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.

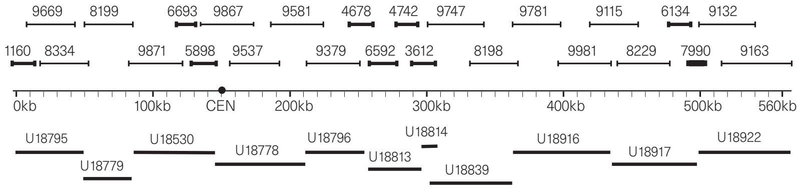

Figure 1 The central line representing *S. cerevisiae* chromosome V is marked in kilobase pairs, starts at the left at the guanine of the *Sau*3A site of the leftmost recombinant yeast DNA, and extends to the right to 570 kb. The centromere, *CEN*, is represented by a solid circle at ~151 kb. Above the line, placed across their map positions, are the individual recombinant yeast DNAs that were sequenced; 16 cosmids (thin lines), 8 lambdas (thick lines), and 1 plasmid (very thick line). Genbank accession numbers are placed below the line, above the bars indicating the corresponding map positions. From left to right, Genbank accession numbers are U18795, U18779, U18530, U18778, U18796, U18813, U18814, U18839, U18916, U18917 and U18922. There is some deliberate overlap between Genbank entries to maintain contiguity.

There are only two places on chromosome V where the quality of the sequence is not high. In the first case, at about 312 kb, there are ~50 bp of unique sequence bounded on both sides by poly (dA): poly (dT). *Taq* polymerase, and the other DNA polymerases tested, frequently terminated within the homopolymer, and seldom reached the short unique sequence. Therefore, we have only a few reads across the unique sequence. In the second case, at about 450 kb, there is a 5-kb stretch that contains many delta and delta-like sequences interspersed with a small amount of unique sequence. The clone containing this segment was shotgun sequenced to an average of 16-fold redundancy, yet there were relatively few reads in this region. Therefore, for PCR amplification, this 5-kb region was divided into many virtual parts, based on the positions of the unique sequences. Several custom primer pairs, and internal sequencing primers, were designed and synthesized for each part[7]. These were used in PCR amplification reactions with total yeast genomic DNA as the template. We have sequenced carefully across this region. For most bases, there is sequence from both strands.

There were three special cases that warrant further attention. First, a point mutation had occurred during either cloning or subsequent propagation in *Escherichia coli*. In an overlap region shared by two recombinant DNAs (lambda 5898 and cosmid 9867; Fig. 1), the sequence should be the same, but in this case there was one reproducible base difference. Lambda 5898 has a guanine residue where cosmid 9867 has an adenine residue. When total yeast genomic DNA was used as a template for PCR amplification, the product of which was used as a template for dye-terminator sequencing, the base at that position was an adenine. The traces showed no indication of a naturally occurring polymorphism. We therefore conclude that the guanine in lambda 5898 was the result of a mutation. In the second special case, we examined the ORFs for any apparently premature, in-frame stop codon, and found two that were puzzling. The first was the TGA stop codon at position (rounded-off) 352 kb. There are three rightward-reading frames, and this stop codon (TGA) is in the first of these. Following the TGA in this frame, there is a lysine codon (AAA) and a methionine codon (ATG) followed by a long ORF. In the second reading frame there are three stop codons: two (TGA, TAA) are next to each other, the third is five triplets further on; this frame is truly stopped. However, there is a long ORF starting with a methionine codon considerably upstream and ending at the double stop codons TGA, TAA. The third reading frame has many stop codons. The two ORFs share one base: A, the first base (ATG) of the first reading frame and the last base (AAA) of the second reading frame. In general, yeast ORFs are separated by several hundred bases. Except for a -1 frame shift in the first reading frame, there would be one ORF rather than two. However, despite sequencing through this position many times, the sequence, the TGA and the five A bases in a row remained invariant. The second apparently premature stop codon occurred in the ORF that corresponds to *FLO8* at ~375 kb. The TAG stop codon between YER108c and YER109c appears not just in the recombinant yeast DNA, but also in PCR amplifications from total yeast DNA. Thus it is possible that *FLO8* is not functional in yeast strain AB972. In the third special case, in the entire collection of recombinant yeast DNAs, only lambda 3612 (Fig. 1) covers this region. We found lambda 3612 to be highly unstable, giving rise to non-random DNA deletions at extremely high frequency. Starting with 30 individual plaques from a primary stock, only one yielded lambda 3612 DNA without a detectable deletion, as judged by *Eco*RI/*Hin*dIII double-digestion patterns, but even that gave rise to deleted DNAs upon subsequent growth. Therefore, all of the lambda 3612 sequence came (uncharacteristically) from just one preparation of DNA.

