molecular weight. The chromosome IX band was excised under long-wave ultraviolet transillumination to minimize DNA damage. Chromosomal DNA was purified by melting and phenol extraction, sonicated and end repaired. Two libraries were prepared: fragments 1.4–2 kb in length were cloned into M13mp18, and fragments 6–9 kb long were cloned into the phagemid vector pBS. Over 10,000 independent M13 clones were sequenced and assembled into a database using the program XBAP[19]. Sequencing strategy and methods used for sequence assembly are as described[19–25]. The lambda-clone consensus sequences were also entered into this database, which contained several thousand contigs at this stage, most of which contained a single gel read. We concluded that the chromosome IX DNA preparation was approximately 30% contaminated with DNA from other chromosomes, and that this was the source of most single-read contigs. This contamination, together with repetitive sequences in the database, caused many problems with the data assembly.

To overcome this problem, all single-read contigs were removed from the working copy of the chromosome IX database and collected in a secondary database. As further data was generated, the secondary database was periodically rescreened, and single reads were re-entered if they found matches in the primary database. This reduced the number of reads in the primary database to approximately 7,000, which represented coverage of the chromosome five times over. The database still contained several hundred contigs. At this stage a cosmid library covering most of chromosome IX became available[17]. The chromosomal 'shotgun' data were 'seeded' with reads from cosmid clones selected to give coverage over regions not previously sequenced by lambda clones. This approach also allowed the chromosome to be split up into manageable sections to solve double-stranding and compression problems. A minimal 'shotgun' of 300–500 reads was performed on each cosmid clone. Data from these cosmids were entered into the chromosome IX 'shotgun' database to contiguate the entire chromosome, and into separate cosmid databases for ease of handling, together with overlapping reads from the whole-chromosome shotgun. Each cosmid-sized project was contiguated, double stranded and all compressions were resolved.

Three regions of the chromosome remained unrepresented in either cosmid or lambda libraries: the left and right telomeres, and a region near the centre of the chromosome flanked by lambda clones 6569 and 3299. The right telomere was sequenced by primer walking using a plasmid clone[26]. The left telomere was finished using data from the whole-chromosome 'shotgun' and some primer walking from polymerase chain reaction (PCR) products generated from PFGE-purified chromosome IX DNA. The gap near the centre of the chromosome was filled using data from the whole-chromosome 'shotgun' and by sequencing a 1 kb fragment generated by PCR from genomic DNA. The gap between the lambda clones 6569 and 3299 was approximately 7 kb.

Received 24 July 1996; accepted 16 January 1997.

1. Davison, A. J. *DNA Sequence* 1, 389-394 (1991).
2. Telford, E.A.R. *et al. Virology* 189, 304-316 (1992).
3. Rawlinson, W.D. *et al. J. Virol* 70, 8833–8849 (1996).
4. Fleischmann, R. D. *et al. Science* 269, 496–512 (1995).
5. Fraser, C. M. *et al. Science* 270, 397–403 (1995).
6. Termier, M. & Kalogeropoulos, A. *Yeast* 12, 369–384 (1996).
7. Seufert, W. *Nucleic Acids Res.* 18, 3653 (1990).
8. Oliver, S. G. *et al. Nature* 357, 38–46 (1992).
9. Bussey, H. *et al. Proc. Natl Acad. Sci. USA* 92, 3809–3813 (1995).
10. Murakami, Y. *et al. Nature Genet.* 10, 261–268 (1995).
11. Fitzgerald-Hayes, M. *Yeast* 3, 187–200 (1987).
12. Mortimer, R. K. *et al.* http://genome.www.stanford.edu/saccdb/edition12.html (1995).
13. Cooper, T. G., Gorski , M. & Turoscy, V. *Genetics* 92, 383–396 (1979).
14. Turoscy, V., Chisholm, G. & Cooper, T. G. *Genetics* 108, 827–831 (1984).
15. Sharp, P. M. & Lloyd, A. T. *Nucleic Acids Res.* 21, 179–183 (1993).
16. Vaudin, M. *et al. Nucleic Acids Res.* 23, 670–674 (1995).
17. Riles, L. *et al. Genetics* 134, 81–150 (1993).
18. Schwartz, D. C. & Cantor, C. R. *Cell* 37, 67–75 (1984).
19. Dear, S. & Staden, R. *Nucleic Acids Res.* 19, 3907–3911 (1991).
20. Craxton, M. *Methods: A Companion to Methods in Enzymology* Vol. 3 (ed. Roe, B.) 20–26 (Academic, San Diego, 1991).
21. Halloran, N., Du, Z. & Wilson, R. K. in *Methods in Molecular Biology* Vol. 10, *DNA Sequencing: Laboratory Protocols* (eds Griffin, H. G. & Griffin, A. M.) 297–316 (Humana, Clifton, NJ, 1992).
22. Smith, V. *et al. Methods Enzymol.* 218, 173–187 (1993).
23. Hawkins, T. *DNA Sequence* 3, 65–69 (1992).
24. Staden, R. *Methods Mol. Biol.* 25, 27–36 (1994).
25. Staden, R. *Methods Mol. Biol.* 25, 37–67 (1994).
26. Louis, E. *Biochemica* 3, 25–26 (1995).

