The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV

C. Jacq[1], J. Alt-Mörbe[2], B. Andre[3], W. Arnold[4], A. Bahr[5], J. P. G. Ballesta[6], M. Bargues[7], L. Baron[8], A. Becker[4], N. Biteau[8], H. Blöcker[9], C. Blugeon[1], J. Boskovic[6], P. Brandt[9], M. Brückner[10], M. J. Buitrago[11], F. Coster[12], T. Delaveau[1], F. del Rey[11], B. Dujon[13], L. G. Eide[14], J. M. Garcia-Cantalejo[6], A. Goffeau[12], A. Gomez-Peris[15], C. Granotier[8], V. Hanemann[16], T. Hankeln[5], J. D. Hoheisel[17], W. Jäger[9], A. Jimenez[6], J.-L. Jonniaux[12], C. Krämer[5], H. Küster[4], P. Laamanen[18], Y. Legros[8], E. Louis[19], S. Möller-Rieker[5], A. Monnet[8], M. Moro[20], S. Müller-Auer[10], B. Nußbaumer[4], N. Paricio[7], L. Paulin[18], J. Perea[1], M. Perez-Alonso[7], J. E. Perez-Ortin[15], T. M. Pohl[21], H. Prydz[14], B. Purnelle[12], S. W. Rasmussen[22], M. Remacha[6], J. L. Revuelta[11], M. Rieger[10], D. Salom[15], H. P. Saluz[16], J. E. Saiz[11], A.-M. Saren[18], M. Schäfer[10], M. Scharfe[23], E. R. Schmidt[5], C. Schneider[20], P. Scholler[17], S. Schwarz[17], A. Soler-Mira[6], L. A. Urrestarazu[3], P. Verhasselt[24], S. Vissers[3], M. Voet[24], G. Volckaert[24], G. Wagner[10], R. Wambutt[23], E. Wedler[23], H. Wedler[23], S. Wölfl[16], D. E. Harris[25], S. Bowman[25], D. Brown[25], C. M. Churcher[25], R. Connor[25], K. Dedman[25], S. Gentles[25], N. Hamlin[25], S. Hunt[25], L. Jones[25], S. McDonald[25], L. Murphy[25], D. Niblett[25], C. Odell[25], K. Oliver[25], M. A. Rajandream[25], C. Richards[25], L. Shore[25], S. V. Walsh[25], B. G. Barrell[25], F. S. Dietrich[26], J. Mulligan[26], E. Allen[26], R. Araujo[26], E. Aviles[26], A. Berno[26], J. Carpenter[26], E. Chen[26], J. M. Cherry[26], E. Chung[26], M. Duncan[26], S. Hunicke-Smith[26], R. Hyman[26], C. Komp[26], D. Lashkari[26], H. Lew[26], D. Lin[26], D. Mosedale[26], K. Nakahara[26], A. Namath[26], P. Oefner[26], C. Oh[26], F. X. Petel[26], D. Roberts[26], S. Schramm[26], M. Schroeder[26], T. Shogren[26], N. Shroff[26], A. Winant[26], M. Yelton[26], D. Botstein[26], R. W. Davis[26], M. Johnston[27], L. Hillier[27], L. Riles[27] and other members of the Genome Sequencing Center[27], K. Albermann[28], J. Hani[28], K. Heumann[28], K. Kleine[28], H. W. Mewes[28], A. Zollner[28] & P. Zaccaria[28]