A comparison of the DNA base sequence of chromosome V to that of the other *S. cerevisiae* chromosomes shows that there are two stretches that have similar genes in the same order on two other yeast chromosomes. A portion of the left arm of chromosome V, containing *CYC*7 and *RAD*, shows the same relative gene order as chromosome X, but in the opposite orientation, as noted previously[8]. In addition, a 60-kb region of the left arm of chromosome IX contains nine genes or ORFs for which each has an apparent homologue within a 60-kb region of the right arm of chromosome V. The nine putative protein pairs and their calculated similarity (identicality)[9] are: (1) YIL045w/YER054c, 63 (44) %; (2) YIL050w/YER059w, 71 (52) %; (3) YIL051c/YER057c, 87 (70) %; (4) YIL053w/YER062c, 97 (92) %; (5) YIL056w/YER064c, 63 (47) %; (6) tRNA-ser, 100%; (7) YIL057c/YER067w, 85 (67) %; (8) *RNR*3/YER070w, 90 (82) %; and (9) YIL074c/YER081w, 95 (91) %. On chromosome V itself, the *FCY*2 protein product[10] and the putative YER060w translation product are 87 (75) % related.

When considering the accuracy of our 569,202-bp sequence of *S. cerevisiae* chromosome V, we must emphasize that (essentially) all of the sequence was determined from recombinant DNA propagated in *E. coli*. Even if our sequence of the recombinant DNAs were 100% accurate, there may be sequence differences between the recombinant DNAs and the yeast genome. We identified one apparent point mutation solely because it occurred within a region common to two recombinant DNAs. Other point mutations, occurring during cloning or propagation in *E. coli*, would probably not be detected.

There is a much more dramatic example of a discrepancy between the sequence of a recombinant DNA and the yeast genome. We deposited our chromosome V sequence in Genbank and SacchDB (http://genome-www.stanford.edu/) in December 1994. On 18 April 1996, we received an e-mail from J.-L. Souciet, J. de Montigny and S. Potier (CNRS, Strasbourg) politely telling us that ~2 kb were missing from our sequence. They had found that, in addition to *FCY2* (YER056c; encoding a purine-cytosine permease[10]) at ~267 kb and a closely related ORF (YER060w) at ~275 kb, there is also a third closely related ORF in this region (Genbank accession no. X97346) (Fig 2.). Our sequence across the apparent deletion came from two libraries made from lambda 6592. The sequence is 12-fold deep, and there are no traces that diverge from our Genbank sequence. The Genbank sequence is therefore an accurate sequence of this particular recombinant DNA: lambda 6592. We had intended to sequence cosmid 9380. Repeated attempts to prepare 9380 yielded only minute amounts of DNA. We abandoned cosmid 9380, and instead sequenced three lambda DNAs: 4678, 6592 and 4742. From the tiny amount of cosmid 9380 DNA, we constructed (and sequenced) a *Sau*3A library. The data (traces) from the library of *Sau*3A-cleaved 9380 DNA were used within the assemblies

