4. Riles, L. *et al. Genetics.* 134, 81–150 (1993).
5. Louis, E. J. & Haber, J. E. *Genetics* 131, 559–574 (1992).
6. Louis, E. J., Naumova, E. S., Lee, A., Naumov, G. & Haber, J. E. *Genetics* 136, 789–802 (1994).
7. Lashkari, D. A., Hunicke–Smith, S. P,. Norgren, R. M., Davis, R.W. & Brennan, T. *Proc. Natl Acad. Sci.* 92, 7912–7915 (1995).
8. Stiles, J. I., Friedman, L. R., Helms, C., Consaul, S. & Sherman, F. *J. Mol. Biol.* 148, 331–346 (1981).
9. Smith, T., Waterman, M. & Burks, C. *Nucleic Acids Res.* 13, 645–656 (1985).
10. Weber, E., Rodriguez, C., Chevallier, M. R. & Jund, R. *Mol. Microbiol.* 4, 585–596 (1990).
11. Fleischmann, R. D. *et al. Science* 269, 496–512 (1995).
12. Fraser, C. M. *et al. Science* 270, 397–403 (1995).
13. Oefner, P. S. *et al. Nucleic Acids Res.* 24, 3879–3886 (1996).
14. Olson, M. V. *et al. Proc. Natl. Acad. Sci. USA* 83, 7826–7830 (1986).
15. Fleer, R., Nicolet, C. M., Pure, G. A. & Friedberg, E. C. *Mol. Cell. Biol.* 7, 1180–1192 (1987).
16. Gleeson, T.J. & Staden, R. *Comp. Appl. Biosci.* 7, 398 (1991).

Acknowledgements. We thank the members of the informal, international group who sequenced the entire yeast genome for their generosity and cooperation that enabled the Yeast Genome Project to be completed a year earlier than scheduled and under budget. In particular, we thank L. Riles and M. Olson for sending us their set of recombinant yeast DNAs and unpublished mapping data; E. Gilbertson for assistance in preparing this manuscript; and N. Schroff and A. Wynant for technical assistance. This work was supported by grants from the NIH (U. S. Public Health Service).

Correspondence and requests for materials should be addressed to R. W. D. (e-mail: gilbert@genome.stanford.edu). The sequence of yeast chromosome V has been deposited in Genbank.

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII

H. Tettelin[1], M. L. Agostoni Carbone[2], K. Albermann[3], M. Albers[4], J. Arroyo[5], U. Backes[4], T. Barreiros[6], I. Bertani[7], A. J. Bjourson[8], M. Brückner[9], C.V. Bruschi[7], G. Carignani[10], L. Castagnoli[11], E. Cerdan[12], M. L. Clemente[10], A. Coblenz[4], M. Coglievina[7], E. Coissac[13], E. Defoor14, S. Del Bino[1], H. Delius[15], D. Delneri[7], P. de Wergifosse[1], B. Dujon[16], P. Durand[17], K. D. Entian[18], P. Eraso[19], V. Escribano[19], L. Fabiani[20], B. Fartmann[21]*, F. Feroli[10], M. Feuermann[22], L. Frontali[20], M. García-Gonzalez[5]*, M. I. García-Sáez[5], A. Goffeau[1], P. Guerreiro[6], J. Hani[3], M. Hansen[4], U. Hebling[15], K. Hernandez[23], K. Heumann[3], F. Hilger[17], B. Hofmann[15], K. J. Indge[24], C.M. James[24], R. Klima[7], P. Kötter[18], B. Kramer[21]*, W. Kramer[21], G. Lauquin[25], H. Leuther[4], E. J. Louis[26], E. Maillier[13], A. Marconi[20], E. Martegani[27], M. J. Mazón[19], C. Mazzoni[20], A. D. K. McReynolds[8], P. Melchioretto[2], H. W. Mewes[3], O. Minenkova[11], S. Müller-Auer[9], A. Nawrocki[28], P. Netter[13], R. Neu[4], C. Nombela[5], S.G. Oliver[24], L. Panzeri[2], S. Paoluzi[11], P. Plevani[2], D. Portetelle[17], F. Portillo[19], S. Potier[22], B. Purnelle[1], M. Rieger[9], L. Riles[29], T. Rinaldi[20], J. Robben[14], C. Rodrigues-Pousada[6], E. Rodriguez-Belmonte[12], A. M. Rodriguez-Torres[12], M. Rose[18], M. Ruzzi[30], M. Saliola[20], M. Sánchez-Perez[5], B. Schäfer[4], M. Schäfer[9], M. Scharfe[31], T. Schmidheini[23], A. Schreer[4], J. Skala[28], J. L. Souciet[22], H.Y. Steensma[32,33], E. Talla[1], A. Thierry[16], M. Vandenbol[17], Q. J. M. van der Aart[32], L. Van Dyck[1], M. Vanoni[27], P. Verhasselt[14], M. Voet[14], G. Volckaert[14], R. Wambutt[31], M. D. Watson[34], N. Weber[23], E. Wedler[31], H. Wedler[31], P. Wipfli[23], K. Wolf[4], L. F. Wright[8], P. Zaccaria[7], M. Zimmermann[4], A. Zollner[3] & K. Kleine[3]

