where T_c decreases as the layer thickness of the nonsuperconducting $PrBa_2Cu_3O_7$ increases (*21–23*).

REFERENCES AND NOTES

1. J. G. Bednorz and K. A. Muller, *Z. Phys. B* **64**, 189 (1986).
2. T. A. Vanderah, *Chemistry of Superconductor Materials* (Noyes, Park Ridge, NJ, 1992).
3. A. Schilling *et al.*, *Nature* **363**, 56 (1993).
4. H. Ihara *et al.*, *Jpn. J. Appl. Phys.* **33**, L503 (1994).
5. C.-Q. Jin, S. Adachi, X.-J. Wu, H. Yamauchi, S. Tanaka, *Physica C* **223**, 238 (1994).
6. M. A. Alario-Franco *et al.*, *ibid.* **222**, 52 (1994).
7. Z. Hiroi *et al.*, *Nature* **364**, 315 (1993).
8. J. N. Eckstein *et al.*, *Appl. Phys. Lett.* **57**, 931 (1990).
9. T. Terashima *et al.*, *Phys. Rev. Lett.* **65**, 2684 (1990).
10. M. Y. Chern, A. Gupta, B. W. Hussey, *Appl. Phys. Lett.* **60**, 3045 (1992).
11. D. P. Norton, B. C. Chakoumakos, J. D. Budai, D. H. Lowndes, *ibid.* **62**, 1679 (1993).
12. C. Niu and C. M. Lieber, *J. Am. Chem. Soc.* **114**, 3570 (1992).
13. M. Yoshimoto, H. Nagata, J. Gong, H. Ohkubo, H. Koinuma, *Physica C* **185**, 2085 (1991).
14. M. Kanai, T. Kawai, S. Kawai, *Appl. Phys. Lett.* **58**, 771 (1991).
15. T. Siegrist *et al.*, *Nature* **334**, 231 (1988).
16. T. Kawai, Y. Egami, H. Tabata, S. Kawai, *ibid.* **349**, 200 (1991).
17. M. Lagues *et al.*, *Science* **262**, 1850 (1993).
18. X. Li, T. Kawai, S. Kawai, *Jpn. J. Appl. Phys.* **33**, L18 (1994).
19. S. D. Obertelli, J. R. Cooper, J. L. Tallon, *Phys. Rev. B* **46**, 14928 (1992).
20. W. A. Fietz and W. W. Webb, *Phys. Rev.* **178**, 657 (1969).
21. D. H. Lowndes, D. P. Norton, J. D. Budai, *Phys. Rev. Lett.* **65**, 1160 (1990).
22. J.-M. Triscone *et al.*, *ibid.* **64**, 804 (1990).
23. Q. Li *et al.*, *ibid.*, p. 3086.

24. We thank P. H. Fleming for assistance with substrate preparation. This research was sponsored by the Division of Materials Sciences, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems.

9 June 1994; accepted 9 August 1994

Complete Nucleotide Sequence of *Saccharomyces cerevisiae* Chromosome VIII

M. Johnston, S. Andrews, R. Brinkman, J. Cooper, H. Ding,
J. Dover, Z. Du, A. Favello, L. Fulton, S. Gattung, C. Geisel,
J. Kirsten, T. Kucaba, L. Hillier, M. Jier, L. Johnston,
Y. Langston, P. Latreille, E. J. Louis,* C. Macri, E. Mardis,
S. Menezes, L. Mouser, M. Nhan, L. Rifkin, L. Riles, H. St. Peter,
E. Trevaskis, K. Vaughan, D. Vignati, L. Wilcox, P. Wohldman,
R. Waterston, R. Wilson, M. Vaudin

The complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII reveals that it contains 269 predicted or known genes (300 base pairs or larger). Fifty-nine of these genes (22 percent) were previously identified. Of the 210 novel genes, 65 are predicted to encode proteins that are similar to other proteins of known or predicted function. Sixteen genes appear to be relatively recently duplicated. On average, there is one gene approximately every 2 kilobases. Although the coding density and base composition across the chromosome are not uniform, no regular pattern of variation is apparent.

To identify all of the genes that constitute a simple eukaryotic cell, an international collaborative effort is under way to determine the sequence of the *Saccharomyces cerevisiae* genome. This is an important goal because of the central importance of yeast as a model organism for the study of functions basic to all eukaryotic cells. The sequences of the first two yeast chromosomes to be completed (*1, 2*) have revealed that more than two-thirds of yeast genes have not been previously recognized and are thus novel, and the functions of more than half of these cannot be predicted, because they are not similar to proteins of known function. Here, we describe the DNA sequence of yeast chromosome VIII, which provides another 210 previously unrecognized genes and further illuminates features of yeast chromosome organization.

The sequence was determined (*3*) from the set of 23 partially overlapping phage λ and cosmid clones shown in Fig. 1 that were previously mapped by Riles *et al.* (*4*). The order of Hind III and Eco RI sites predicted from the sequence is consistent with the physical map of these sites determined independently by Riles *et al.* (*4*), which confirms that the sequence was assembled correctly. We estimate the accuracy of the sequence to be better than 99.99% (*5*). The genes and other features of the chromosome VIII sequence are listed in Table 1.

The sequence contains 269 nonoverlapping open reading frames (ORFs) greater than 300 base pairs (bp). On the basis of the analysis of Dujon *et al.* (*2, 6*), approximately 7% of these are likely to be false genes. Thirteen of these ORFs (4.8%) are predicted to be interrupted by introns at the extreme 5′ end of each gene. The average gene size is 482 codons; the longest ORF (YHR099w) spans 11,235 bp (3745 codons).