[1]*Laboratoire de Génétique Moléculaire, URA 1302 du CNRS, Ecole Normale Supérieure, 46 rue d'Ulm 75230 Paris Cedex 05, France*
[2]*Labor für DNA-Analytik, Wipperstrasse 2, 79100 Freiburg, Germany*
[3]*Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Campus Plaine CP244, Boulevard du Triomphe, 1050 Bruxelles, Belgium*
[4]*Lehrstuhl für Genetik, Fakultät für Biologie, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany*
[5]*Institut für Molekulargenetik, Universität Mainz, Becherweg 32, D-55099 Mainz, Germany*
[6]*Centro de Biologia Molecular, CSIC and UAM, Canto Blanco, 28049 Madrid, Spain*
[7]*Department of Genetics, University of Valencia, Campus of Burjasot, E-46100 Burjasot, Spain*
[8]*Pharmacia Biotech, Parc technologique, rue R. Razel, 91898 Orsay Cedex, France*
[9]*GB-Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany*
[10]*Genotype GmbH, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany*
[11]*Departemento de Microbiologia y Genetica, Universidad de Salamanca, Avenida del Campo Charro E-37007 Salamanca, Spain*
[12]*Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud, 2/20, 1348 Louvain-la-Neuve, Belgium*
[13]*Unité de Génétique Moléculaire des Levures, Département de Biotechnologies, Institut Pasteur, F-75724 Paris Cedex 15, France*
[14]*The Biotechnology Centre of Oslo, Ganstadalleen 21, 0371 Oslo, Norway*
[15]*Departamento de Bioquimica y Biologia Molecular, Facultad de Biologia, Universidad de Valencia, E-46100 Burjasot, Spain*
[16]*Hans-Knöll-Institut, Beutenbergstrasse 11, D-07745, Jena, Germany*
[17]*Molecular Genetic Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, D-69210 Heidelberg, Germany*
[18]*DNA Syntheses and Sequencing Laboratory, Institute of Biotechnology, University of Helsinki, P.O. Box 56, Viikinkaari 9, FIN-00014, Helsinki, Finland*
[19]*Department of Yeast Genetics, Institute of Molecular Medecine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK*
[20]*L.N.C.I.B., Area Science Park, Padriciano 99, I-34012 Trieste, Italy*
[21]*GATC GmbH, Fitz-Arnold-Strasse 23, 78467 Konstanz, Germany*
[22]*Carlsberg Laboratory, Gamle Carlsberg Vej 10, DK-2500 Copenhagen Valby, Denmark*
[23]*AGON GmbH, Glienicker Weg 185, D-12489 Berlin, Germany*
[24]*Katholieke Universiteit Leuven, Laboratory of Gene Technology, Willem de Croylaan, 42, B-3001 Leuven, Belgium*
[25]*The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[26]*Department of Biochemistry, Stanford University, Beckman Center, Stanford CA 94305-5307, USA*
[27]*The Genome Sequencing Center, Department of Genetics, Washington University, School of Medicine, 630 S. Euclid Avenue, St Louis, Missouri 63110, USA*
[28]*Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, D-82152 Martinsried bei München, Germany.*

The complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome IV has been determined. Apart from chromosome XII, which contains the 1–2 Mb rDNA cluster, chromosome IV is the longest *S. cerevisiae* chromosome. It was split into three parts, which were sequenced by a consortium from the European Community, the Sanger Centre, and groups from St Louis and Stanford in the United States. The sequence of 1,531,974 base pairs contains 796 predicted or known genes, 318 (39.9%) of which have been previously identified. Of the 478 new genes, 225 (28.3%) are homologous to previously identified genes and 253 (32%) have unknown functions or correspond to spurious open reading frames (ORFs). On average there is one gene approximately every two kilobases. Superimposed on alternating regional variations in G+C composition, there is a large central domain with a lower G+C content that contains all the yeast transposon (Ty) elements and most of the tRNA genes. Chromosome IV shares with chromosomes II, V, XII, XIII and XV some long clustered duplications which partly explain its origin.

The technique of determining the DNA sequence of large genomes has been unchanged for 21 years[1]. Sequencing the yeast genome required considerable organization by the European Union, which initiated the grouping of 35 laboratories to sequence the first yeast chromosome[2] and coordinate an international effort to sequence the others. Chromosome IV had already been characterized both physically and genetically[3,4], and our sequence data presented are in good agreement with these preliminary data.

The average base composition of chromosome IV is 37.9% G+C, which is lower than for most of the yeast chromosomes (for example, 38.5% for chromosome III (ref. 2) and 38.3% for chromosome II (ref. 5)). Along the 1,513,914 base pairs of the chromosome there are alternating regions about 50 kilobases long of high and low G+C content (Fig. 1). This periodicity is not clearly associated with a variation in gene density, as has been observed for some other chromosomes[6,7]. The central domain of chromosome IV (coordinates 500,000 to 1,215,000) has a much lower G+C value (37.4%) than the two flanking regions (38.2%); a similar observation has been made for the much smaller chromosome VI(ref. 8).