[1]Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix-du-Sud 2/20, B-1348 Louvain-la-Neuve, Belgium
[2]Dipartimento di Genetica e di Biologia dei Microrganismi, Università di Milano, via Celoria 26, I-20133 Milano, Italy
[3]Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany
[4]Institut für Biologie IV, Mikrobiologie, Worringerweg, RWTH-Aachen, D-52056, Germany
[5]Departamento de Microbiología II and Centro de Secuenciación de DNA de la UCM, Faculdad de Farmacia, Universidad Complutense, E-28040 Madrid, Spain
[6]Laboratorio de Genetica Molecular, Instituto Gulbenkian de Ciencia, Ap 14, E-2781 Oeiras Codex, Portugal
[7]Microbiology Group, International Center for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy;
[8]The Biotechnology Center for Animal and Plant Health, The Queens University of Belfast, Newforge Lane, Belfast, BT9 5PX, UK
[9]Genotype GmbH, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany
[10]Dipartimento di Chimica biologica, Università di Padova, via Trieste 75, I-35121 Padova, Italy
[11]Dipartimento di Biologia, Università di Roma 'Tor Vergata', via della Ricerca Scientifica, I-00133 Roma, Italy
[12]Departamento de Biologia Celular y Molecular, Facultad de Ciencias, Universidad de La Coruna, Campus de La Zapateira s/n, La Coruna, Spain;
[13]Centre de Génétique Moléculaire, CNRS, Laboratoire Associé à l'Université Pierre et Marie Curie, F-91198 Gif-sur-Yvette Cedex, France
[14]Katholieke Universiteit Leuven, Laboratory of Gene Technology, Willem de Croylaan 42, B-3001 Leuven, Belgium
[15]Deutsches Krebsforschungszentrum, Department for Applied Virology, D-69120 Heidelberg, Germany
[16]Unité de Génétique Moléculaire des levures (URA 1149 of CNRS and UPR 927 of University P.M. Curie, Paris), Department of Biotechnologies, 25 rue du Dr. Roux, Institut Pasteur, F-75724 Paris Cedex 15, France
[17]Unité de Microbiologie, Faculté Universitaire des Sciences Agronomiques de Gembloux, Avenue Maréchal Juin 6, B-5030 Gembloux, Belgium
[18]Institut für Mikrobiologie der Johann Wolfgang Goethe-Universität Frankfurt /Main, Marie-Curie Strasse 9, D-60439 Frankfurt, Germany;
[19]Instituto de Investigaciones Biomédicas del C.S.I.C.y Departamento de Bioquímica, Facultad de Medicina de la U.A.M., E-28029 Madrid, Spain
[20]Dipartimento di Biologia Cellulare e dello Sviluppo, Università di Roma La Sapienza, P. le Aldo Moro 5, I-00185 Roma, Italy
[21]Institut für Molekulare Genetik, Georg-August-Universität, Grisebachstr. 8, D-37077 Göttingen, Germany
[22]Laboratoire de Microbiologie et Génétique URA 1481, Université Louis Pasteur/CNRS, Institut de Botanique, rue Goethe 28, F-67083 Strasbourg Cedex, France
[23]Microsynth GmbH, Schützenstrasse 15, CH-9436 Balgach, Switzerland
[24]Department of Biochemistry and Applied Molecular Biology, UMIST, PO Box 88, Sackville Street, Manchester M60 1QD, UK
[25]Institut de Biochimie Cellulaire, CNRS, rue Camille Saint-Saèns 1, F-33077 Bordeaux Cedex, France
[26]Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, OX3 9DU, UK
[27]Dipartimento di Biochimica e Fisiologia Generali, Università di Milano, via Celoria 26, I-20133 Milano, Italy
[28]Institute of Microbiology, Wroclaw University, Przybyszewskiego 63, P-51148 Wroclaw, Poland
[29]Genome Sequencing Center and Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63110, USA
[30]Dipartimento di Agrobiologia e Agrochimica, Università della Tuscia, via S. Camillo de Lellis, I-01100 Viterbo, Italy
[31]AGON GmbH, Glienicker Weg 185, D-12489 Berlin, Germany
[32]Institute for Molecular Plant Sciences, Leiden University, Wassenaarseweg 64, NL-2333 AL Leiden, The Netherlands
[33]Department of Microbiology and Enzymology, Delft University of Technology, Julianalaan 67, NL-2628 BC Delft, The Netherlands
[34]Department of Biological Sciences, University of Durham, South Road, Durham, DH1 3LE, UK