Table 1. List of genes and features of chromosome VIII. The number of the cosmid (as submitted to GenBank) and its accession number are listed above the elements included in that database entry. **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 is listed. **Column 2**: Genes are named according to established convention: Y designates yeast; H designates chromosome VIII; L and R designate the left or right chromosomal arm, respectively; w and c designate that the gene is encoded on the top or bottom strand, respectively; and a superscript "s" denotes genes predicted to be spliced. Genes are numbered from the CEN toward each TEL (telomere). Transfer RNA names also follow convention: t designates tRNA; the next letter is the one-letter code for the amino acid inserted by the tRNA (abbreviations for the amino acid residues are A, Ala; F, Phe; H, His; P, Pro; Q, Gln; S, Ser; T, Thr; and V, Val.); the letters in parentheses are the codon recognized by the tRNA; and w and c designate that the tRNA is on the top (w) or bottom (c) strand. Retrotransposon LTRs in brackets are partial elements. **Column 3**: Genetic names of genes previously identified. Note that one previously identified gene does not have a locus name (YHR042w) and that two genes (HXT5/YHR096c and ACT5/YHR129c) were named during the course of this work. **Column 4**: A description of the function of the genes. A description of the protein most similar to the other genes is also listed. Genes with no listing in this column have no homologs (BLASTX score usually less than 70). **Column 5**: The BLASTX (*18*) score for the alignment of the encoded protein to its closest homolog. Note that BLASTX scores are not listed for previously identified genes, because the two sequences are identical. BLASTX scores greater than 100 are generally considered to indicate a significant relation between two proteins; scores between 70 and 100 are considered suggestive of a relation. **Column 6**: Database accession number of the closest homolog. In the few cases where comparison of predicted proteins to the BLOCKS database (*19*) revealed potential similarities not found by BLAST, the number of the BLOCKS entry is given.

Fifty-nine of the genes (22%) were previously identified (that is, already present in the public databases). Another 65 of the ORFs (24%) are predicted to encode proteins that are similar to genes of known or predicted function (see Table 1 for a list). Thus, the function of only 46% of the encoded proteins is known or can be predicted (in some cases, only the biological process that the protein is involved in is

Genome Sequencing Center and Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA.

*Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, England.