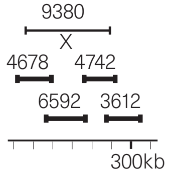

Figure 2 The bottom line is a schematic map of yeast chromosome V from 240 kb to 310 kb. Individual, sequenced recombinant DNAs are placed above the line and across the appropriate map positions. Cosmid 9380 covers map positions ~240 kb to 300 kb and overlaps lambda 3612. The latter is the only recombinant DNA in the entire Olson collection[4,14] that covers map positions 300 kb to 310 kb. Because 9380 could only be produced in minute amounts, three lambda DNAs (thick lines) were substituted: 4678, 6592 and 4742. The X marks the position of the ~2 kb deletion in 6592 DNA.

of the three lambda DNAs. There were some data from the 9380 library that were left over, mostly vector without insert, low-quality traces, *etc*., which we put aside. We searched our 'left-over' 9380 traces for homology to unique sequences in Genbank X97346 and found one excellent (97.3% identity) match for 294 bp (bounded by *Sau*3A sites). We conclude that lambda 6592 has a deletion relative to the yeast genome and that one additional ORF should be added, bringing the ORF total to 272.

To complete the sequence of chromosome V, an insertion of 2,011 bp (Genbank accession number X97346) should be made at base 275,951 of our sequence as has been done in SacchDB. In addition, H. Wedler and R. Wambutt have sequenced the left (Genbank accession no. U73806) and right (Genbank accession no. U34775) telomeres of yeast chromosome V. Within SacchDB, 2,477 bp have been placed to the left of the leftmost base (the G of the leftmost *Sau*3A site) of our sequence. That G is no longer base 1 but base 2,478. Within the left telomere, there is an ORF, YEL077c, which brings the current ORF total to 273. Concomitantly in SacchDB, 3,181 bp have been placed to the right of our sequence.

Basically, there are two types of errors: random and systematic. If there is a random error in an individual sequence read, we will find and correct that error because we sequenced both strands to high redundancy (average of 12.5-fold). It is much more difficult to identify a systematic error that is inherent in, for example, the dye-primer chemistry, polyacrylamide gel electrophoresis, *Taq* polymerase, or base-calling software, that systematically misreads or deletes a base(s) within a particular sequence. *Taq* polymerase seemed to have systematic difficulties synthesizing across short repeating units; not only is the number of repeats often ambiguous, the sequence traces following a repeat are often diminished in signal quality. We believe that this observation reflects an inherent characteristic of *Taq* polymerase. Of the several DNA polymerases tested in an attempt to solve this problem, Amplitaq FS polymerase (Perkin-Elmer 402079) yielded the best number of good sequence calls, but did not solve the problem completely. However, the problem in counting the short repeating units unambiguously may not be a sequencing problem but in some cases may reflect true biological heterogeneity. A second possible systematic error arises from the well-known guanine compressions. Guanine compressions are usually identified when the base-calling software identifies fewer guanines on one strand than cytosines on the opposing strand. However, if two (or more) guanine compressions are positioned symmetrically on opposing strands, the compression on one strand is compensated by an analogous compression on the other strand. There are no 'extra' cytosines, and the existence of the compressions could be missed.

One reason for sequencing all of the *S. cerevisiae* DNA is that yeast is important as a model organism. A second reason is to test the approaches to, and develop technologies for, large-scale DNA sequencing in preparation for the sequencing of the human genome. In this regard, we would like to describe some important lessons learned during the sequencing of yeast chromosome V. First, 800 kb were shotgun sequenced to achieve 569,202 bp of contiguous sequence, an inefficiency of 40%. Considerable time and money would have been saved if the ends of the recombinant yeast DNAs had been mapped relative to each other (a 'sequence-ready' contig of cosmid DNAs). Second, a large amount of freezer space was used in archiving recombinant M13 DNAs, a small percentage of which were later used as templates for finishing. An important reason in the delay of finishing was the cost of oligonucleotide primers for PCR. Finishing has been made economical by the availability of low-cost oligonucleotides[7], so long-term storage of M13 DNAs is no longer necessary. Third, when the Yeast Genome Project was started, the conventional wisdom had that it was necessary to sequence a set of overlapping cosmids. However, we now know that the sequence of DNA as large as bacterial genomes can be assembled using a shotgun approach[11,12]. If we started again, we would purify *S. cerevisiae* chromosome V directly by pulse-field gel electrophoresis, hydrodynamically shear the DNA to an average size of 1 kb (ref 13) and shotgun clone the sheared DNA directly into the M13 sequencing vector. The yeast genome could probably have been sequenced by the direct shotgun cloning of total genomic DNA to generate one M13 sequencing library. Individual cosmid and lambda clones could have been used to fill holes and resolve ambiguities. □