The low G+C content of the central domain seems to be correlated with the presence of Ty elements. All nine Ty1 or Ty2 elements, including a truncated form of Ty1, are localized between coordinates 450,000 and 1,190,000 (Fig. 1). Yeast transposons seem to insert into specific chromosomal regions[9,10] where they are localized preferentially upstream of tRNA genes, as they might interact with the RNA polymerase III machinery[11]. The density of tRNA genes in the central domain of chromosome IV is twice that in the flanking regions in which no Ty elements are found. A total of 27 tRNA genes are localized on each strand of the chromosome, 17 of which are located in the central domain. Of the 27 tRNA genes, 18 are in the vicinity of long terminal repeats (LTRs). Thus most of the tRNA genes, LTRs and Ty elements,

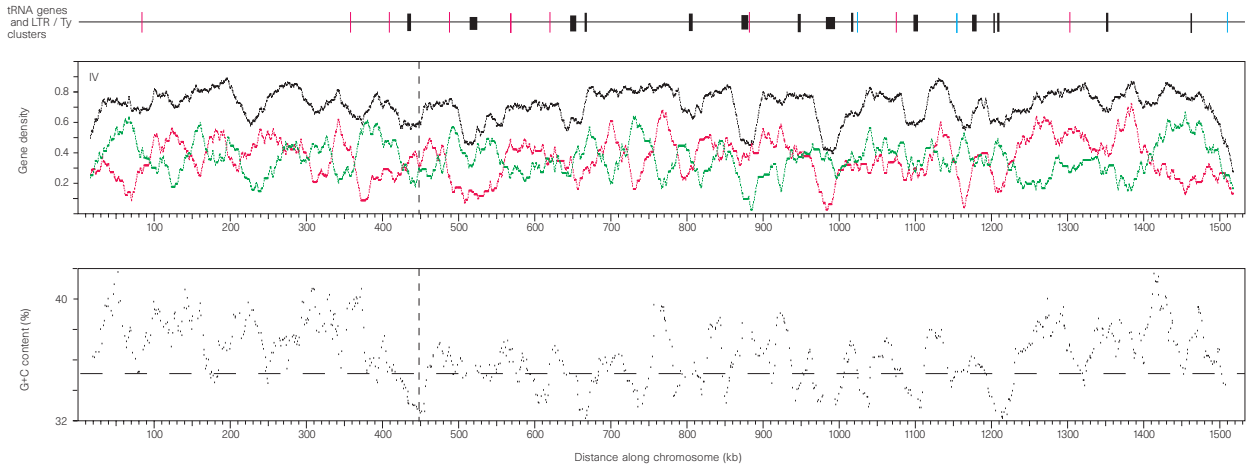

Figure 1 Overall molecular architecture of chromosome IV shows positions of tRNA genes, solo LTR or Ty elements (thin vertical lines), or clusters of them (thick vertical lines), along the chromosome map. Panels show variation of gene density (top) and base composition (bottom) along the sequence-based map of chromosome IV (scale in kilobases from the left telomere). Vertical broken lines indicate the centromere. Gene density is expressed as the probability for each nucleotide to be part of an ORF, andwas calculated using sliding windows of 30 kb (in steps of 0.5 kb) for the Watson strand alone (red line), the Crick strand alone (green line), and the sum of both (black line). G+C content was calculated from the silent positions of codons using a sliding window of 13 consecutive ORFs; the horizontal broken line indicates average G+C content (%) at silent positions of codons.

together with a lower G+C content, are found in this central domain. In chromosome II the 13 tRNA genes and three Ty elements are in AT-rich regions[5].

The left telomere of chromosome IV is very similar to other yeast telomeres. Adjacent to the $C_{1-3}A$ repeat are the usual STR-A, STR-B, STR-C, STR-D and the core X elements (435–904) shared by most of the telomeres[12]. The left end of chromosome IV shares with the right end of chromosome X a large, nearly identical block of sequence similarity more than 19 kilobases long. This duplication includes five ORFs, which code for almost identical products. Indeed the sequences are so similar that we needed to exclude the possibility of contamination of the cosmid contig of chromosome IV by DNA sequences from chromosome X. To confirm our data, we established the genomic sequence of the junction between the duplicated sequence and the rest of the chromosome. Such subtelomeric duplications have often been observed in the yeast genome, suggesting either recent or continuous exchange of genetic information[13]. The right telomere has a less conventional structure with an internal TG_{1-3} repeat.

