus sequence is required for chromosomal origin function in *Saccharomyces cerevisiae*. *Molec. cell. Biol.* **12**, 4305–4313 (1992).
10. Riles, L. *et al.* Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics* **134**, 81–150 (1993).
11. Iwasaki, T., Shirahige, K., Yoshikawa, H. & Ogasawara, N. The direct cloning of the yeast genome using the gap-filling method and the complete physical mapping of *Saccharomyces cerevsiae* chromosome VI. *Gene* **11**, 81–87 (1992).
12. Mortimer, R.K. *et al.* Fungi: *S. cerevisiae* (Nuclear genes). in *Genetic maps locus maps of complex genome, Sixth Edition Book 3, Lower Eukaryotes* (ed. O'Brien, S.J.) 36–56 (Cold Spring Harbor Laboratory Press, New York,1993).
13. van Houten, J.V. & Newlon, C.S. Mutational Analysis of the consensus sequence of replication origin from yeast chromosome III. *Molec. cell. Biol.* **8**, 3917–3925 (1990).
14. Kipling, D. & Kearsey, S. Reversion of autonomous replicating sequence mutations in *Saccharomyces cerevsiae*: Creation of eukaryotic replication origin with prokayotic vector DNA. *Molec. cell. Biol.* **10**, 265–272 (1990).
15. Amati, B.D. & Gasser, S.M. Chromosomal ARS and CEN elements bind specifically to the yeast nuclear scaffold. *Cell* **54**, 967–978 (1988).
16. Hoffman, J.F.-X., Laroche, T., Brand, A.H. & Gasser, S.M. RAP-I factor is necessary for DNA loop formation *in vitro* at the silent mating type locus HML. *Cell* **57**, 725–737 (1989).
17. Shore, D., Stillman, B.J., Brand, A.H. & Nasmyth, K.A. Identification of silencer binding proteins from yeast: Possible roles in SIR control and DNA replication. *EMBO J.* **6**, 461–467 (1987).
18. Diffley, J.F.-X. & Stillman, B. Interactions between purified cellular proteins and yeast origin of replication. *Cancer Cells* **6**, 235–243 (1988).
19. Spitzner, J.R. & Muller, M.T. A consensus sequence for cleavage by vertebrate DNA topoisomerase II. *Nucl. Acids Res.* **16**, 5533–5555 (1988).
20. Zorio, D.A.R., Cheng, N.N., Blumenthal, T. & Spieth, J. Operons as a common form of chromosomal organization in *C.elegans*. *Nature* **372**, 270–272 (1994).
21. Saito, I. & Stark, G.R. Charomids: Cosmid vectors for efficient cloning and mapping of large and small restriction fragments. *Proc. natn. Acad. Sci. U.S.A.* **83**, 8664–8668 (1986).
22. Hanahan. D. Studies on transformation of *Eschrichia coli* with plasmids. *J. molec. Biol.* **166**, 557–580 (1983).
23. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular cloning: a laboratory manual.* 2nd edn (Cold Spring Harbor Press, Cold Spring Harbor, 1989).
24. Hunkapiller, T., Kaiser, R.J., Koop, B.F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59–67 (1991).
25. Pearson W. Rapid and sensitive sequence comparison with FASTP and FASTA. in *Methods in Enzymology* **183** (ed. Doolittle, R.F.), 63–98 (1990).