Acknowledgements. We thank A. Fraser for cosmid DNA preparation; the subcloning group for lambda and cosmid library preparation; the staff in the gel pouring and media kitchens for their help; the staff in the computer support and software development groups and R. Staden for software support; L. Riles and M. Olsen for lambda clones and cosmids; E. Louis for telomeric plasmids; B. Dujon for providing Fig. 1, which was prepared using unpublished software developed by C. Marck in collaboration with B. Dujon; and our colleagues for proof reading and criticism of this manuscript.This work was funded by the Medical Research Council and the Wellcome Trust.

Correspondence and requests for materials should be addresed to B.B. (e-mail: barrell@sanger.ac.uk). Clone accession numbers and other information can be found on http://www.sanger.ac.uk/yeast/home. html.

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII

M. Johnston[1], L. Hillier[1], L. Riles[1], other members of the Genome Sequencing Center[1], K. Albermann[2], B. André[3], W. Ansorge[4], V. Benes[4], M. Brückner[5], H. Delius[6], E. Dubois[7], A. Düsterhöft[8], K.-D. Entian[9], M. Floeth[8], A. Goffeau[10], U. Hebling[6], K. Heumann[2], D. Heuss-Neitzel[8], H. Hilbert[8], F. Hilger[11], K. Kleine[2], P. Kötter.[9], E. J. Louis[12], F. Messenguy[13], H. W. Mewes[2], T. Miosga[14], D. Möstl[8], S. Müller-Auer[2], U. Nentwich[4], B. Obermaier[15], E. Piravandi[15], T. M. Pohl[16], D. Portetelle[11], B. Purnelle[10], S. Rechmann[4], M. Rieger[5], M. Rinke[15], M. Rose[9], M. Scharfe[17], B. Scherens[18], P. Scholler[19], C. Schwager[4], S. Schwarz[19], A. P. Underwood[12], L. A. Urrestarazu[3], M. Vandenbol[11], P. Verhasselt[20], F. Vierendeels[13], M Voet[20], G. Volckaert[20], H. Voss[4], R. Wambutt[17], E. Wedler[17], H. Wedler[17], F. K. Zimmermann[14], A. Zollner[2], J. Hani[2] & J. D. Hoheisel[20]