Using the classical definition of ORFs (one ATG codon followed by at least 99 sense codons), 776 ORFs were recorded in the chromosome; there are also 20 ORFs shorter than 99 amino acids long, making a total of 796 ORFs. Small ORFs of between 25 and 99 codons were extracted and analysed for different properties (codon usage, homologies and ATG environment) to determine their function[7]; 15 had either a putative translation product that is homologous to proteins of other genes, or a codon adaptation index (CAI) greater than 0.2. Five other short ORFs, longer than 91 amino acids, are thought to be 'questionable'. These results support the choice of the threshold of more than 99 codons[14], but show that some short ORFs must be considered[15]. Moreover, this evaluation has to take into account that at least 7% of the ORFs can be considered 'questionable'[6]. Disregarding the retrotransposons, this corresponds to a gene density of one ORF per 2,000 base pairs. The G+C regions correlate roughly, although not precisely, with the regions of increased density. There are approximatively equal numbers of ORFs on the DNA strands (387 on the Watson and 409 on the Crick). However, the gene density is clearly not uniform on a given DNA strand (Fig. 1). This marked preference for an arrangement in which genes are on the same strand has already been observed for other chromosomes[16,17] for which long runs of genes on one strand could be observed. Such gene-rich regions are mainly visible at the two ends of the chromosomes, with more uniform gene density in the central part of the chromosome (550,000 to 1,200,000). Although not as great as predicted by an approximate calculation[18], the high number of Ty and LTR elements in that region should give rise to large numbers of genome rearrangements (inversions, deletions and reciprocal translocations), which could explain this difference with the flanking regions. The construction of gene arrangement has led to general features that probably reflect important functional constraints. Thus the size of regions between ORFs is clearly dependent on the orientation of the flanking genes. In the case of divergent promoters, the mean size is 744 base pairs, whereas it is just 324 bp for convergent terminators. An intermediary situation (593 bp) is found for terminator–promoter combinations. This striking difference in inter-ORF size is probably due to the sequence requirements in the promoter regions for the regulation of gene expression. Nevertheless, the mean size of the inter-ORF region in the case of head-to-head gene orientations is small, and suggests that many divergent genes share common regulation signals. The *GAL1–GAL10* promoter was the first to be described[19], but many other candidates for common regulation have been revealed by the systematic genome sequencing.

Based on the canonical sequences known to control the splicing process, 30 introns can be identified in genes coding for proteins[20]. This represents 4% of genes having an intron in their ORF or in the 5′ untranslated region, a figure close to the situation for the genome as a whole. Of these intron-containing genes, 12 code for ribosomal proteins, three for proteins of the actin family, three for proteins involved in the ubiquitin-dependent protein degradation system, and the rest are distributed between genes that do not necessarily have a high CAI.

Although dependent on the criteria used to estimate the significance of sequence similarities, roughly 30% of the 796 ORFs of chromosome IV are orphans[21] with no sequence relatives in the available databases. This is one of the most exciting findings from the systematic sequencing approach of the yeast genome. Future work will tell us whether some of these genes are really 'yeast specific' and why they have escaped detection by the genetic approaches. The number of sequence orphans will no doubt decrease with the arrival of new sequence data[22]. As an example from chromosome IV, the ORF YDL120w, which had no relatives, is homologous to the human gene recently discovered to be involved in Friedreich's ataxia[23].

Chromosome IV is the longest chromosome in terms of coding sequences, and so might be expected to have features that are scattered on the other chromosomes. One of these features might be the non-uniform organization of the chromosome. The central domain (from 500,000 to

1,250,000) which makes up half of chromosome IV has several distinctive features. First, as with most of the chromosomes, it has more or less regularly spaced regions rich in G+C, but its central domain has a lower G+C content. Second, this central domain contains all of the Ty and most of the LTR elements found on the chromosome. Third, the central domain also contains 18 of the 27 tRNA genes, so its tRNA gene density is twice that of the rest of the chromosome. Finally, the DNA strand distribution of the ORFs is different in the central region when compared with that of the flanking regions. The ORF arrangement of this region might result from a greater genetic plasticity.