Pos.	Gene or element	Locus	Function or homology	BLAST score	Acc. no.
			9196/U11583		
1	TEL		C(1-3)A	repeat	
36	Y' element		Y' subtelomeric repeat		
3310	YHL050c^{s}		Hyp. protein in Y' repeat region (pseudogene?)	1088	spIP24089I
4540	YHL049c		Hyp. protein in Y' repeat region (pseudogene?)	1371	pirIS31214I
5051	X element		X subtelomeric repeat		
6400	YHL048w		YKL219w	653	spIP36034I
7993	*Ty5* LTR				
10211	YHL047c		YKR106w; YCL070c; YCL071c; YCL073c	1372	spIP36173I
12283	YHL046c		Pau1p;YKL224c *et al.*; stress-induced proteins	583	gpIL25123I
12500	YHL045w		YCR103c; YKL223w	163	spIP25609I
13563	YHL044w		YCR007c	130	spIP25354I
14899	YHL043w		YKL219w	179	spIP36034I
15665	YHL042w		YKL219w	178	spIP36034I
17390	YHL041w				
20968	YHL040c		YKR106w	1456	gpIZ28202I
21780	YHL039w				
25506	YHL038c	***CBP2***	Cytochrome b pre-mRNA processing protein		gpIK00138I
26177	YHL037c				
26239	YHL036w		Amino acid permease	151	gpIL25068I
32754	YHL035c		Multidrug resistance protein (ABC transporter)	630	spIP36028I
34075	YHL034c	***SSB1***	Single-strand nucleic acid binding protein		spIP10080I
36023	YHL033c	***RPL4A***	60*S* ribosomal protein L7A-1, same as *MAK7*		spIP17076I
38506	YHL032c	***GUT1***	Glycerol kinase		spIP32190I
39484	YHL031c				
40082	YHL030w				
47966	YHL029c				
48761	YHL028w		Ser-Thr rich		
51109	YHL027w	***RIM1***	Pos. regulator of meiosis (Cys-His Zn fingers)		spIP33400I
54023	YHL026c				
			9433/U11582		
54848	YHL025w	***SNF6***	Transcription factor		spIP18888I
56646	YHL024w		RNA binding proteins	90	spIQ01130I
62560	YHL023c				
62752	tH(CUC)w		tRNA-His		
64154	YHL022c	***SPO11***	Sporulation protein		spIP23179I
65855	YHL021c				
67452	YHL020c	***OPI1***	Neg. regulator of phospholipid biosyn.		spIP21957I
69544	YHL019c		Clathrin coat associated protein AP54	156	spIQ00776I
69704	YHL018w		Dimerization cofactor of NF1-a	85	spIP80095I
70272	YHL017w		Probable transmembrane protein YKL039w	150	pirIS37739I
74240	YHL016c	***DUR3***	Urea active transporter		spIP33413I
75408	YHL015w		S10P family of 40*S* ribosomal proteins	337	spIP23403I
77310	YHL014c		Glycogen phosphorylase; GTP-binding protein	60	spIP00489I
78349	YHL013c				
78931	YHL012w		UDP-glucose pyrophosphorylase	228	spIP08800I
81611	YHL011c		Phosphoribosyl pyrophosphate synthetase	518	spIP11908I
83716	YHL010c				
			L5018/U11581		
85055	YHL009c		bZIP DNA-binding protein	124	spIP19880I
85367	tV(GUU)c		tRNA-Val		
85383	[sigma]				
85534	tau				
	Ty4				
91755	tau				
91767	delta				
92095	[delta]				
94505	YHL008c		Potential formate transporter NirC (*E. coli*)	62	spIP35839I
97932	YHL007c	***STE20***	Protein Ser-Thr kinase, pheromone response		gbIL04655I
98789	YHL006c				
99214	YHL005c				
			9780/U10555		
99213	YHL004w	***MRP4***	Mitochondrial ribosomal protein		spIP32902I
101877	YHL003c		Hypothetical protein YKL008c	1549	spIP28496I
102605	YHL002w		SH3 domain	151	spIP29354I
104270	YHL001w^{a}		Hypothetical protein YKL006w	677	spIP36105I
105579	CDEIII				
	CEN				
105689	CDEI				
106048	YHR001w		Hyp. prot. YKR003w; oxysterol-binding prot.	1596	spIQ02201I
108805	YHR002w		Mitochondrial carrier/Grave's disease prot.	192	gpIX66035I
111310	YHR003c		Hypothetical protein YKL027w	344	gpIZ28027I
113087	YHR004c				
114910	YHR005c	***GPA1***	G protein alpha subunit		spIP08539I
116172	tT(ACT)c		tRNA-Thr		
116745	delta				
117807	YHR006w		Zn finger protein (C2H2 type) Stp1p (yeast)	507	spIQ00947I
121676	YHR007c	***ERG11***	Cyto. P-450 L1 (Lanosterol 14-a-demethylase)		spIP10614I
			L2825/U10400		
123583	YHR008c	***SOD2***	Superoxide dismutase		spIP00447I
125658	YHR009c				
126513	YHR010w^{a}		Ribosomal protein L27	424	pirIS00401I
127772	YHR011w		Seryl-tRNA synthetase	369	gpIX75627I
129473	YHR012w^{a}				
131438	YHR013c	***ARD1***	Arrest-defective protein		spIP07347I
132038	YHR014w	***SPO13***	Meiosis-specific sporulation protein		spIP23624I
133099	tS(TCT)c		tRNA-Ser		
133665	delta				
134313	tQ(CAA)w		tRNA-Gln		
134545	YHR015w		Poly(A)-binding protein	627	gpID26442I
138446	YHR016c^{a}		SH3 domain in COOH-terminus	111	gpIX59932I
138685	YHR017w				
141393	YHR018c	***ARG4***	Arginosuccinate lyase		spIP04076I
			8082/U10399		
143549	YHR019c		Filarial antigen (nematode); Asp-tRNA-synthetase	937	gpIJ03266I
143987	YHR020w		Multifunctional aminoacyl tRNA-synthetase	956	spIP28668I
146305	tA(GCT)c		tRNA-Ala		
146322	sigma				
148660	YHR021c^{s}		40*S* ribosomal prot. S27; potential Zn finger	429	spIP35997I
150336	YHR022c		*RAS*-related protein	68	gpIU02928I
151657	YHR023w	***MYO1***	Myosin		spIP08964I
159183	YHR024c	***MAS2***	Mitochondrial processing peptidase		spIP11914I
159429	YHR025w	***THR1***	Homoserine kinase		gpIM37692I
160835	YHR026w	***PPA1***	Proteolipid protein of proton ATPase		spIP23968I
164702	YHR027c				
167425	YHR028c	***DAP2***	Dipeptidyl aminopeptidase B		spIP18962I
168552	YHR029c		Thymidylate synthase (putative)	112	gpIX59273I
			8179/U00062		
170335	YHR030c	***SLT2***	Protein Ser-Thr kinase		gpIX59262I
172961	YHR031c		Pif1p (mito. DNA repair/recomb. prot.)	388	spIP07271I
173335	YHR032w				
175539	YHR033w		Pro1p (gamma-glutamyl kinase)	997	spIP32264I
177990	YHR034c				
178210	YHR035w		Sec23p (yeast protein transport protein)	90	spIP15303I
180336	YHR036w				
181968	YHR037w	***PUT2***	P5C dehydrogenase		gpIU00062I
184057	YHR038w				
186800	YHR039c		Aldehyde dehydrogenase	159	spIP17445I
187915	YHR040w		Hit1p, required for high-temperature growth	98	pirIS30869I
189855	YHR041c^{s}	***SRB2***	Transcription factor		spIP34162I
190534	YHR042w		NADPH–cytochrome P-450 reductase		gpID13788I
193536	YHR043c				
194799	YHR044c				
195542	YHR045w				
198276	YHR046c		Inositol monophosphatase, QUTG protein	189	pirIS11944I
201301	YHR047c	***AAP1***	Ala-Arg aminopeptidase (Zn metalloprotease)		gbIL12542I
204598	YHR048w		Various drug resistance proteins	293	pirIJC1173I
206453	YHR049w				
207646	YHR050w		Smf1p (mitochrodrial membrane protein)	441	bbsI119299
209697	YHR051w	***COX6***	Cytochrome c oxidase subunit VI		spIP00427I
210840	YHR052w				
			8025/U00061		
212720	YHR053c	***CUP1***	Copper metallothionein		spIP07215I
214249	YHR054c		ORFX in *CUP1* repeat region		
214718	YHR055c	***CUP1***	Copper metallothionein		spIP07215I
217681	YHR056c		ORFX' (extended) in *CUP1* repeat region		
218844	YHR057c	***CYP2***	Peptidyl-prolyl cis-trans isomerase		spIP23285I
219885	YHR058c				
220109	YHR059w				
220726	YHR060w				
222479	YHR061c				
223759	YHR062c				
225170	YHR063c				
227244	YHR064c		Hsp70 heat shock protein	432	spIP22202I
229164	YHR065c		RNA helicase (DEAD box)	562	spIP34580I
229336	YHR066w				
230971	YHR067w				
232134	YHR068w				
234659	YHR069c		Hyp. protein upstream of abl (human)	275	gbIU07561I
234882	YHR070w				
237005	YHR071w		G1/S cyclin	74	spIP24867I
237940	tF(TTC)1c^{a}		tRNA-Phe		
237995	[delta]				
			9205/U10556		
239099	YHR072w	***ERG7***	Lanosterol synthase		gpIU04841I
242583	YHR073w		Oxysterol-binding protein	172	spIP22059I
246194	YHR074w		Spore outgrowth factor B (*B. subtilis*)	83	spIP08164I
249642	YHR075c				
251102	YHR076w				
255650	YHR077c		Highly acidic COOH-terminus		
256361	YHR078w				
261571	YHR079c	***IRE1***	Protein kinase		spIP32361I
266839	YHR080c				
267539	YHR081w				