[1] The Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, 630 S. Euclid Avenue, St. Louis, Missouri 63110, USA
[2] Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany
[3] Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles - Campus Plaine CP 244 Boulevard du Triomphe, B-1050 Bruxelles, Belgium
[4] EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany
[5] Genotype, Klingenstrasse 35, D-69434 Hirschhorn and Angelhofweg 39, D-69259 Wilhelmsfeld, Germany
[6] DNA Sequencing, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany
[7] Laboratoire de Microbiologie de l' Université Libre de Bruxelles, B-1070 Brussels, Belgium
[8] QIAGEN GmbH, Max-Volmer-Strasse 4, D-40724 Hilden, Germany
[9] Johann Wolfgang Goethe-Universität Frankfurt, Institut für Mikrobiologie, Marie-Curie-Strasse 9; Geb. N250, D-60439 Frankfurt/Main, Germany
[10] Unité de Biochimie Physiologique, Faculté des Sciences Agronomiques, Université Catholique de Louvain, Place Croix du Sud, 2-20, B-1348 Louvain-la-Neuve, Belgium.
[11] Faculté Universitaire des Sciences Agronomiques, Unité de Microbiologie, 6, avenue Maréchal Juin, 5030 Gembloux, Belgium
[12] Department of Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK
[13] Research Institute of the CERIA-COOVI, B-1070 Brussels, Belgium
[14] Institut für Mikrobiologie und Genetik, Schnittspahnstrasse 10, D-64287 Darmstadt, Germany
[15] MediGene GmbH, Lochhamer Strasse 11, 82152 Martinsried, Germany
[16] GATC Gesellschaft für Analyse-Technik und Consulting mbH, Fritz-Arnold-Strasse 23, 78467 Konstanz, Germany
[17] AGON GmbH, Glienicker Weg 185, D-12489 Berlin, Germany
[18] Laboratorium voor Erfelijkheidsleer en Microbiologie van de Vrije Universiteit Brussel Vlaams Interuniversitair Instituut voor Biothechnologie, Departement Microbiologie, Avenue E. Gryson 1, B-1070 Brussels, Belgium
[19] Katholieke Universiteit Leuven, Laboratory of Gene Technology, Willem de Croylaan 42, B-3001 Leuven, Belgium
[20] Molecular Genetic Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany

The yeast *Saccharomyces cerevisiae* is the pre-eminent organism for the study of basic functions of eukaryotic cells[1]. All of the genes of this simple eukaryotic cell have recently been revealed by an international collaborative effort to determine the complete DNA sequence of its nuclear genome. Here we describe some of the features of chromosome XII.

The nucleotide composition of the chromosome, which is 38.48% G+C overall, and gene density vary across the chromosome. This has been observed for other yeast chromosomes[2–6] (Fig. 1). There are three main regions deficient in G+C, centred at approximately 150, 685 and 1,043 kilobases; one of these, as expected, coincides with the centromere. There is only one main peak of high G+C content, at approximately 473 kb, centred over the rDNA repeats. There does not seem to be any regularity in the variation in nucleotide composition, as may be the case for some other yeast chromosomes[3].

Like other yeast chromosomes, 72% of chromosome XII is predicted to code for protein (considering only two copies of the rDNA cluster). The sequence contains 534 open reading frames (ORFs) of 100 or more sense codons (excluding the 13 ORFs contained within yeast transposable elements), distributed roughly equally on the two strands (255 on the Watson (top) strand and 279 on the Crick (bottom) strand. The average ORF size is 485 codons. The largest gene in the chromosome, *YLR106c*, containing 4,910 codons, is the largest in the yeast genome. The average distance between ORFs is 545 base pairs for the 121 divergently transcribed genes (promoters abutting), 282 bp for the 120 convergently transcribed genes (terminators abutting), and 493 bp for the 208 genes that are transcribed in the same direction (promoter abutting terminator). Of the ORFs, 17 (3.2% of the total) contain introns, all of which are at the extreme 5′ end of the gene (except for *YLR464w*, a probable pseudogene). Two genes (*YLL057c* and *YLR388w*) may contain introns in the 5′-untranslated region of their mRNA. As expected[7], about half of the intron-containing genes (9) encode ribosomal proteins.

Only 170 (31.8%) of the genes were previously identified. Of the 364 newly identified genes, 34 (6.4% of the total genes) are obviously similar to proteins of known function, and 54 (10.1%) are weakly similar to proteins of known function. Thus a function is known or can be predicted for 48.3% of the encoded proteins. A further 69 genes (12.9%) encode proteins similar to proteins of unknown function; 207 (38.8%) of the predicted proteins are not similar to other proteins.

Included in the predicted ORFs are 55 that are ‘questionable’, that is, they consist of fewer than 150 codons and have a codon adaptation index (CAI)[8] of less than 0.110, or they overlap with another ORF. Of the 40 questionable ORFs that overlap with another ORF, the true gene can be predicted for 27 of these pairs, which include either a gene whose product is known (16 pairs) or whose predicted product is similar to another protein in the databases (11 pairs). There are therefore 13 overlapping ORFs that are suspect, although which of these ORFs is actually a gene awaits experimental determination.