Analysis of structural relationships inside the yeast genome might provide an insight into eukaryotic genome organization and evolution. Redundancy is one of the most salient features of the yeast genome structure[24], and the DNA sequence of the whole yeast genome reveals several types of redundancy, probably originating from different biological processes. The most common form of redundancy involves individual genes that have a homologue in the genome; about 20% of the genes of chromosome IV are in this class. Second, there are clusters of very similar copies of a gene, often arranged in tandem; for example, there are five copies of the *ENA1* (or *PMR2*) gene on chromosome IV. Third, subtelomeric duplications are frequent and involve large regions of chromosomes that are very similar in both coding and non-coding regions[13]. Finally, clustered duplications are characterized by clusters of homologous genes in the same order, usually in the same orientation, and interspersed by long DNA fragments. Such paralogous regions have already been described between chromosomes III and XIV (ref. 25) and between chromosomes V and X (ref. 26). Only in the case of the duplication between chromosomes III and XIV is the gene order conserved. It is 15 kilobases long and contains four genes. The clustered duplications on chromosome IV are made up of at least 336 kilobase pairs, including 49 pairs of homologous genes. Chromosome IV shares large ordered cluster of homologous genes with chromosomes II, V, VIII, XII and XIII (ref. 27). A careful analysis of these duplications will no doubt tell us a great deal about the evolution of the yeast genome. In the largest interchromosomal clustered duplications, involving chromosome IV (coordinates 449,752–569,763) and chromosome II (238,164–407,122), the 18 gene pairs are all transcribed in the same direction. When known, most of the genes from a pair code for proteins with homologous but not identical functions (for example *GAL1* and *GAL3*)[19]. Homologous genes from a clustered duplication can also be completely identical or totally different in their function. An extreme case of divergence involves YDR037w, which codes for a lysyl tRNA synthetase, and YBR060w_A, which is part of chromosome II, in which many stop codons interrupt an ORF, of which parts are homologous to YDR037w. Such a pseudogene could not be detected by searching the DNA sequence of chromosome II, as it has very short ORFs and no ATG codon. To our knowledge, this is the most degenerated yeast pseudogene yet discovered. This observation suggests that similar degenerated pseudogenes may have escaped previous analyses, and hence that the total number of pseudogenes may be underestimated.

A pair of genes from a clustered duplication can also differ in the presence of an intron. Both YDR055w and YBRO78w are homologous to the gene *SPS2*, but only YBR078w has an intron, and the CAI of the two genes differs from 0.27 (YDR055w) to 0.61 (YBR078w), suggesting an unusual evolutionary process. The compared analysis of the interspersed DNA fragments is also very informative. For example, chromosome II has a Ty element where an LTR element is present at the equivalent position on chromosome IV, suggesting that the Ty element was lost from chromosome IV after the duplication process.

The greatly different degrees of similarities between the different gene pairs composing a duplicated region indicate that at least some of the duplications have evolved at very different rates, suggesting in some cases that gene conversion processes[28] have interfered with slower evolutionary processes. A careful quantitative analysis of the relative evolution rates of the different elements will be required to establish a chronological order of the different events. Nevertheless, evidence suggests that a first duplication event has been followed by the dispersal of the duplicated elements by the insertions of DNA fragments of various sizes and gene compositions. Most of these clustered duplications in chromosome IV are localized in the pericentromeric region. The centromere itself is included in the longest duplicated region, which occurs between chromosomes IV (coordinates 450,000–570,000) and II (238,000–407,000). Similar localizations of clusters have already been noticed on other chromosomes[7,17,25]. To explain the proximity of the centromere, it was suggested that the gene dispersion of the initial cluster of duplicated genes might be slower in the centromeric regions than nearer the telomere owing to the adverse effects of rearrangements on chromosome segregation[25]. Alternatively, the centromeric duplications might have been essential steps in the construction of the yeast genome[25]. In agreement with these ideas, the central domain of chromosome IV contains few traces of clustered redundancies, perhaps because of its genetic plasticity. These preliminary observations indicate that the availability of the complete sequence of the yeast genome will allow a greater understanding of the processes involved in creating the genome architecture. □