Pos.	Gene or element	Locus	Function or homology	BLAST score	Acc. no.
271549	YHR082c		Protein Ser-Thr kinase	136	gp|M20487|
272628	YHR083w				
274175	YHR084w	***STE12***	Transcriptional activator		sp|P13574|
276765	YHR085w				
			9332/U00060		
278154	YHR086w	***NAM8***	RNA binding protein		gp|U00060|
280821	YHR087w				
281496	YHR088w				
283299	YHR089c	***GAR1***	snRNP required for pre-rRNA processing		sp|P28007|
284626	YHR090c				
286771	YHR091c		Arginyl-tRNA synthetase	472	sp|P11875|
288813	YHR092c	***HXT4***	Hexose transporter		sp|P32467|
289144	YHR093w				
292627	YHR094c	***HXT1***	Hexose transporter		sp|P32465|
292945	YHR095w				
296449	YHR096c	***HXT5***	Hexose transporter	576	sp|P32467|
298611	YHR097c[a]				
301936	YHR098c				
302763	YHR099w				
			8263/U00059		
314675	YHR100c				
315970	YHR101c[a]				
316574	YHR102w		Protein Ser-Thr kinase	325	sp|Q03497|
320416	YHR103w				
323411	YHR104w		Aldo-keto reductase	495	sp|P31867|
324768	YHR105w		Bact. reg. prot. (helix-turn-helix, arsR group)		BL00846
325600	YHR106w		Thioredoxin reductase	457	gp|Z23109|
328038	YHR107c	***CDC12***	Cell division cycle protein		sp|P32468|
328305	YHR108w				
330312	YHR109w				
332284	YHR110w		Glycoprotein 25L; involved in protein sorting?	149	sp|P27869|
333074	YHR111w		Molybdopterin biosynthesis protein moeB	313	sp|P12282|
335665	YHR112c		Cystathionine gamma-synthase	221	sp|P00935|
336339	YHR113w		Vacuolar aminopeptidase	249	sp|P14904|
338085	YHR114w		SH3 domain	100	sp|P27870|
341361	YHR115c				
341667	YHR116w				
342351	YHR117w		Mito. protein import receptor; TPR repeats	616	sp|P07213|
345624	YHR118c				
346045	YHR119w		Trithorax protein (COOH-terminus)	232	sp|P20659|
349576	YHR120w	***MSH1***	DNA mismatch repair protein		sp|P25846|
352758	YHR121w				
			9315/U10398		
353627	YHR122w				
354817	YHR123w[a]	***EPT1***	Ethanolaminephosphotransferase		sp|P22140|
356563	YHR124w				
358571	tF(TTC)2c[a]		tRNA-Phe		
358698	[delta]				
358861	YHR125w				
359081	[delta]				
360183	YHR126c		Tir2p (Cold shock-induced protein)	81	sp|P33890|
360915	YHR127w				
362012	YHR128w	***FUR1***	Uracil phosphoribosyltransferase		sp|P18562|
364155	YHR129c	***ACT5***	Actin-related protein; centractin	564	gp|Z14978|
365302	YHR130c				
367864	YHR131c		Highly acidic COOH-terminus		
369795	YHR132c		Carboxypeptidases	279	sp|P15089|
371597	YHR133c				
371749	YHR134w				
374310	YHR135c	***YCK1***	Casein kinase homolog I		sp|P23291|
375100	YHR136c				
375709	YHR137w				
377699	YHR138c				
379199	YHR139c	***SPS100***	Sporulation-specific wall maturation prot.		sp|P13130|
380575	YHR140w				
382751	YHR141c[a]	***RPL4B***	60*S* ribosomal prot. L41, same as *MAK18*		gp|D10578|
			9666/U10397		
383538	YHR142w				
385510	YHR143w		Ser-Thr rich		
388726	YHR144c	***DCD1***	dCMP deaminase		sp|P06773|
388995	tP(CCA)c[a]		tRNA-Pro; probable *SUF8* gene		
389337	YHR145c		(spans most of delta element)		
389509	delta				
390300	YHR146w				
393283	YHR147c	***MRP-L6***	Mitochondrial ribosomal protein L6		sp|P32904|
393534	YHR148w		40*S* ribosomal protein YS11 (YP28)	136	sp|P05755|
396659	YHR149c				
397251	YHR150w				
400848	YHR151c				
401434	YHR152w	***SPO12***	Sporulation protein		sp|P17123|
402682	YHR153c	***SPO16***	Sporulation protein		sp|P17122|
402966	YHR154w				
407103	YHR155w		Sip3p (Snf1p interacting protein)	363	gp|U03376|
412406	YHR156c				
412907	YHR157w	***REC104***	Meiotic recombination protein		sp|P33323|
417179	YHR158c				
417549	YHR159w				
420072	YHR160c		Aminopeptidase P & proline dipeptidase		BL00491
422286	YHR161c				
			9986/U00027		
423072	YHR162w		Rat brain 0-44 mRNA, segment 2	221	gp|M13095|
423630	YHR163w				
429177	YHR164c		DNA-binding prot. for G-rich single strands	147	gp|L14754|
436947	YHR165c	***PRP8***	U5 snRNP, pre-mRNA splicing factor		sp|P33334|
439049	YHR166c	***CDC23***	Cell division cycle protein		sp|P16522|
439341	YHR167w				
440376	YHR168w		GTP-binding proteins	214	sp|P20964|
442179	YHR169w		RNA helicase (DEAD box)	319	sp|P34580|
443826	YHR170w				
445710	YHR171w		Molybdopterin biosynthesis protein ChlN	141	sp|P12282|
448332	YHR172w				
451150	YHR173c				
451324	YHR174w	***ENO2***	Enolase 2 (2-phosphoglycerate dehydratase)		pir|A01148|
452869	YHR175w				
454226	YHR176w		Flavin-containing monooxygenase	97	gp|L10037|
456589	YHR177w				
459294	YHR178w		Zinc finger (6-Cys) protein	95	sp|P08657|
462497	YHR179w	***OYE2***	NADPH oxidoreductase (Old Yellow enzyme)		sp|Q03558|
			9186/U00028		
465173	YHR180w				
466528	delta				
466906	[sigma)				
466985	tT(ACA)w		tRNA-Thr		
467223	YHR181w				
468214	YHR182w				
470955	YHR183w		6-phosphogluconate dehydrogenase	800	gp|M80598|
472739	YHR184w				
			9998/U00030		
475335	YHR185c				
475782	tV(GTG)c		tRNA-Val		
480619	YHR186c				
480985	YHR187w				
483808	YHR188c				
484023	YHR189w				
484840	YHR190w	***ERG9***	Farnesyl-diphosphate farnesyltransferase		gb|X59959|
486626	YHR191c				
486821	YHR192w				
488231	YHR193c				
488652	YHR194w				
490742	YHR195w				
491926	YHR196w				
493891	YHR197w				
497275	YHR198c		YHR199c gene product	160	gp|U00030|
498417	YHR199c		YHR198c gene product	160	gp|U00030|
499074	YHR200w				
501138	YHR201c	***PPX1***	Exopolyphosphatase		gp|L28711|
502383	YHR202w				
505525	YHR203c[a]	***RPS7A***	Ribosomal protein S7		gp|M64293|
506314	YHR204w		Alpha-mannosidase	81	gp|U03458|
			9177/U00029		
509361	YHR205w	***SCH9***	cAMP-dependent protein kinase		gp|X57629|
512727	YHR206w		Heat shock transcription factor	239	sp|P10961|
516480	YHR207c				
517527	YHR208w		Teratocarcinoma protein	475	sp|P24288|
519432	YHR209w		Hyp. yeast prot. between DMC1-BMH1	158	gb|L11229|
521732	YHR210c		UDP-glucose-4-epimerase (GalE, Gal10p)	304	sp|P04397|
525387	YHR211w		Flo1p (flocculation prot.; *FLO8* gene?)	1075	sp|P32768|
538089	YHR212c		RAA19 gene on chr. I right arm (identical)	555	gp|L28920|
539146	YHR213w		Flo1p (flocculation protein)	653	sp|P32768|
541646	YHR214w				
543605	delta				
	Ty1				
549631	delta				
552094	YHR215w	***PHO12***	Acid phosphatase	2479	sp|P35842|
554391	YHR216w		IMP dehydrogenase (*PUR5*?)	1351	gp|L22608|
556098	X element		X subtelomeric repeat		
556640	Y' element		Y' subetelomeric repeat		
557037	YHR217c				
558009	YHR218w[a]		Hyp. protein in Y' repeat region (pseudogene?)	1871	sp|P24089|
560168	YHR219w		Hyp. protein in Y' repeat region (pseudogene?)	3143	pir|S28368|
562451	TEL		TG(1-3) repeat		