Chromosome XII contains 22 tRNA genes, of which 7 are predicted to contain introns. Most of the tRNA genes are widely separated, although there are two clusters of three tRNA genes, each in a region of 9 kb to 13 kb (725,746–734,874 and 784,352–797,247). As expected[9], many (12) of the tRNA genes are near yeast retrotransposons (Ty elements) or their isolated long terminal repeats (LTRs). Three known small nuclear RNAs, *SNR6, SNR30* and *SNR34* are encoded on chromosome XII. Four of the six retrotransposons on chromosome XII are of the Ty1 type and two are Ty2 elements. There are several complete or partial ‘solo’ retrotransposon LTRs, including nine delta elements, four sigma elements, and a tau element.

The subtelomeric regions of chromosome XII are typical[10]. The left subtelomeric region contains a ‘core X’ element, and subtelomeric repeats STR-D, C, B and A, along with two tandem Y′ elements (short versions). The right subtelomeric region contains a core X element, the STR elements listed above, and 3–4 tandem Y′ (long version) elements. The sequence of the first 1.5 and the last Y′ elements were fused to give two copies of the Y′ element in the presented sequence. Proximal to the core X are shared homologies with several other telomere regions. As with several other chromosomes, both chromosome ends contain members of the *PAU/TIP/SRP* gene family[11].

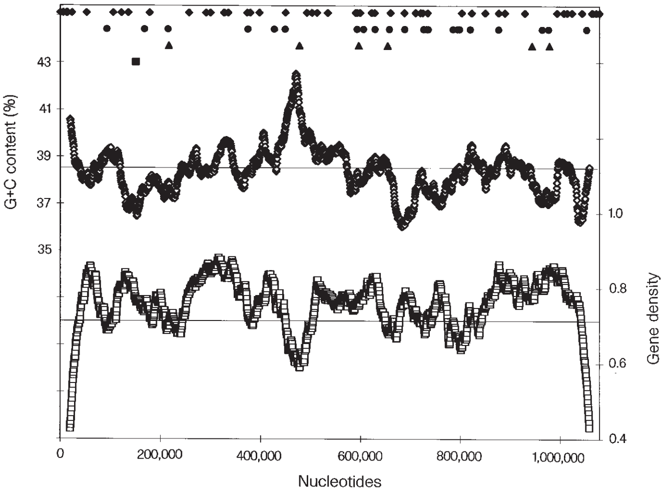

Figure 1 Top, non-coding elements of chromosome XII including autonomously replicating sequence (ARS) (filled diamonds); tRNA (filled circles); Ty element (filled triangles); centromere (filled square). Middle, G+C content as a percentage (open diamonds). Bottom, gene density (open squares).

Chromosome XII is estimated to contain 100–200 copies of the 9,137 base pair rDNA repeat[9,12–14]. Only one complete copy (the leftmost repeat in Fig. 2) and one nearly complete copy of the rDNA (the rightmost repeat in Fig. 2) are represented in the assembled sequence. In some strains these repeats seem to be interrupted by non-rDNA sequence[14].

The boundaries of the rDNA repeats are in non-transcribed regions downstream of the 35S rDNA (the left, or centromere-proximal, boundary) and 5S rDNA (the right, or centromere-distal, boundary)[15]. The structure of the left boundary of the rDNA (nucleotide 21,811 in Fig. 2) is straightforward; the right boundary[16] is more complicated. Immediately to the right of the rDNA repeats are several copies of a 3.6-kb repeat (one of which is interrupted by a Ty element) that includes the *ASP3* gene[17] and ends with a nearly complete 5S rDNA gene (5S$^{\text{var}}$ in Fig. 2). The precise number of copies of this 3.6-kb repeat in the genome is not known. The rightmost rDNA repeat ends in a 5S rDNA that adjoins a 3.6-kb repeat. Thus this rightmost rDNA repeat is lacking the 759 bp of sequence between the end of 5S rDNA and the end of the rDNA repeat (equivalent to nucleotides 30,186–30,947 in Fig. 2). The structure of the right rDNA junction differs in other yeast strains[15].