Methods

The sequence was assembled from a set of 44 partly overlapping cosmids and lambda phages from two independent contigs of chromosome IV. The 650-kb cosmid contig corresponding to the left part of chromosome IV was constructed mainly from a specific cosmid library obtained from a gel-purified chromosome portion (J. D. H. *et al*., unpublished), and a few other cosmid clones from this contig came from a library[29]. The rest of the chromosome sequence was established from a cosmid-lambda phage library (L. Riles & M. Olson, unpublished, and ref. 30). The two cosmid contigs were made from two closely related yeast strains: AB972, derived from S288C[30], and FY1679, a diploid strain issued from the cross between FY23 and FY73, both of which are isogenic with S288C except for the markers indicated. Sequence analysis of a large overlapping fragment (of about 170 kb) confirmed that the strains AB972 and FY1679 are very similar, as the number of base differences was below the estimated error rate. Only the extreme left telomeric regions of the two strains clearly differ, probably in the number of their TG_{1-3} repeats (C. B. and C. J., unpublished). The telomeres were isolated independently and sequenced 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[13]. Two gaps in the 650-kb left cosmid contig (constructed from the strain FY1679) were filled with lambda clones from the library constructed from AB972. They correspond to the regions 9,756–11360 and 363,100–368,150 . The left 600-kb region was sequenced according to the rules followed by the European consortium and the 20 cosmids and phages were distributed to 18 contractors, whereas the central part was sequenced by the Sanger Centre (EMBL database SCCHRIV, accession no. Z71256); the rest of the chromosome was sequenced by groups from Washington University in St Louis and Stanford University in the United States.

There were very few base differences in the overlapping fragment sequenced in parallel by the Sanger Centre and by the European consortium, demonstrating that both approaches are reliable. However, a verification procedure was necessary because of the greater heterogeneity of the European approach. This was done on 25 regions of the left part of the chromosome, according to the protocol of G. V. (manuscript in preparation). This allows direct polymerase chain reaction (PCR) sequencing of a 300-bp region of the yeast genome limited by two previously designed oligonucleotides. We could thus correct a sequence in which a bacterial transposon had been inserted during the cloning process but no real sequence error could be detected at this final step of the sequencing project. Sequence errors could only be corrected after examination of the raw sequence data. From these data, the error rate of this part of the yeast chromosome IV sequence presented is less than four errors per 10 kb. In the central part of the chromosome the error rate is estimated as less than one error per 10 kb. Specific strategies were developed to sequence difficult parts of the chromosome. Thus, for example, to finish the regions between the two transposons located in cosmid 8142 (http://www.sanger.ac.uk/~yeastpub/svw/sequencing.html), a PCR product covering this region from strain MCYC2576 was sequenced. This strain, a gift from E. Louis, did not contain the transposons.

Received 24 July 1996; accepted 11 March 1997.

1. Sanger, F. & Coulson, A.R. *J. Mol. Biol.* 94, 441–447 (1975).
2. Oliver, S. *et al. Nature* 357, 38–46 (1992).
3. Mortimer, R. K. *et al*. http://genome-www.stanford.edu/sacchdb/edition 12.html (1995).
4. Olson, M. V. in *The Molecular and Cellular Biology of the Yeast* Saccharomyces*: Genome Dynamics, Protein Synthesis and Energetics* (eds Broach, J. R., Jones, E. W. and J. R. Pringle) 1–39 (Cold Spring