*Present addresses: MWG-BIOTECH GmbH, Anzinger Strasse 7, D-85560 Ebersberg, Germany (B. F.); Centro de Biotecnología, Camagüey, Cuba (M. G.-G.); Abteilung Klinische Biochemie, Zentrum Innere Medizin, Georg-August-Universität, Robert-Koch-Strasse 40, D-37075 Göttingen, Germany, (B. K.).

The complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII has 572 predicted open reading frames (ORFs), of which 341 are new. No correlation was found between G+C content and gene density along the chromosome, and their variations are random. Of the ORFs, 17% show high similarity to human proteins. Almost half of the ORFs could be classified in functional categories, and there is a slight increase in the number of transcription (7.0 %) and translation (5.2 %) factors when com-

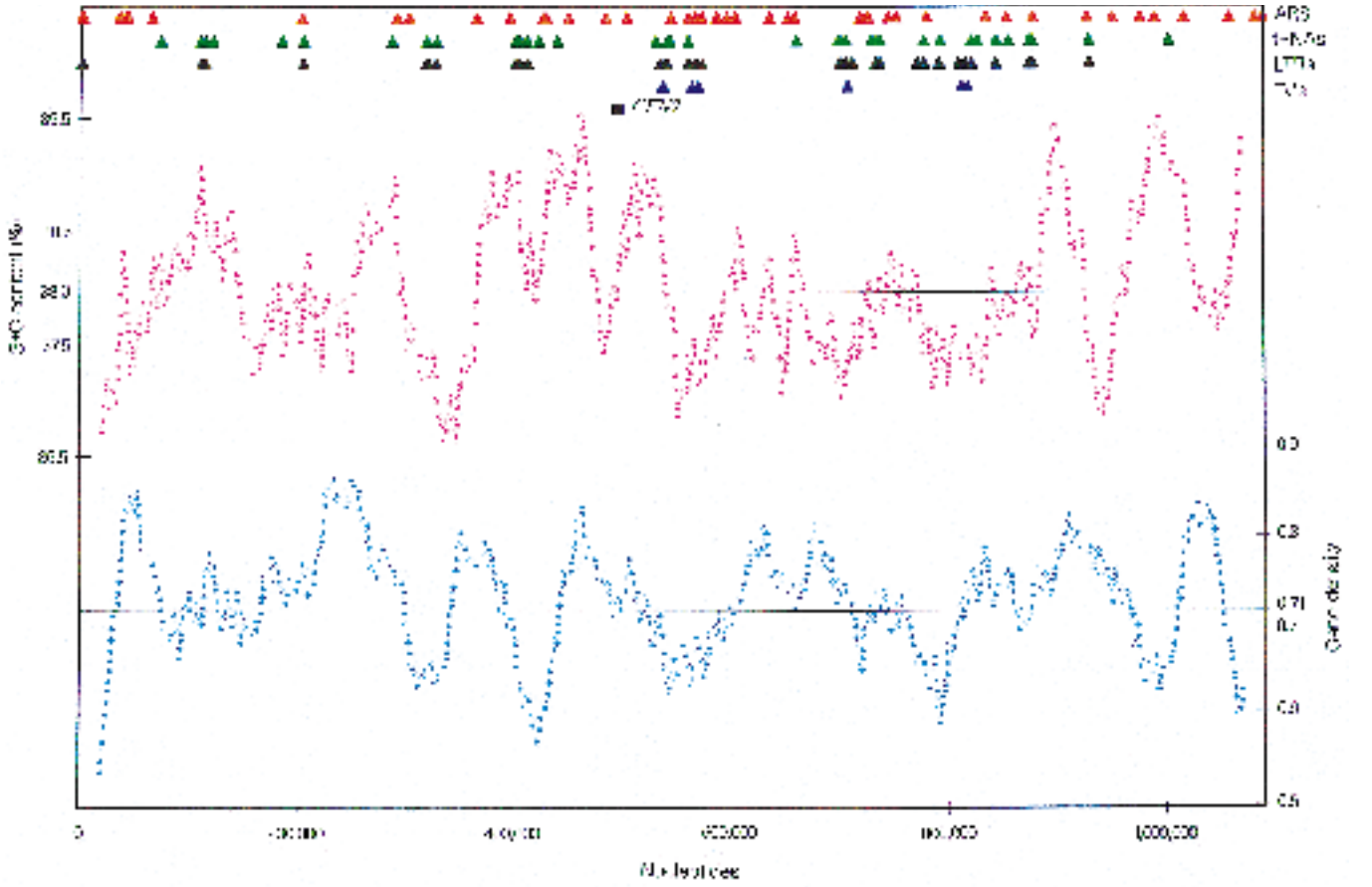

Figure 1 Top, position of genetic elements along chromosome VII. Middle, compositional variation curve; each point represents the average G+C content in a 40-kb sliding window (steps are 500 bp); the horizontal line represents the average G+C (38%). Bottom, gene density is expressed as the fraction of nucleotides within ORFs versus the total number of nucleotides in sliding windows of 40 kb (steps are 500 bp); the horizontal line represents the average gene density (0.71).

pared with the complete *S. cerevisiae* genome. Accurate verification procedures demonstrate that there are less than two errors per 10,000 base pairs in the published sequence.

Before the publication in 1992 of the yeast chromosome III sequence[1], the only available *S. cerevisiae* genome sequence of appreciable size was a contig of 24 kilobases (kb) from chromosome VII (refs 2, 3). A 60-kb physical map covering the left arm of this chromosome between *CEN7* and *TRP5* markers had been built, allowing sequencing and transcriptional mapping of nine ORFs located in the 24-kb region spanning the *PMA1* and *ATE1* loci. Analysis of these data led to the estimation of a minimum number of 5,300 expressed genes in yeast. In this centromeric region a recombination frequency of 1 cM corresponded to an average distance of 3.3 kb, compared with 2.9 kb for the complete chromosome. These extrapolations have been confirmed by the complete 1,090,936 nucleotide sequence of chromosome VII, the fourth longest in *S. cerevisiae*.