The 5S rDNA gene in the 3.6-kb repeats lacks the non-transcribed regions of the gene. It begins two nucleotides upstream of the 5′ end of 5S rRNA, and is missing the last four nucleotides of the 5S rRNA. These genes are labelled ‘5S$^{\text{var}}$’ in Fig. 2 to indicate that they are incomplete. Immediately downstream of this gene is a run of 10 T residues that is reminiscent of the transcription termination sequence of RNA polymerase III (there are 29 T residues downstream of the 5S rDNA gene in the rDNA repeats). Because this gene seems to be missing the promoter and much of the terminator, it might represent a reverse-transcribed copy of the 5S rRNA that integrated into the genome. Nevertheless these genes produce 5S rRNA transcripts[18].

One possible explanation of the structure of the right junction is that a reverse-transcribed copy of 5S rRNA is inserted into the genome near the right border of the rDNA cluster. This gene could then have been part of a 3.6-kb duplication. It is then easy to imagine a recombination event between an intact 5S rDNA gene in one of the rDNA repeats and a 5S$^{\text{var}}$ rDNA in one of the 3.6-kb repeats that generated the right rDNA junction we sequenced. Other explanations for the origin of this junction have been proffered[16].

To speed the completion of the sequence of this large chromosome, two groups collaborated on its sequencing. The rDNA repeats on chromosome XII served as a convenient point to divide the effort: the EU sequencing network[19] determined the sequence of the chromosome to the

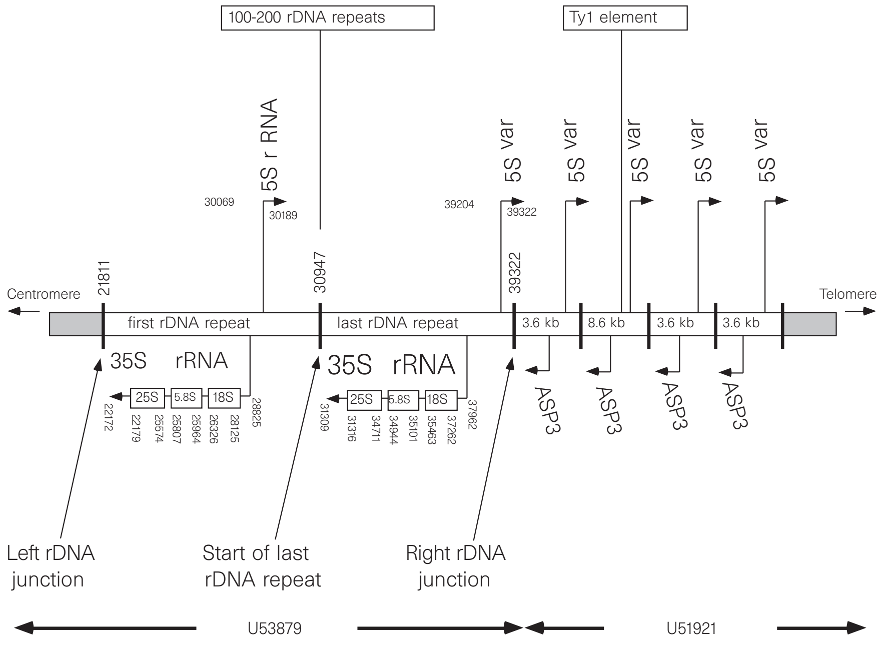

Figure 2 Diagram of the rDNA repeats and surrounding sequence, as assembled for cosmids YSCL9634 (left, GenBank accession no. U53879) and YSCL9362 (right, GenBank accession no. U51921). The numbers shown are the nucleotide coordinates for cosmid YSCL9634. The left rDNA junction (U53879 coordinate 21,811) begins at nucleotide 451,418 of chromosome XII; the right rDNA junction (U53879 coordinate 39,322) is at nucleotide 468,929. The sequence includes 1.92 rDNA repeats, representing the leftmost and rightmost copies in the genome. The remaining 100–200 rDNA repeats in the genome are represented as an insertion at coordinate 30,947. Only one complete 5S rDNA gene (in the left rDNA repeat) is included in this sequence (nucleotides 30,069–30,189); the 5S rDNA genes in the 3.6-kb repeats are variant genes. The 5S rDNA gene in the last rDNA repeat (nucleotides 39,204–39322) includes all 5′ non-translated sequences (like the normal 5S rDNA in the first repeat), but is missing sequences downstream of the 5S rRNA transcript (like the 5S rDNA genes in the 3.6-kb repeats).

left of the rDNA repeats; the sequence to the right was determined at the Washington University Genome Sequencing Center. The sequence of both strands of the entire 1,078,171 base-pair chromosome (but including only two copies of the rDNA repeats) was determined, nearly all the way to the telomeres.