Methods

All of the *S. cerevisiae* recombinant DNAs sequenced in this study were constructed in the laboratory of M. Olson[4,14]. With the exception of plasmid 7990, which was derived from two yeast strains[15], all of the recombinant DNAs were derived from yeast strain AB972. Those recombinant DNAs with number designations less than 8000 are lambdas (except for plasmid 7990), those with numbers over 8000 are cosmids. We sequenced 16 cosmids (8198, 8199, 8229, 8334, 9115, 9132, 9163, 9379, 9537, 9581, 9669, 9747, 9781, 9867, 9871 and 9981), eight lambdas (1160, 3612, 4678, 4742, 5898, 6134, 6592 and 6693), and one plasmid (7990) (Fig. 1). We also obtained some sequence from another five cosmids (8063, 9268, 9380, 9495 which contained a large deletion and 9675) and two lambdas (3955 and 6052). These *S. cerevisiae* recombinant DNAs (except 9495) are available from the American Type Culture Collection.

The 'shotgun' sequencing strategy was to reduce randomly the size of the yeast recombinant DNAs ('inserts') to approximately 1 kb. The inserts were ligated to the M13 sequencing vector by using a 'linker–adaptor' system, which minimizes the formation of chimaeric DNAs. The recombinant M13 'sequencing library' was electroporated into *E. coli* and plated. Individual M13 plaques were picked and grown, and recombinant M13 DNAs were purified. (Our detailed laboratory protocols are freely available on the World-Wide Web at http://sequence-www.stanford.edu.)

The shotgun sequencing used dye-primer chemistry in cycle sequencing reactions, followed by fluorescence detection using an ABI 373A automated sequencer. Most of the sequence data from individual lanes ('traces' or 'reads') were edited automatically using custom software, with borderline cases being edited manually using the TED software[16]. Individual sequence reads were assembled using the XBAP program[16]. The final sequence was determined by editing manually the assembled reads.

Where there seemed to be overlapping ORFs, in either the same or in the opposite direction, the conventional assumption was made that yeast seldom uses both overlapping frames. Three criteria were used to determine which was the most likely ORF. First, using both FASTA and BLAST programs, each of the overlapping reading frames was examined for homology to known genes in the public databases; an ORF with homology was chosen over one without. Second, each organism has its own distinctive preference for certain codons over others. This preference can be expressed in arithmetic terms, as it is within the GeneFinder program (L. Hiller and P. Green, 1990–1993; documentation, software and yeast codon usage data files. Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA). GeneFinder was used to compare the codon usage for each of the overlapping ORFs. The ORF that more closely matched yeast's codon usage was chosen; in almost all cases, this distinction was unequivocal. Third, the longer ORF was selected. The nomenclature for yeast ORFs is composed of five letter/number combinations: Y (for yeast), E (the fifth letter of the alphabet for V), L or R (for the left or right arm, as defined genetically), a number (counted sequentially from the centromere in both directions), and w or c (for the transcribed strand); for example, the *URA3* gene, encoding orotidine-5′-phosphate decarboxylase, is YEL021w.

Received 24 July 1996; accepted 11 March 1997.

1. Mortimer, R. K., Contopoulou, C. R. & King, J. S. *Yeast* 8, 817–902 (1992).
2. Pearson, W. R. & Lipman, D J. *Proc. Natl Acad. Sci. USA* 85, 2444–2448 (1988).
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* 215, 403–410 (1990).

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-Saëns 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-