Complete Nucleotide Sequence of
Sacccharomyces 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 492 codons; the longest ORF (YHR099w) spans 11,235 bp (3745 codons).

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
known). Nearly half of the ORFs (124, or 46%) are predicted to encode proteins that are not significantly similar to sequences in the public 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 3' elements, one with a 5' element, and two with a partial 5' element and a complete 3' element), which reside 14 to 566 bp upstream of the tRNA genes. Except for the two 3' sequences that are part of the Ty1 element on the right arm of the chromosome, all 3' elements are associated with tRNA genes, as are the three complete or partial 5' elements. The close association of these retrotransposon LTRs with tRNA genes is a general feature of the yeast genome (7). Four complete or partial 5' sequences, two of which are associated with a Ty1 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, but 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 upstream of CUP1 continues into unique sequence in the first copy of the repeat [the right, or centromere (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 terminators (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 approximately 35 kb where 17 of the 18 ORFs are arrayed on the top strand (coordinates 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 chromosomes previously 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 basis 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 composition is clear from the density plot of the overall length of chromosome VIII (Fig. 2, A): G+C composition of the third base of codons in predicted ORFs was calculated over 20-kb windows spaced every 100 bp. 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%). The 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.

![Fig. 1. Genetic and physical map of chromosome VIII.](image-url)

**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 cosmids 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 (genomes-ftp.stanford.edu in the /pub/yeastgenome_seq/chvili directory; ncbi.nlm.nih.gov in the /repository/yeast/chvili directory; mips.embnet.org in the /anonymous/yeast/chvili directory).

![Fig. 2.](image-url)

**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 (YHL036c, YHL0101, 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 YKR003c, 72% identical to YHR001w) are also very near the centromere of chromosome XI (the other homologs in the chromosome pair are 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). The other two 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 would 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, and which 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 65% 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 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 valuable 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

3. The clones sequenced all originate from strain AB1772, 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 cosmids and between a cosmids clone and the left end of the CUP7 repeat were short enough to be recovered as polymerase chain reaction (PCR) products, using as a template the clones that span the gaps (λ 5209 and 4005 and cosmids 9181), which were then sequenced in their entirety. Finally, the sequence of the extreme right end of the chromosome, including the telomeres, was determined from a plasmid clone generated by integration at the TG₃₃ repeats of the telomere followed by deletion of 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-
Specific Cleavage of Model Recombination and Repair Intermediates by the Yeast Rad1-Rad10 DNA Endonuclease

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

10. The “top” strand refers to the strand with polarity 5’ to 3’ (left to right of the chromosome as oriented (according to convention) on the genome map of R. K. Mortimer et al. [Yeast 8, 817 (1992)].
11. The distribution of G+C content for chromosome VIII was found to be statistically different (p < 0.01) from that of a random sequence with the same nucleotide content. Further, the analysis confirmed that the three major peaks in the chromosome VIII G+C content plots are significantly different from that of the random sequence (three to four times as many standard deviations from the mean as peaks in the random sequence) (L. Hiller and G. Martin, in preparation).
17. These comparisons were performed at the National Center for Biotechnology Information with the BLAST network service.
20. We thank H. Bussey for providing the sequence of yeast chromosome I and E. Sonhammer and R. Durbín for modifying 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 September 1994

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-strand 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-