The 781,865 nucleotides determined at Washington University came from 24 partly overlapping cosmid and two lambda clones[20]. The sequence of each clone was determined by a ‘shotgun’ strategy followed by directed sequencing[2]. The sequence of each clone was submitted to GenBank, and the entire non-overlapping sequence was assembled, analysed and annotated[2–5].

The 460,166 nucleotides to the left of the rDNA were determined by the EU network from a set of cosmid clones constructed from gel-purified chromosome XII DNA and mapped specifically for this purpose[21]. Sequencing was done by a directed approach that combines the advantages of primer walking (low redundancy) and ‘shotgun’ sequencing (use of a single primer)[22]. The sequence was determined from ‘shotgun’ sublibraries of 1-kb fragments of the cosmids that were then ordered by hybridization fingerprinting[22]. The sublibraries were arrayed on high-density filters, and sorted into smaller groups by hybridization with restriction fragments of the cosmids . Detailed mapping information was obtained by hybridizations with both oligodeoxynucleotides and pools of clone inserts amplified by the polymerase chain reaction (PCR).

The sequence of the left telomere region, including the TG_{1-3} sequence at the very end of the chromosome, was obtained from clones generated by integrating then excising a plasmid at the telomere, with capture of the flanking sequence[10]. The right telomere sequence was obtained by cycle sequencing of an anchored PCR product of the last Y′ element from a strain whose chromosome end was specifically marked by unique vector sequence[23]. The sequence of the very end of the Y′ element (about 130 bp short of the end of the chromosome) was not determined.

Only the sequence of the leftmost rDNA repeat (see Fig. 2) and about 300 nucleotides across the junction of the first and second repeat was determined. It was assembled appropriately to give the two rDNA repeats presented in Fig. 2 and in the database (GenBank accession no. U53879). The right junction sequence was not present in the cosmid closest to the rDNA on the right (YSCL9362; GenBank accession no. U51921), nor in two phage lambda clones that were mapped to this region. The structure of the junction was inferred from our ability to obtain a product of the expected size (the size of a 3.6-kb repeat) in PCR using an oligonucleotide primer in the 3.6-kb repeat (lying just to the right of $5S^{var}$) and a primer unique to the rDNA repeat (lying just to the left of 5S rDNA). Our sequence was assembled from these results, and found to match the sequence of the previously determined junction[16].

The complete, assembled, non-overlapping sequence of chromosome XII can be obtained at: http://speedy.mips.biochem.mpg.de/mips/yeast/ and http://genome-www.stanford.edu/Saccharomyces/.

Verification of 71,072 bp of sequence determined by the EU network (64,001 bp of overlaps between cosmids sequenced independently, and 7,071 bp of selected region that were resequenced) revealed five mistakes per 10 kb, but most errors were clustered in just a few regions. Only 14 differences were found in 175,891 nucleotides that were sequenced independently by both groups; six of these were sequencing errors, leading to an error frequency of only one mistake per 29 kb. The origins of the remaining eight discrepancies were determined by sequencing PCR products of the genome of the two strains used to generate the clones. Seven of the differences are due to changes that arose in the clones, presumably during propagation in *Escherichia coli*; only one results from differences between the two yeast strains (which are isogenic, but were propagated separately for many years) used to generate the two sets of clones. Thus the number of errors in the sequence is equivalent to the number of errors resulting from propagation of the DNA in *E. coli* and yeast. □

Received 31 July 1996; accepted 11 March 1997.

1. Jones, E. W., Pringle. J. R. & Broach J. R. *The Molecular and Cellular Biology of the Yeast* Saccharomyces, Vols 1–3, (Cold Spring Harbor Laboratory Press, NY, 1991–1996).
2. Johnston, M. *et al. Science* 265, 2077–2082 (1994).