known). Nearly half of the ORFs (124, or 46%) are predicted to encode proteins that are not significantly similar to sequences in the puhlic databases. Finally, 21 genes (7.8%) are predicted to encode proteins that are similar to proteins of unknown function. Only two of these (YHR069c and YHR162w) are similar to gene products of other organisms; most of the rest (13 of 19) lie very near the ends of the chromosome, where large segments are extensively duplicated in analogous regions of other yeast chromosomes.

Eleven transfer RNA (tRNA) genes were identified, three of which are interrupted by introns. Nine of these are preceded by complete or partial copies of the long terminal repeats (LTRs) of yeast retrotransposons (six with partial or complete δ elements, one with a σ element, and two with a partial σ element and a complete δ element), which reside 14 to 566 bp upstream of the tRNA genes. Except for the two δ sequences that are part of the *Ty1* element on the right arm of the chromosome, all δ elements are associated with tRNA genes, as are the three complete or partial σ elements. The close association of these retrotransposon LTRs with tRNA genes is a general feature of the yeast genome (7). Four complete or partial τ sequences, two of which are associated with a *Ty4* element on the left arm and one *Ty5* LTR (8) were also identified.

The *CUP1* gene, encoding copper metallothionein, is contained in a 1998-bp repeated sequence that also includes an ORF of unknown function upstream of *CUP1* (YHR054c, previously called ORFX). The repeated region has been estimated to span 29.9 kb in the strain we used (4), which would encompass 15 repeats, hut the number of repeats varies among yeast strains (9). We sequenced into the repeat region from each end and determined the sequence of one complete repeat. However, because the ORF upsttеam of *CUP1* continues into unique sequence in the first copy of the repeat [the right, or centtomere (CEN) distal copy], we included two copies of the repeat in the final sequence in order to include this novel ORF (YHR056c). Thus, the sequence includes two copies of the *CUP1* gene (YHR053c and YHR055c).

The coding sequence comprises 69.2% of the chromosome, with one gene every 2087 bp. The average distance between genes is 629 bp, with differences in the spacing between genes with divergent promoters (731 bp) and genes with convergent tetminators (479 bp). There are more genes on the top strand (*10*) [144 on the top (w) strand and 124 on the bottom (c) strand], but nearly all the excess w strand genes are accounted for by a stretch of apptoximately 35 kb where 17 of the 18 ORFs are arrayed on the top sttand (cootdinates 439341 to 474454). Disregarding this unusual cluster of genes, there are nearly equal numbers of genes on each strand. These properties of the sequence are similar to those found for the two yeast chtomosomes pteviously sequenced (*1*, *2*).

The base composition of the chromosome is clearly not uniform over its length (Fig. 2, A and B): there are two major G+C–rich peaks toward the left end of the chromosome and several minor peaks in the right half of the chromosome. On the hasis of statistical analysis, we are confident that at least the two major G+C–rich peaks and the one major G+C–poor peak in the left half of the chromosome are significant (*11*). A similar degree of nonuniformity in base

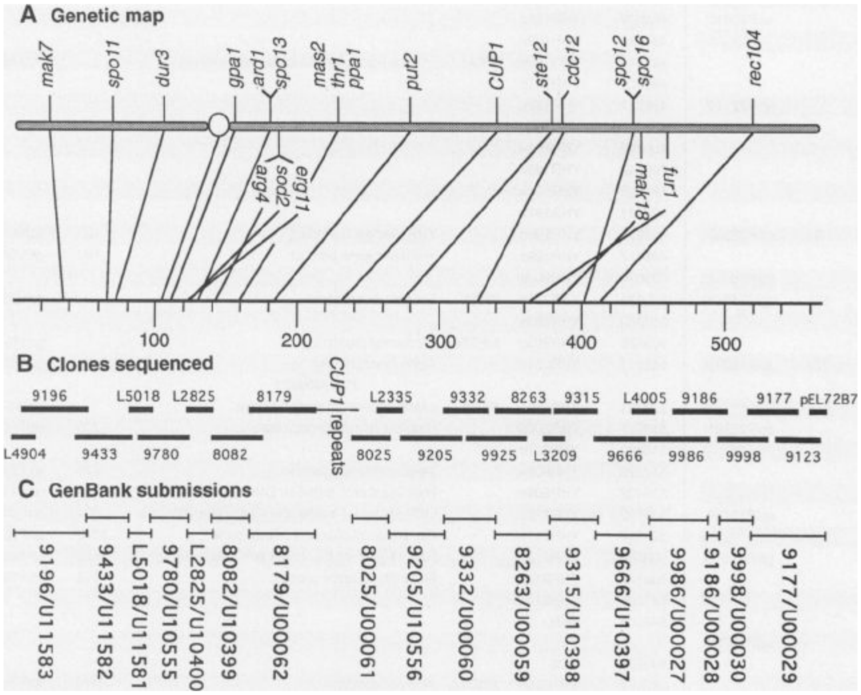

Fig. 1. Genetic and physical map of chromosome VIII. **(A)** Genetic map of the loci identified in the DNA sequence. The true location of these genes is indicated by lines connecting them to the scale (in base pairs). Note the two minor discrepancies in the genetic map. **(B)** Physical map of cosmid and phage λ clones used to determine the sequence. **(C)** Map of the extent of DNA sequence included in each GenBank entry. The GenBank entry name and accession number are listed below each line. In addition, the entire (nonoverlapping) sequence (562,638 bp) is available via anonymous ftp (genome-ftp.stanford.edu in the /pub/yeast/genome_seq/chrVIII directory; ncbi.nlm.nih.gov in the /repository/yeast/CHVIII directory; mips.embnet.org in the /anonymous/yeast/chrviii directory).