Chromosome VII contains 564 ORFs of more than 99 codons, plus eight smaller, previously identified, ORFs. Of these 572 ORFs, 19 are predicted to carry an intron at their extreme 5′ end. The *RPL6A* gene is interrupted by two introns, one of which codes for the small nuclear RNA39. Of the 572 ORFs, 152 (26.5 %) had previously been characterized biochemically, and an additional 79 (14 %) had been characterized phenotypically. Of these 231 known genes, disruption phenotypes have been reported in 140 cases, of which 37 are lethal. An additional 61 ORFs (11 %) show a high similarity (FASTA score above 300 or one third of self score) to another ORF of known function. However, if the threshold FASTA score is lowered to 150, which is a significant value in many cases, another 20 ORFs (3.5 %) could be envisaged to be of predictable function, raising the total number of ORFs of known or predictable function to 312 (54.6 %). Of the remaining ORFs, 74 (12.9 %) are similar to protein sequences of unknown function, and 186 ORFs (32.5 %) show weak or no significant similarity to any other protein sequence in the public data libraries. Finally, the expression of 63 ORFs (11 %) is questionable owing to their partial overlap with other ORFs (44) or to a combination of small size (less than 150 codons) and low codon adaptation index (CAI below 0.110)[4]. However, these criteria are not absolute as at least four expressed ORFs from chromosome VII do not meet them: *AGA2* (87 codons; CAI, 0.089), *SOH1* (127 codons; CAI, 0.096), *SPT4* (102 codons; CAI, 0.109) and *VMA21* (77 codons; CAI, 0.109). Almost 30 % of the ORFs from chromosome VII are redundant, as 166 show high similarity (FASTA score greater than 300 or one third of the self score) with other yeast genes.