Harbor Laboratory Press, NY, 1991.
5. Feldmann, H. *et al. EMBO J.* 13, 5795–5809 (1994).
6. Dujon, B. *et al. Nature* 369, 371–378 (1994).
7. Galibert, F. *et al. EMBO J.* 15, 2031–2049 (1996).
8. Murakami, Y., Naitou, M., Hagiwarar, H. & Shibata, T. *Nature Genet.* 10, 261–268 (1995).
9. Ji, H. *et al. Cell* 73, 1007–1018 (1993).
10. Lochmüller, H., Stucka, R. & Feldmann, H. *Curr. Genet.* 16, 247–252 (1989).
11. Voytas, D .F. & Boeke, J. D. *Trends Genet.* 9, 421–427 (1993).
12. Pryde, F. E., Huckle, T. C. & Louis, E. J. *Yeast* 11, 371–382 (1995).
13. Louis, E. J. & Borts, R. H. *Genetics* 139, 125–136 (1995).
14. Termier, M. & Kalogeropoulos, A. *Yeast* 12, 369–384 (1996).
15. Navarre, C., Caty, P., Leterme, S., Dietrich, F. & Goffeau, A. *J. Biol. Chem.* 267, 21262–21268 (1994).
16. Bussey, H. *Proc. Natl Acad. Sci. USA* 92, 3809–3813 (1995).
17. Johnston, M. *et al. Science* 265, 2077–2081 (1994).
18. Kupiec, M. & Petes, T.D. *Mol. Cell. Biol.* 8, 2942–2952 (1988).
19. Johnston, M. & Davis, R.W. *Mol. Cell. Biol.* 4, 1440–1448 (1984).
20. Kalogeropoulos, A. *Yeast* 11, 555–565 (1995).
21. Dujon, B. *Trends Genet.* 12, 263–270 (1996).
22. Casari, G., De Daruvar, A., Sander, C. & Schneider, R. *Trends in Genet.* 12, 244–245 (1996).
23. Campuzano, V. *et al. Science* 271, 1423–1427 (1996).
24. Oliver, S.G. *Nature* 379, 597–600 (1996).
25. Lalo, D., Stettler, S., Mariotte, S., Slonimski, P. P. & Thuriaux, P. *C. R. Acad. Sci.* 316, 367–373 (1993).
26. Melnick, L.M. & Sherman, F. *J. Mol. Biol.* 233, 372–388 (1993).
27. Heumann, K. & Mewes, H. W. *Nature Genet.* (submitted).
28. Jinks-Robertson, S. & Petes, T.D. *Proc. Natl Acad. Sci. USA* 82, 3350–3354 (1985).
29. Thierry, A., Gaillon, L., Galibert, F. & Dujon, B. *Yeast* 11, 121–135 (1995).
30. Riles, L. *et al. Genetics* 134, 81–91 (1993).

Acknowledgements. The laboratory consortium operating under contracts with the European Commission was initiated and organized by A. G. A. Vassaroti and P. Mordant were in charge of the administrative coordination. This work was supported by the European Commission under the Biotech programmes, the Groupe de Recherche et d'Etudes sur le Génome, the Centre National de la Recherche Scientifique, the Wellcome Trust, the Région Wallone, the Belgian Federal Services for Science Policy, the Research Fund of the Katholieke Universiteit Leuven, the Région de Bruxelles Capitale, the Fundacion Ramon Aceres and Comision Interministerial de Ciencia y Tecnologia and the Bundesminister für Forschung und Technologie.

Correspondence and requests for materials should be addressed to C.J. (e-mail: jacq@biologie.ens.fr).

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V

F. S. Dietrich*, J. Mulligan*, K. Hennessy, M. A. Yelton*, E. Allen, R. Araujo, E. Aviles, A. Berno, T. Brennan, J. Carpenter, E. Chen, J. M. Cherry, E. Chung, M. Duncan, E. Guzman, G. Hartzell, S. Hunicke-Smith, R. W. Hyman, A. Kayser, C. Komp, D. Lashkari, H. Lew, D. Lin, D. Mosedale, K. Nakahara, A. Namath, R. Norgren, P. Oefner, C. Oh, F. X. Petel, D. Roberts, P. Sehl, S. Schramm, T. Shogren, V. Smith, P. Taylor, Y. Wei, D. Botstein & R. W. Davis

Stanford DNA Sequencing and Technology Center, 855 California Avenue, Palo Alto, California 94304 and Departments of Biochemistry and Genetics, Stanford University Medical School, Stanford, California 94305, USA
**These authors were in charge of the project at different times during the study.*

Here we report the sequence of 569,202 base pairs of *Saccharomyces cerevisiae* chromosome V. Analysis of the sequence revealed a centromere, two telomeres and 271 open reading frames (ORFs) plus 13 tRNAs and four small nuclear RNAs. There are two Ty1 transposable elements, each of which contains an ORF (included in the count of 271). Of the ORFs, 78 (29%) are new, 81 (30%) have potential homologues in the public databases, and 112 (41%) are previously characterized yeast genes.