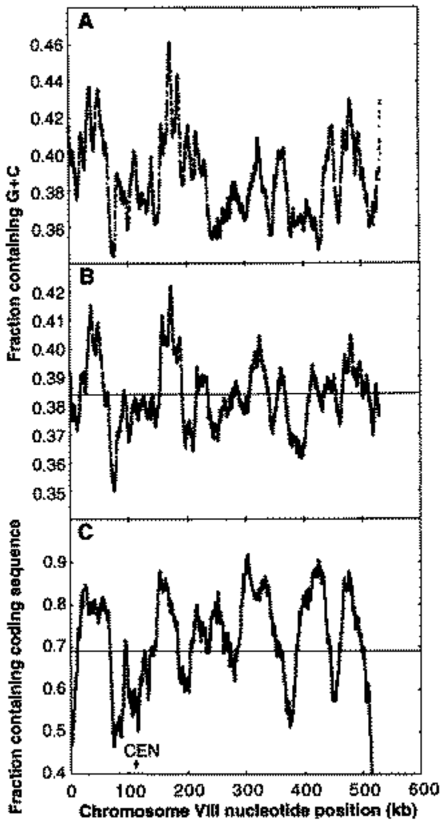

Fig. 2. Plot of coding density and G+C composition over the length of chromosome VIII. **(A)** G+C composition of the third base of codons in predicted ORFs was calculated over 20-kb windows spaced every 100 bp. **(B)** Overall G+C composition was calculated over 20-kb windows spaced every 100 bp. The horizontal line marks the average G+C composition (38.45%). **(C)** Coding density was calculated over 20-kb windows spaced every 100 bp. The horizontal line marks the average coding density (69.2%). For all three plots, similar results were obtained if the window size was varied between 10 and 50 kb or if the window size was the next 15 ORFs.

composition was observed for chromosomes III and XI (*2*, *12*). Although the regional variations in chromosome XI seem to occur in an almost regular pattern, those in chromosome VIII appear less regular. Thus, a regular periodicity of base composition does not appear to be a universal feature of yeast chromosomes. These base composition and gene density variations could be of functional importance (that is, having to do with processes such as replication or chromosome packaging) or could reflect the evolutionary history of the chromosome.

Similarly, the amount of protein coding sequence is not uniformly distributed over the length of chromosome VIII: there are six or seven regions of the chromosome with a coding density that is higher than average (Fig. 2C), a phenomenon also noted for chromosome XI (*2*). Perhaps not surprisingly, the G+C–rich regions correlate roughly, though certainly not precisely, with the regions of increased coding density, as was also noted for chromosome XI (*2*).

Several regions of chromosome VIII are duplicated on chromosomes I, III, or XI. The most extensive of these is an approximately 30-kb region very near the right telomere (bases 525393 to 555891) that is more than 90% identical to the similar region on the right arm of chromosome I. In addition, a smaller portion of this region of the right arms of chromosomes I and VIII is also duplicated on the left arm of chromosome I (*13*). This duplication, which was previously recognized (*14*), includes six genes whose order and orientation are preserved in the two copies. A *Ty1* element present in the duplicated region of chromosome VIII was probably originally present and subsequently lost from the homologous region of chromosome I, because chromosome I retains one of the LTRs of the retrotransposon at this location. A remarkable feature of this duplication is that its borders coincide almost precisely with the coding sequence (YHR211w at the left border and YHR216w at the right border). In addition, the high degree of sequence conservation between these regions of chromosomes I and VIII extends through a noncoding sequence, which suggests that this is a relatively recent duplication. Alternatively, the duplication could be more ancient, but extensive enough for the duplicated regions to pair infrequently in mitosis or meiosis and to be homogenized by gene conversion. A few other comparable duplications have been recognized on other yeast chromosomes (*10*, *15*).

There are also several shorter duplicated segments of the subtelomeric region of the left arm of chromosome VIII at analogous positions of chromosomes III and XI. [This is in addition to the X and Y' subtelomeric repeats, which are present at the ends of nearly all yeast chromosomes (*7*, *16*).] These duplicated segments, which are scattered throughout the region between coordinates 5000 and 13000, vary in identity from about 54 to about 94% and are largely limited to four ORFs (YHL045 to YHL048).

Six other individual genes on chromosome VIII appear to be recently duplicated. This is clearly recognizable at the DNA level [BLASTN score cutoff of 300 (*17*)], in contrast to duplications of clearly older origin, which can be recognized only at the protein level. In each case, the duplicated sequences are confined to nearly the entire coding region of the duplicated gene. Four of the duplicated genes (YHL003c, YHL001w, YHR001w, and YHR003c) reside near the centromere, and three of the four homologs of these genes (YKL008c, 70% identical to YHR003c; YKL006w, 96% identical to YHL001w; and YKR003w, 72% identical to YHR001w) are also very near the centromere of chromosome XI [the other homolog is also on chromosome XI but is somewhat distant from the centromere, and the duplication is much less extensive and much less conserved (YKL027w, 57 to 63% identical to YHR003c over less than half the length of these genes)]. Two other duplicated genes (YHL047w and YHR021c) are dispersed on chromosome VIII, though homologs (YKL156w and YKL157w, respectively) are adjacent on chromosome XI. Thus, a total of 16 genes on chromosome VIII appear to be recently duplicated. In addition, another obvious case of less recent gene duplication on chromosome VIII is a cluster of three hexose transporter genes (YHR092c/HXT4, YHR094c/HXT1, and YHR096c/HXT5). The amount of redundancy recognized in the yeast genome will undoubtedly grow as the sequence of additional chromosomes becomes available.