The average ORF size is 468 codons, the longest (YGL195w, *GCN1*) being 2,672 codons. The average distance between ORFs located on the same strand is 514 base pairs when not containing tRNAs or long terminal repeats (LTRs) and six are longer than 2,000 bp. In the case of divergent promotors, the spacing is 553 bp, with two being longer than 2,000 bp. The mean size allocated to convergent terminators is only 304 bp, and one of these is longer than 2,000 bp. The mean G+C content of inter-ORFs regions is 33 %. All of these values are very similar to those found for the complete genome[5].

An attempt has been made to classify chromosome VII ORFs in functional categories (Table 1). Note that any given protein may belong to more than one category. The percentage of ORFs from chromosome VII in each functional category correlates with the functional distribution of ORFs within the complete genome, with a slightly higher content in the transcription (13 transcription factors) and protein synthesis (18 ribosomal proteins) categories.

The putative transmembrane spans have been computed with the KKD algorithm[6] using a rather low threshold[7] that takes into account not only the fully hydrophobic spans, but also the predicted amphipathic α-

Table 1 Functional categories of yeast ORFs from chromosome VII

Functional category	ORFs	
	Number	Percentage
Metabolism	55	9.0
Energy	19	3.1
DNA synthesis	10	1.6
Transcription	43	7.0
Protein synthesis	32	5.2
Protein destination	18	2.9
Transport facilitation (permeases)	18	2.9
Intracellular traffic	17	2.8
Cell structure	19	3.1
Organelle assembly	11	1.8
Signal transduction	5	0.8
Cell division	24	3.9
Cell rescue	12	1.9
Retrotransposons	13	2.1
Unclassified	318	51.8
Total	614	

Table 2 Similarity of yeast chromosome VII ORFs of unknown function with human proteins

Yeast ORF	Human protein
YGL150c	SNF2a transcription activator that cooperates with the oestrogen and retinoic acid receptors
YGL125w	methylenetetrahydrofolate reductase
YGL106w	calmodulin
YGL003c	probable cell-division control protein CDC 55
YGR034w	new ribosomal protein similar to the human ribosomal protein L26
YGR043c	transaldolase
YGR217w	first putative calcium channel in yeast similar to the human voltage-dependent L-type calcium channel a1 subunit
YGR231c	prohibitin, which inhibits DNA synthesis and regulates proliferation
YGR256w	phosphogluconate dehydrogenase

The similarity threshold is a FASTA score higher than 300 or higher than one third of self score.

helices which, when present in a bundle, can contribute to the formation of a polar channel within the lipid bilayer. Of the 572 ORFs, 359 (63 %) show no predicted transmembrane spans or are known to be soluble, 79 (14 %) carry at least three putative spans or are known to be membrane bound, and 134 (23 %) have one or two predicted hydrophobic α-helices, a feature which does not necessarily mean that they are membrane-bound. All ORFs have been submitted to PSORT analysis[8] to predict their subcellular localization. If we consider a high certainty score to be at least 0.8 for a given localization, and a low one to be less than 0.5 for all other possible localizations, only 76 ORFs match these criteria: 37 ORF products predicted to be in the nucleus, 16 in the endoplasmic reticulum, nine in the mitochondria, nine in the plasma membrane, two in the vacuole, two in the peroxisome, and one secreted outside the cell. For the 23 chromosome VII proteins of known subcellular localization, the PSORT prediction was correct in 14 cases (61 %). Note that three of the six ORFs that exhibited an erroneous nuclear localization were ribosomal proteins which are known to carry a nuclear localization signal to allow the biosynthesis of the ribosome particles[9].

By applying the program PYTHIA[10] to search for simple repeats within a gene, we detected at least 19 genes with regularly repeated nucleotides corresponding to repeated amino acids in the encoded proteins. Some of these regions are composed of single amino-acid repeats, including 21 aspartates (in YGL227w), 13 aspartates (YGL058w, *RAD6*), 11 and 8 glutamines (YGL066w), 18 asparagines (YGR233c), 15 asparagines (YGL014w), 10 asparagines (YGL013c, *PDR1*), 23 serines (YGR023w) or 16 serines (YGR130c). We also identified more complex repeats, such as $(T\ S/N\ ATT\ A/E\ S\ X_4)_{11}$ in YGR296w (=Y′), $(QQQP)_9$ in YGL122c (*NAB2*), $(DEEE)_3$ in YGL164w, $(AQ)_{14}$ in YGL181w (*GTS1*), $(TSSS)_9$ in YGL028c, or $(HN)_5$ in YGL178w (*MPT5*).