As part of an international collaborative effort to sequence the total genome of the yeast *Saccharomyces cerevisiae,* we have deduced the DNA sequence of 569,202 base pairs of yeast chromosome V. We used an overlapping set of recombinant yeast cosmid and lambda clones that together cover the entire chromosome (except for the extreme ends of the telomeres). A line drawing of chromosome V and the identification of the recombinant DNAs sequenced are shown in Fig. 1. The sequence was broken arbitrarily into 11 slightly overlapping pieces for ease of handling and deposited in Genbank (see Fig. 1 for accession numbers).

Sequencing was accomplished in two phases: the 'shotgun' phase, using dye-primer chemistry, and the 'finishing' phase, using the polymerase chain reaction (PCR) and dye-terminator chemistry. There were no gaps in the sequence at the end of shotgun sequencing and assembly. The assembled, continuous sequence of chromosome V has 569,202 bp, starting from the guanine residue of the *Sau*3A site on the left vector boundary of the leftmost clone (1160 in Fig. 1). The 569-kilobase sequence is based on the results from 32,631 individual lanes of sequencing gels, or reads. The average depth of coverage was 12.5-fold. The minimum acceptable coverage was three, with at least one read from each strand.

After shotgun sequencing and assembly, problems remained in the sequence at a frequency of (roughly) two per kilobase and were of several types. They included the inability to count unambiguously the number of repeating units, such as poly (dA), and guanine compressions. There were also small regions in which only one of the two strands had been sequenced. These difficulties were resolved during the finishing phase.

After finishing, the 569-kb contig was checked against three external sets of data. First was the use of tetrad segregation data to derive a genetic map for yeast[1]. The chromosome V gene order based on DNA sequence was in complete agreement with the tetrad segregation data. There were two locations on the genetic map (*CEN*V at 151 kb and *PRO*3 at 200 kb) where closely spaced loci had been mapped against distant markers and not against each other, resulting in ambiguities of relative locus order[1], which were resolved using the DNA sequence. The gene order across the centromere is *GLC*3 tRNA-Arg *GCN*4 *CEN*V *MNN*1. In the region of *PRO*3, at 200 kb, the gene order is *PRO*3 *GPA*2 *GCD*11 *CHO*1 *GAL*83. Second, our sequence was compared to the *S. cerevisiae* sequences already deposited in Genbank, using both the FASTA and BLAST programs[2,3]. In the rare cases of sequence difference, we re-examined our trace files. Remaining ambiguities were resolved using the same methodology as finishing. Third, we checked our data against the primary *Eco*RI/*Hin*dIII double-digestion fragment maps of the recombinant yeast DNAs[4]. Our sequence was examined for *Eco*RI and *Hin*dIII cleavage sites. Of 534 mapped fragments, there were only five discrepancies, which is a tribute to the care taken in preparing the cleavage sites map[4]. The five apparent discrepancies between the double-digest map[4] and our sequence are: the map had doublets where the sequence predicts singlets after bases 272, 193; 280,936; and 441,102; the map has a fragment that was not found in the sequence after base 414,946; and the sequence is missing a cleavage site after base 506,807.

We examined all six possible reading frames of the 569-kb sequence for ORFs of at least 300 bp that began with a start codon and ended with a stop codon. As a special case, an ORF could be interrupted if there were yeast splice donor/acceptor/branchpoint sequences present at the appropriate intervals. The remaining sequence was examined using FASTA and BLAST for homology to sequences in the public databases. This enabled us to find small ORFs, as well as the centromere, 13 tRNAs, two Ty1 elements (which each contain an ORF), four small nuclear RNAs, many delta and delta-like elements, and the highly conserved X and Y' sequences characteristic of yeast telomeres (see refs 5, 6) at the far left and right ends.

Initially, 271 ORFs were identified in the 569-kb sequence, although this number has changed as evaluation continued. The 271 ORFs make up roughly 70% of the sequence, with an average of 2.1 kb per ORF. The 'average' ORF (1.4 kb) encodes 475 amino acids. Of the ORFs, 112 (41%) have been characterized previously, 81 (30%) have apparent homologues in the public databases, and 78 (29%) are new; six (2%) are spliced. Of the 81 apparent homologues, 55 of these are to other *S. cerevisiae* sequences.

The fractional G+C content of the 569,202 bp of chromosome V is 0.384. The combined ORF DNAs have a fractional G+C content of 0.401, and the combined 'non-ORF' DNA has a G+C content of 0.351.