We imagine two ways these duplications could have arisen. First, some of these genes could represent processed genes that were inserted into the genome relatively recently, a view that is consistent with the conservation of sequence only in the coding regions. However, all of these cases would appear to be created by integration of full-length complementary DNAs, because none appear to be pseudogenes and this is unexpected in this model. In addition, one of the homologous gene pairs includes introns in both genes (which are 63% identical; their exons are 96% identical), which suggests that at least these genes were not duplicated by this mechanism. Alternatively, the clustering of four of the duplicated genes near the centromeres of their respective chromosomes compels us to consider the idea that entire genomic regions were duplicated. This centromeric duplication would appear to be ancient, because the DNA sequence has clearly diverged outside the coding regions, but the high degree of DNA sequence conservation in the coding region would appear to be at odds with this view.

Analysis of the sequence of chromosome VIII corroborates our current view of the organization of yeast chromosomes. The high coding density and close spacing of genes on chromosome VIII is similar to that of the other two yeast chromosomes sequenced, and the degree of genetic redundancy is also similar. However, the apparent organization of chromosome XI into regularly spaced intervals of G+C–rich and G+C–poor segments does not appear to hold for chromosome VIII, making the generality of this phenomenon unlikely. The most immediate and wide-ranging impact of this work is likely to be the identification of the 210 novel genes found on chromosome VIII, most of which we are unable to predict a function for at the present time. The sophisticated genetic techniques available for manipulating yeast cells provide the possibility of determining the function of many of these genes. It seems certain that *S. cerevisiae* will become even more important for understanding the function of eukaryotic cells as the sequence of more chromosomes is made available to the scientific community by the several groups collaborating internationally to complete the sequence of the entire yeast genome.

REFERENCES AND NOTES

1. S. G. Oliver *et al.*, *Nature* **357**, 38 (1992).
2. B. Dujon *et al.*, *ibid.* **369**, 371 (1994).
3. The clones sequenced all originate from strain AB972, which is derived from the common laboratory strain S288C (*4*). The sequence of the entire yeast DNA insert of each cosmid clone was determined. We sequenced the yeast DNA inserts in the phage λ clones after converting them into plasmids by recombination in yeast [J. Erickson and M. Johnston, *Genetics* **134**, 151 (1993)]. Gaps that exist between two pairs of cosmid clones and between a cosmid clone and the left end of the *CUP1* repeat were short enough to be recovered as polymerase chain reaction (PCR) products, using as a template the clones that span the gaps (λ 3209 and 4005 and cosmid 9181), which were then sequenced in their entirety. Finally, the sequence of the extreme right end of the chromosome, including the telomere, was determined mined from a plasmid clone generated by integration at the TG_{1-3} repeats of the telomere, followed by excision of the plasmid and capture of the flanking sequences (E. Louis, unpublished results). The details of the sequencing strategy have been described elsewhere [R. Wilson *et al.*, *Nature* **368**, 32 (1994)]. Briefly, 1- to 2-kb sheared fragments of the substrate DNA (cosmid, plasmid, or PCR product) were subcloned into M13 and sequenced on automated fluorescent DNA sequencing machines with universal primer. The sequence was assembled into contigs after 600 to 800 random subclones of each cosmid (fewer for the smaller λ clones and PCR products) had been sequenced (approximately six- to eightfold redundancy in the data). At this point, a directed sequencing strategy was used to join contigs, to sequence regions not represented on both strands, and to resolve discrepancies in the sequence. The sequence of both strands of each clone was determined (the sequence of overlapping re-

gions of cosmids was finished for only one clone), and all ambiguities in the sequence were resolved betore the sequence of a clone was considered finished. The finished sequences were compared with the public sequence databases for protein and nucleic acid homologies [SWISSPROT (retease 28.0), PIR (retease 40.0), and GENPEPT (release 82.0)], with BLASTX (for protein similarities) and BLASTN (for nucleotide simrarities) (*18*) and searched for tRNAs with TRNASCAN [G. Fichanl and C. Burks, *J. Mol. Biol.* **220**, 659 (1991)]. The sequence ot each cosmid was atso compared lo the yeast sequences in GenBank, and discrepancies were examined in our sequence and corrected whan possible (however, we judged that very tew of these differences were dua to mistakas in our sequence). The finished saquances wera assembted and interectively annotated with AScDB, a version of the *Caenorhabditis elegans* database program ACeDB (R. Durbin and J.-T. Mieg, unpublished rasults) modified (by E. Sonhammer and R. Durbin and L. Hilfer) for use with yeast data. Al this point, any polenlial tramestiift errors were recognized, and the appropriate regions were resequenced to resolve the problems. Portions ot the chromosome (usualty individual cosmids) were submilled to GenBank, as shown in Fig. 1 (enlry names and accession numbers are atso lisled in Table 1). Only a smalt number ot overlapping bases were included in each database entry lo facilitate joining ot the sequences or to keep a gane intact. In addition, lhe entire (nonoverlapping) 562,638 bp of DNA lhat comprise chromosome VIlt are available via anonymous fite transfer prolocol (ftp) (genome-ftp.slantord.edu in the direclory: /pub/yeast/genome_saq/chrVIIt; ncbi.ntm.nih.gov in the directory: /repository/yeast/CHVIII). Alt ORFs containing at least 100 codons (including the ATG and translation termination codons) were identified. This anelysis was done in batch with two scripts (ASCPREP1 and ASCPREP2; L. Hiltier, unpubtished resutts) that prepara the sequence and the database search results for entry into AScDB, which was used interactivaly to ennotate tha sequence. Genes were chosen with the help of the GENEFINDER program (P. Green and L. Hitler, unpublished results) modifred (by L. Hittier, E. Sonhammer, and R. Durbin) tor use with *S. cerevisiae*. Alt genes larger than 100 codons were annolated, except in the case of overlapping genes, where the longest gene or tha gene that had homology to another gene was chosen. The first ATG codon in an ORF was always chosen as the beginning ot lhe gene. Splice sites were used as necessary and when possible to construct a gene; a TACTAAC box 5 to 134 basas upstream of tha 3′ splice site [B. C. Rymond and M. Rosbash, in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, E. Jones, J. Pringte, J. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), vol. 2, pp. 143–192] was damanded in each case. We sought delta (δ), sigma (σ), and tau (τ) alements by comparing the sequence using BLASTN and FASTA against a representative member of each element.