A systematic search for similarities with human proteins was performed for the 572 ORFs of yeast chromosome VII. A total of 95 ORFs (16.6 %) show a very significant similarity (FASTA score higher than 300 or higher than one third of self score) with human proteins, of which these 79 (13.8%) correspond to known yeast proteins whose function is often closely related to the function of the human homologue. Similarities of some of the 16 previously unknown ORFs (2.8 %) with human proteins are shown in Table 2.

Several chromosome VII ORFs show a high degree of similarity with interesting proteins from other organisms. These include: YGL236c, similar to the glucose-inhibited division (gidA) protein of the bacterium *Escherichia coli*; YGL201c, similar to the intestinal DNA replication protein of the rat *Rattus norvegicus*; and YGL054c, similar to the Cni protein necessary for anterior–posterior and dorsal–ventral patterning in the fruit fly *Drosophila melanogaster*.

Another interesting feature of *S. cerevisiae* chromosome VII is the existence of a pseudogene, which has been confirmed by direct polymerase chain reaction (PCR) sequencing on the yeast genome. This pseudogene, YGL259w, contains two frameshifts, and only one of the three ORFs is longer than 99 codons. However, all three parts show a high similarity with YIR039c, a hypothetical aspartyl proteinase. Another curious feature is the presence of three ORFs in the same frame separated by two stop codons. The rightmost one, YGL238w, corresponds to the *CSE1* chromosome segregation gene. The leftmost ORF, YGL241w, shows 17% identity over 1,053 amino acids with *CSE1*, whereas the central ORF shows no similarity with this protein. Finally, a possibly unique feature in the yeast genome is the tail-to-tail arrangement of the *SMD1* (YGR074w) and *PRP38* (YGR075c) genes, with their respective ORFs terminating on opposite strands without any intervening nucleotide between the stop codons[11]. This region, as well as the *CSE1* region, have been verified by direct PCR sequencing on the yeast genome.

Chromosome VII contains six yeast retrotransposons: three Ty1s, one Ty2, one Ty3 and a pseudo-Ty, which contains, in addition to the normal frameshift separating the Ty1A and Ty1B coding sequences, two frameshifts splitting both ORFs in two parts. Of 35 tRNA genes identified, eight are interrupted by introns. All LTRs on chromosome VII are associated with a tRNA (for review, see ref. 12), except for the two Ty5 LTRs located close to the left telomere; however, 12 tRNAs are not associated with LTRs. The positions of tRNAs, Tys, LTRs and putative ARS consensus have been compared to G+C content and gene density along the chromosome (see Fig. 1). No clear correlation could be identified regarding the location of these genetic elements. Furthermore, no statistical correlation exists between the G+C content and the gene density curves (correlation coefficient, 0.04) when highly overlapping, neighbouring sliding windows (98.7 % overlap) are used. These G+C content and gene density variations are not significant, as the same analysis performed on several random mixes of the original sequence yields similar G+C content and gene density variations. A similar graph was obtained using non-overlapping neighbouring windows along the sequence after removal of Ty and LTR elements. In this case, a good correlation was found between the G+C content of the genes and the gene density (correlation coefficient, 0.98), as well as between the G+C content of silent positions of codons in genes and the gene density (correlation coefficient, 0.66). Finally, there is no significant difference in the G+C content of the coding regions on each strand: 40.07 % G+C and 305 ORFs on the Watson strand, and 39.95 % G+C and 267 ORFs on the Crick strand.

The physical map of chromosome VII has been constructed independently from the genetic map using the meganuclease I-*Sce*I to produce *in vitro* nested fragmentations of the chromosome[13,14]. The cosmids chosen for sequencing were screened from two genomic libraries[14] and completed using a few cosmids from the physical map constructed by L. Riles (unpublished). The left telomere has been cloned[15] and the right telomeric sequence was obtained from a PCR fragment amplified from a strain carrying the pEL61 plasmid[15] integrated in the subtelomeric region.