3. Dujon, B. *et al. Nature* 369, 371–378 (1994).
4. Feldmann, H. *et al. EMBO J.* 13, 5795-5809 (1994).
5. Galibert, A. *et al. EMBO J.* 15, 2031–2049 (1996).
6. Sharp, P. & Lloyd, A. *Nucleic Acids Res.* 21, 179–183 (1993).
7. Rodriguez-Medina, J. R. & Rymond, B. C. *Mol. Gen. Genet*. 243, 532–539 (1994).
8. Sharp, P. M. & Li, W. H. *Nucleic Acids Res.* 15, 1281–1295 (1987).
9. Olson, M. V. in *The Molecular and Cellular Biology of the Yeast* Saccharomyces, Vol. 1 (ed. Broach J. R., Pringle. J. R. & Jones, E. W.) 1–39 (Cold Spring Harbor Laboratory Press, NY, 1991).
10. Louis, E. J. & Borts, R. *Genetics* 139, 125–136 (1995).
11. Viswanathan, M., Muthukumar, G., Lenard, J. & Cong, Y. S. *Gene* 148, 149–153 (1994).
12. Pasero, P. & Marilley, M. *Mol. Gen. Genet.* 236, 448–452 (1993).
13. Chindamporn, A., Iwaguchi, S., Nakagawa, Y., Homma, M. & Tanaka, K. *J. Gen. Microbiol.* 139, 1409–1415 (1993).
14. Rustchenko, E. P. & Sherman, F. *Yeast* 10, 1157–1171 (1994).
15. Zamb, T. & Petes, T. D. *Cell* 28, 355–364 (1982).
16. McMahon, M. E., Stamenkovich, D. & Petes, T. D. *Nucleic Acids Res*. 12, 8001–8016 (1984).
17. Kim, K. W., Kamerud, J. Q., Livingston, D. M. & Roon, R. J. *J. Biol. Chem.* 263, 11948–11953 (1988).
18. Piper, P. W., Lockheart, A. & Patel, N. *Nucleic Acids Res.* 12, 4083–4096 (1984).
19. Vassarotti, A. *et al. J. Biotechnol.* 41, 131–137 (1995).
20. Riles, L. *et al. Genetics* 134, 81–150 (1993).
21. Scholler, P., Schwarz, S. & Hoheisel, J. D. *Yeast* 11, 659–666 (1995).
22. Scholler, P. *et al. Nucleic Acids Res.* 23, 3842–3849 (1995).
23. Louis, E. J. *Biochemica* 3, 25–26 (1995).

Acknowledgements. We thank J. Warner for help with interpreting the rDNA sequence. This work was supported by the NIH National Center for Human Genome Research and the European Commision; the Belgian Federal Services for Science Policy (D.W.T.C.) and the Service Fédéraux des Affaires Scientifiques, Techniques et Culturelles, Pôles d'attraction Inter-universitaire, the Region de Bruxelles-Capitale, Region Wallonne, the Research Fund of the Katholieke Universiteit Leuven and The Wellcome Trust.

Correspondence and requests for materials should be addressed to M.J. (e-mail: mj@genetics.wustl.edu/)

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII

S. Bowman, C. Churcher, K. Badcock, D. Brown, T. Chillingworth, R. Connor, K Dedman, K. Devlin, S. Gentles, N. Hamlin, S. Hunt, K. Jagels, G. Lye, S. Moule, C. Odell, D. Pearson, M. Rajandream, P. Rice, J. Skelton, S. Walsh, S. Whitehead & B. Barrell

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Systematic sequencing of the genome of *Saccharomyces cerevisiae* has revealed thousands of new predicted genes and allowed analysis of long-range features of chromosomal organization. Generally, genes and predicted genes seem to be distributed evenly throughout the genome, having no overall preference for DNA strand. Apart from the smaller chromosomes, which can have substantially lower gene density in their telomeric regions[1–3], there is a consistent average of one open reading frame (ORF) approximately every two kilobases. However, one of the most surprising findings for a eukaryote with approximately 6,000 genes was the amount of apparent redundancy in its genome. This redundancy occurs both between individual ORFs and over more extensive chromosome regions, which have been duplicated preserving gene order and orientation[4–6]. Here we report the entire nucleotide sequence of chromosome XIII, the sixth-largest *S. cerevisiae* chromosome, and demonstrate that its features and organization are consistent with those observed for other *S. cerevisiae* chromosomes. Analysis revealed 459 ORFs, 284 have not been identified previously. Both intra- and interchromosomal duplications of regions of this chromosome have occurred.