4. L. Riles *et al.*, *Genetics* **134**, 81 (1993); L. Riles and M. Olson, unpubtished resutts.
5. This is a conservative accuracy astimala based on our analysis ot the yeast sequence as welt as ot the *C. elegans* sequence lhal has been determined in our sequencing center. We identified mistakes in the yeast sequence by comparing our sequence to sequences already in GenBank and by recognizing apparenl frameshift errors. In 425 kb of yeasl sequence checked in this way, 24 potential errors were identified (two by comperison to sequences in GenBank and 22 by recognition ot apparant framashifts)—approximately one error in 17 kb (most ot lhese errors were corrected). An independent comparison of 17,208 bp of *C. elegans* sequence to an independently datermined sequenca already in GanBank ravealed one error (L. Hillier, unpubtished results), corroborating our estimate ot approximalely one mistake per 17 kb.
6. B. Dujon, personat communicalion.
7. M. V. Olson, in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, E. Jones, J. Pringla, J. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), vol. 1, pp. 1–40.
8. D. F. Voytas and J. D. Boeke, *Nature* **358**, 717 (1992).
9. R. R. Butt and D. J. Ecker, *Microbiol. Rev.* **51**, 351 (1987).
10. The "top" strand refers to tha strend with polarity 5′ to 3′ (left to right) of the chromosome as oriented (according to convention) on the genetic map of R. K. Mortimer *et al.* [*Yeast* **8**, 817 (1992)].
11. The dislribution ot G+C contant tor chromosoma VIIt was tound lo be statistically different (α > 0.01) from that ot a random sequenca with the same nucleotide content. Further, tha analysis confirmed that the three major peaks in tha chromosome VIIt G+C content plots are significantly different trom that of the random sequence (three to four times as many standard deviations from the mean as peaks in the random sequence) (L. Hillier and G. Marth, in preparation).
12. P. M. Sharp and A. T. Lloyd, *Nucleic Acids Res.* **21**, 179 (1993).
13. H. Bussey, personal communication.
14. H. Y. Steensma, P. De Jonge, A. Kaptein, D. B. Kaback, *Curr. Genet.* **16**, 131 (1989).
15. D. Lalo, S. Stettier, S. Mariotte, P. Slonimski, P. Thuriaux, *C. R. Acad. Sci. Paris* **316**, 367 (1993).
16. E. J. Louis, E. S. Naumova, A. Lee, G. Naumov, J. E. Haber, *Yeast* **10**, 271 (1994).
17. These comparisons wera performed at the National Center for Biotechnology Information with the BLAST natwork service.
18. S. F. Altschul, W. Gish, W. Millar, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
19. S. Hanikoff and J. G. Henikoff, *Genomics* **19**, 97 (1994).
20. Wa thank H. Bussey for providing tha sequence of yeast chromosome I and E. Sonhemmar and R. Durbin tor modifing GENEFINDER for use with yeast data. Supported by a grant from the NIH National Center for Human Genome Research. E.J.L. received support from the Wellcome Trust.

16 August 1994; accepted 1 Septembar 1994

Specific Cleavage of Model Recombination and Repair Intermediates by the Yeast Rad1-Rad10 DNA Endonuclease

A. Jane Bardwell,*† Lee Bardwell,*‡ Alan E. Tomkinson, Errol C. Friedberg§

The *RAD1* and *RAD10* genes of *Saccharomyces cerevisiae* are required for both nucleotide excision repair and certain mitotic recombination events. Here, model recombination and repair intermediates were used to show that Rad1-Rad10–mediated cleavage occurs at duplex–single-strand junctions. Moreover, cleavage occurs only on the strand containing the 3′ single-stranded tail. Thus, both biochemical and genetic evidence indicate a role for the Rad1-Rad10 complex in the cleavage of specific recombination intermediates. Furthermore, these data suggest that Rad1-Rad10 endonuclease incises DNA 5′ to damaged bases during nucleotide excision repair.

The *S. cerevisiae RAD1* and *RAD10* genes are involved in both nucleotide excision repair (*1*) and mitotic recombination (*2*–*9*). *RAD1* is the probable homolog of the human *XPF* (*ERCC4*) gene, which is defective in the cancer-prone disease xeroderma pigmentosum (*10*, *11*); *RAD10* is homologous to human *ERCC1* (*12*). Rad1 and Rad10 proteins form a stable complex (*13*, *14*) that catalyzes the endonucleolytic degradation of single-stranded bacteriophage DNA but is inactive on linear duplex DNA (*15*, *16*). Here we demonstrate that rather than exhibiting a generalized single-strand DNA endonuclease activity as previously indicated (*15*, *16*), Rad1-Rad10 protein is a duplex–3′ single-sttand junction-specific endonuclease. The characterization of this structure-specific activity greatly clarifies the role of Rad1-Rad10 protein in recombination and DNA repair.

Single-stranded, duplex, or partial duplex model DNA substrates were generated from synthetic oligonucleotides 18 to 50 nucleotides in length (Table 1). Rad1-Rad10 endonuclease did not degrade a single-stranded 49-nucleotide oligomer (S1 in Table 1 and Fig. 1, A and B) or a 49–base pair (bp) duplex structure (D in Table 1 and Fig. 2, A and B). However, when S1 was annealed to shorter complementary oligonucleotides, partial duplex molecules containing 3′ single-stranded tails (TD1 and TD2 in Table 1) were cleaved by the enzyme (Fig. 1A), whereas substrate TD3 (Table 1) containing a 5′ single-stranded tail was not (Fig. 1A). In a similar manner, substrate S3 (Table 1) was not cleaved as a single-stranded oligonucleotide (Fig. 2B), nor as a partial duplex derivative with a 5′ single-stranded tail (TD4 in Table 1 and Fig. 1A). A partial duplex derivative with a 3′ tail was cleaved (TD5 in Table 1 and Fig. 1A).

Analyses with denaturing gels demon-

A. J. Bardwelt, L. Bardwell, E. C. Friedberg, Laboralory ot Molecular Pathology, The University of Texas Southwestarn Medical Center at Dallas, Dallas, TX 75235, USA.
A. E. Tomkinson, Institule tor Biotechnology, Cenler for Molecular Medicine, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78245, USA.

*These authors contributed equatty to lhis sludy.
†Presenl address: Genelabs Technologies Inc., 505 Penobscot Driva, Redwood City, CA 94063, USA.
‡Presenl address: Departmenl of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA.
§To whom correspondence should be addressed.