The quality of the sequence of chromosome VII was assessed using dif-

ferent approaches. As well as partial overlaps between the regions sequenced by two laboratories, putative frameshift checking, alignment of the sequence with previously published data and a few random resequencing verifications were performed on cosmid subclones. A new method for verifying specific regions of the sequence was developed for this chromosome (G. V. *et al.*, manuscript in preparation) and applied to several other chromosomes. The extrapolation from the number of discrepancies observed in the overlaps (102,049 nucleotides, 9.4 %) to the whole sequence suggests that the nucleotide sequence of chromosome VII is 99.974 % accurate. The quality of the coding regions, where frameshifts are quite easy to check, is probably much higher than that of the intergenic regions. Indeed, the quality assessment procedure led to the correction of a total of 90 errors mainly located in the coding regions. A total of 56,344 bp (5.2 %) have been resequenced, and the comparison with the original data makes it possible to estimate that about 120 errors remain in the chromosome VII sequence. Parts of the sequence were published independently[16–30] before assembly of the contig and application of the final quality controls; several other manuscripts are in the press. □

Received 19 July 1996; accepted 11 March 1997.

1. Oliver, S.G. et al. *Nature* 357, 38-46 (1992).
2. Capieaux, E., Ulaszewski, S., Balzi, E. & Goffeau, A. *Yeast* 7, 275–280 (1991).
3. Chen, W. *et al. Yeast* 7, 287–299 (1991).
4. Dujon, B. *et al. Nature* 369, 371–378 (1994).
5. Dujon, B. *Trends Genet.* 12, 263–270 (1996).
6. Klein, P., Kanehisa, M. & Delisi, C. *Biochim. Biophys. Acta* 815, 468–476 (1985).
7. Goffeau, A., Slonimski, P., Nakai, K. & Risler, J.L. *Yeast* 9, 691–702 (1993).
8. Nakai, K. & Kanehisa, M. *Genomics* 14, 897–911 (1992).
9. Woolford, J. L. & Warner, J. R. in *The Molecular and Cellular Biology of the Yeast Saccharomyces* Vol.1 (ed. Jones, E., Pringle, J. & Broach, J.) 597–598 (Cold Spring Harbor Laboratory Press, NY, 1991).
10. Milosavljevic, A. & Jurka, J. *Comput. Adv. Biosci.* 9, 409–411 (1993).
11. Rymond, B. C. *Proc. Natl Acad. Sci. USA* 90, 848–852 (1993).
12. Olson, M. V. in *The Molecular and Cellular Biology of the Yeast Saccharomyces* Vol. 1 (ed. Jones, E., Pringle, J. & Broach, J.) 1–40 (Cold Spring Harbor Laboratory Press, NY, 1991).
13. Tettelin, H. *et al.* in *Methods in Molecular Genetics*, (ed. Adolph, K.W.) (Academic,London, 1995).
14. Thierry, A., Gaillon, L., Galibert, F. & Dujon, B. *Yeast* 11, 121–135 (1995).
15. Louis, E. & Borts, R. *Genetics* 139, 125–136 (1995).
16. Arroyo, J. *et al. Yeast* 11, 587–591 (1995).
17. Bertani, I. *et al. Yeast* 11, 1187–1194 (1995).
18. Coglievina, M. *et al. Yeast* 11, 767–774 (1995).
19. Coissac, E., Maillier, E., Robineau, S. & Netter, P. *Yeast* 12, 1555–1562 (1996).
20. Escribano, V., Eraso, P., Portillo, F. & Mazón, M.J. *Yeast* 12, 887–892 (1996).
21. Guerreiro, P. *et al. Yeast* 11, 1087–1091 (1995).
22. Guerreiro, P. *et al. Yeast* 12, 273–280 (1996).
23. Hansen, M. *et al. Yeast* 12, 1273–1277 (1996).
24. James, C.M., Indge, K.J. & Oliver, S.G. *Yeast* 11, 1413–1419 (1995).
25. Klima, R. *et al. Yeast* 12, 1033–1040 (1996).
26. Rodriguez–Belmonte, E. *et al. Yeast* 12, 145–148 (1996).
27. Skala, J., Nawrocki, A. & Goffeau, A. *Yeast* 11, 1421–1427 (1995).
28. Tizon, B. *et al. Yeast* 12, 1047–1051 (1996).
29. Vandenbol, M., Durand, P., Portetelle, D. & Hilger, F. *Yeast* 11, 1519–1523 (1995).
30. Van der Aart, Q.J.M., Kleine, K. & Steensma, H.Y. *Yeast* 12, 385–390 (1996).