Chromosome XIII of *S. cerevisiae* is 924,430 base pairs long, and contains 459 ORFs. Eight of these are TyA and TyB ORFs from four Ty1 retrotransposons present on chromosome XIII in strain AB972, two of which are located on each arm of the chromosome, and these are excluded from further analyses. The average gene density on this chromosome is one ORF for every 1,997 base pairs of DNA, which correlates well with that observed for other *S. cerevisiae* chromosomes, with 74.2% of DNA on this chromosome contributing to ORFs. An average chromosome XIII ORF is 494 codons long.

Of the 451 *S. cerevisiae* ORFs on chromosome XIII, 167 (37.0%) encode previously identified proteins. A further 281 (62.3%) predicted genes have not been previously sequenced; 121 (26.8%) of these ORFs have similarities to genes for which some biochemical information is available. However, several of this category of ORF have their best protein similarity to a protein of unknown function. A total of 160 ORFs (35.5%) encode predicted proteins that are not significantly similar to proteins of known function. Because of the rapidly advancing progress of other systematic sequencing projects, many of these ORFs have homology to hypothetical proteins both in yeasts and higher organisms. A total of 51 predicted genes have similarity only to predicted proteins of unknown function. Although the majority are most similar to another *S. cerevisiae* hypothetical protein (50.9%), several have their best homology to an ORF identified in systematic sequencing of the yeast *Schizosaccharomyces pombe*[7] (11.8%), or to predicted proteins in the nematode *Caenorhabditis elegans*[8] (17%). Thus they are members of gene families whose function is currently unknown. There were no significant protein sequence similarities for 109 ORFs, of which 11 are thought to be questionable ORFs based on their length, codon adaptation index (CAI) value and positional base preferences.

During the systematic sequencing of other chromosomes, several putative pseudogenes were identified[1,9]. These consisted of ORFs separated by a stop codon or frameshift from upstream or downstream sequences that shared a common homology to a single *S. cerevisiae* ORF. Most of these pseudogenes identified occur close to the telomeres of chromosomes. Three ORFs on chromosome XIII (YMR084W, YMR085W and YMR326C) have been classified as putative pseudogene ORFs. Of these, only YMR326C is located close to one of the chromosome telomeres; all three have strong similarity to sequences found elsewhere in the *S. cerevisiae* genome. These frameshifts have been confirmed by sequencing genomic DNA.

The average intergenic distance between adjacent ORFs depends on their relative orientation. This is certainly the case on chromosome XIII, in which 204 ORFs are arranged in tandem with an average intergenic distance of 450 base pairs. Of these, 110 are divergent and are an average 616 bp apart, and 111 are convergent and an average of 260 bp apart. This is consistent with a greater sequence requirement for the regulation of gene expression from promoter elements than for transcription termination.

Of the 451 ORFs on chromosome XIII, 24 (5.3%) are predicted to contain introns. There seems to be no preference for DNA strand, with 229 genes coded on the Watson strand and 222 on the Crick strand. There is no evidence of any significant clustering of related genes. However, there are several instances in which two very similar ORFs occur close to one another in tandem; for example, YMR169C and YMR170C/ALD2 (aldehyde dehydrogenases), and YMR006C and YMR008C/PLB1 (lysophospholipases).

The longest ORF on chromosome XIII is *HFA1*, (which is homologous to *FAS3*), a putative acetyl-CoA carboxylase that had been sequenced previously[10] (2,123 codons). A total of 39 ORFs on this chromosome are more than 1,000 codons in length. *S. cerevisiae* genes of less than 100 codons with no homology are difficult to detect[11]. On chromosome XIII, 10 ORFs shorter than 100 amino acids in length have been identified. The smallest of these is YMR248C, which is just 55 amino acids long, and may be spliced to a second small ORF immediately upstream. The smallest ORF on this chromosome that encodes a previously characterized protein is *COX7*, which is 59 amino acids long and encodes cytochrome oxidase polypeptide VII (ref.12).

Chromosome XIII encodes 21 predicted tRNA genes, of which six are spliced. In addition to the four Ty1 retrotransposons, several long terminal repeat (LTR) sequences are present, providing evidence of previous trans-