Acknowledgements. This study is part of the third and final phase of the European Yeast Genome Sequencing Project carried out under the administrative coordination of A. Vassarotti and the Université Catholique de Louvain. We thank P. Mordant for accounting and help; F. Foury for support and advice; G. Gérard for statistical analyses; and our colleagues for help and discussion. We would also like to acknowledge the contribution of S. Kiefer who died in a car accident in 1994. This work was supported by the EU under the BIOTECH II programme and by the Services Fédéraux des Affaires Scientifiques, Techniques et Culturelles; the Pôles d'Attraction Inter-Universitaire; the Région Wallonne; the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture; the Belgian Federal Services for Science Policy; the Research Fund of the Katholieke Universiteit Leuven; the Schweizerisches Bundesamt für Bildung und Wissenschaft; the Fundaçao Calouste Gulbenkian; the International Centre for Genetic Engineering and Biotechnology of Trieste; the Bull HN Information Systems Italy; the Ministerio de Asuntos Exteriores from Spain; the Comisión Interministerial de Ciencia y Tecnología of Spain; and the Groupement de Recherche et d'Etude sur les Génomes.

Correspondence and requests for materials should be addressed to A. G. (e-mail: goffeau@fysa.ucl.ac.be).

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX

C. Churcher[1], S. Bowman[1], K. Badcock[1], A. Bankier[2], D. Brown[1], T. Chillingworth[1], R. Connor[1], K. Devlin[1], S. Gentles[1], N. Hamlin[1], D. Harris[1], T. Horsnell[2], S. Hunt[1], K. Jagels[1], M. Jones[1], G. Lye[1], S. Moule[1], C. Odell[1], D. Pearson[1], M. Rajandream[1], P. Rice[1], N. Rowley[2], J. Skelton[1], V. Smith[2], S. Walsh[1], S. Whitehead[1] & B. Barrell[1]

[1]*The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[2]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

Large-scale systematic sequencing has generally depended on the availability of an ordered library of large-insert bacterial or viral genomic clones for the organism under study. The generation of these large insert libraries, and the location of each clone on a genome map, is a laborious and time-consuming process. In an effort to overcome these problems, several groups have successfully demonstrated the viability of the whole-genome random 'shotgun' method in large-scale sequencing of both viruses and prokaryotes[1–5]. Here we report the sequence of S*accharomyces cerevisiae* chromosome IX, determined in part by a whole-chromosome 'shotgun', and describe the particular difficulties encountered in the random 'shotgun' sequencing of an entire eukaryotic chromosome. Analysis of this sequence shows that chromosome IX contains 221 open reading frames (ORFs), of which approximately 30% have been sequenced previously. This chromosome shows features typical of a small S*accharomyces cerevisiae* chromosome.

The sequence derived for chromosome IX is 439,886 nucleotides in length, and 71.6% codes for proteins or predicted proteins. There are 219 non-overlapping ORFs equal to or greater than 100 amino acids long, and a further two ORFs (YIL060W and YIL059C) that overlap; these are short, and both have a low codon adaptation index (CAI). Although it is unlikely that both are coding, one could not be selected above the other as more likely to encode a protein. A single Ty3-2 retrotransposon containing three ORFs is present on the left arm of chromosome IX (between bases 205,217 and 210,644), leaving 218 *S. cerevisiae*-derived ORFs encoded on this chromosome, of which 116 are on the Crick strand, and 102 (+ 3 transposon ORFs) are on the Watson strand. Of these, 66 (30.3%) have been sequenced previously. A further 68 (31.2%) have some similarity to genes in *S. cerevisiae* and other organisms for which some functional information is available. However, 74 (33.9%) of the predicted genes on this chromosome cannot be assigned even a putative function based on sequence similarity. These can be divided into two groups: those that show no similarity to current database entries (53, 24.3%), and those that are similar to predicted genes of unknown function (21, 9.6%). The remaining 10 (4.6%) are putative pseudogene ORFs.

The average length of a chromosome IX ORF is 476 codons, with an average of one ORF every 1,993 base pairs. The largest ORF on chromosome IX is YIL129C, which encodes a hypothetical protein of 2,376 amino acids. The YIL129C protein is similar to another hypothetical protein encoded on *Caenorhabditis elegans* chromosome III (EMBL database, accession numbers CEF21H11, U11279 and ORF F21H11.2) over a region of 2,009 amino acids. In total, 20 chromosome IX ORFs are longer than 1,000 codons. Short *S. cerevisiae* genes with no homology are difficult to detect[6]. On chromosome IX, five ORFs with less than 100 codons have been identified, but future analysis will probably reveal additional short coding regions. Less than 4% of the ORFs on chromosome IX are predicted to be spliced; eight ORFs contain introns. None of the tRNA genes on this chromosome are spliced.