The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XVI

H. Bussey[1], R.K. Storms[2], A. Ahmed[3], K. Albermann[4], E. Allen[5], W. Ansorge[6], R. Araujo[5], A. Aparicio[5], B. Barrell[7], K. Badcock[7], V. Benes[6], D. Botstein[5], S. Bowman[7], M. Brückner[8], J. Carpenter[5], J.M. Cherry[5], E. Chung[5], C. Churcher[7], F. Coster[9], K. Davis[5], R.W. Davis[5], F.S. Dietrich[5], H. Delius[10], T. DiPaolo[1], E. Dubois[11,12], A. Düsterhöft[13], M. Duncan[5], M. Floeth[13], N. Fortin[1], J.D. Friesen[3], C. Fritz[13], A. Goffeau[9], J. Hall[1], U. Hebling[10], K. Heumann[4], H. Hilbert[13], L. Hillier[14] and other members of the Genome Sequencing Center[14] , S. Hunicke-Smith[5], R. Hyman[5], M. Johnston[14], S. Kalman[5], K. Kleine[4], C. Komp[5], O. Kurdi[5], D. Lashkari[5], H. Lew[5], A. Lin[5], D. Lin[5], E.J. Louis[15], R. Marathe[5], F. Messenguy[11], H.W. Mewes[4], S. Mirtipati[5], D. Moestl[13], S. Müller-Auer[8], A. Namath[5], U. Nentwich[6], P. Oefner[5], D. Pearson[7], F.X. Petel[5], T.M. Pohl[16], B. Purnelle[9], M.A. Rajandream[7], S. Rechmann[6], M. Rieger[8], L. Riles[14], D. Roberts[5], M. Schäfer[8], M. Scharfe[17], B. Scherens[18,19], S. Schramm[5], M. Schröder[5], A. M. Sdicu[1], H. Tettelin[9], L. A. Urrestarazu[20], S. Ushinsky[2], F. Vierendeels[11], S. Vissers[18], H. Voss[6], S.V. Walsh[7], R. Wambutt[17], Y. Wang[1], E. Wedler[17], H. Wedler[17], E. Winnett[1], W-W. Zhong[1], A. Zollner[4], D.H. Vo[1]* & J. Hani[4]*

[1] *Department of Biology, McGill University, 1205 Dr Penfield Avenue, Montreal, H3A 1B1, Canada*
[2] *Department of Biology, Concordia University, Montreal, H3G 1M8, Canada*
[3] *Research Institute, Hospital for Sick Children, Toronto, M5G 1X8, Canada*
[4] *Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany*
[5] *Department of Biochemistry, Stanford University, Stanford, California 94305, USA*
[6] *European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany*
[7] *The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[8] *Genotype GmbH, D-69259 Wilhelmsfeld, Germany*
[9] *Unité de Biochimie Physiologique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium*
[10] *DKFZ, im Neuenheimer Feld 506, D-69120 Heidelberg, Germany*
[11] *Research Institute of CERIA-COOVI, B-1070, Brussels, Belgium*
[12] *Laboratoire de Microbiologie de l'Université Libre de Bruxelles, B-1050, Brussels, Belgium*
[13] *QIAGEN GmbH, Max-Volmer-Strasse 4, D-40724 Hilden, Germany*
[14] *The Genome Sequencing Center, Department of Genetics, Washington University, School of Medicine, 630 S. Euclid Avenue, St Louis, Missouri 63110, USA*
[15] *Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK*
[16] *Gesellschaft für Analyse Technik und Consulting mbH, Fritz-Arnold-Strasse 23, D-78467 Konstanz, Germany*
[17] *AGON, Gesellschaft für molekularbiologische Technologie mbH, Glienicker Weg 185, D-12489 Berlin, Germany*
[18] *Vlaams Interuniversitair Instituut voor Biothechnologie, Department Microbiologie, B-1070, Brussels, Belgium*
[19] *Laboratorium voor Erfelijkheidsleer en Microbiologie van de Vrije Universiteit, Brussels, Belgium*
[20] *Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, B-1050, Brussels, Belgium*
* *These authors contributed equally to the work.*

The nucleotide sequence of the 948,061 base pairs of chromosome XVI has been determined, completing the sequence of the yeast genome. Chromosome XVI was the last yeast chromosome identified[1], and some of the genes mapped early to it, such as *GAL4*, *PEP4* and *RAD1* (ref. 2) have played important roles in the development of yeast biology. The architecture of this final chromosome seems to be typical of the large yeast chromosomes, and shows large duplications with other yeast chromosomes. Chromosome XVI contains 487 potential protein-encoding genes, 17 tRNA genes and two small nuclear RNA genes; 27% of the genes have significant similarities to human gene products, and 48% are new and of unknown biological function. Systematic efforts to explore gene function have begun.

There are 487 open reading frames (ORFs) on chromosome XVI, and 10 Ty-related ORFs. Of these ORFs, 17 have an intron, and ORF RPL6B (YPL198w), which encodes a ribosomal protein, has two introns. ORFs were identified using the working definition that they commence with an ATG and have at least a further 99 contiguous sense codons[3], and were analysed using established procedures[4,5]. Before systematic sequencing, there were 73 genes[6], with 47 genes and their relative positions defining the genetic map, and an additional 26 genes located on the physical map. An additional 92 genes have formal genetic names, some of which had been previously cloned but not mapped to this chromosome or have been studied following their identification by systematic sequencing, for a total of 165 known genes. Thus only 33% of the total ORFs found on chromosome XVI had been identified before the completion of the sequence. Other genetic elements include: 17 tRNAs (9 of which are within 500base pairs of a long terminal repeat (LTR) element of a retrotransposon; 5 Ty retrotransposons; 15 delta elements including partial elements; 4 sigma and 2 tau elements; and 2 snRNAs.

The number of ORFs of known function is 194 (40%), of which 76 are functionally characterized proteins; 88 are known proteins that are not fully characterized; 26 have similarity to proteins of known biochemical and physiological function; 6 are homologous to proteins of known biochemical function; and 55 have a weak homology to known proteins. These weak similarities alone are insufficient to confidently assign function. This leaves 236 ORFs of unknown function (48%), of which 50 have homologues of unknown function, and 186 have no similarity to known proteins. There are 36 questionable ORFs, all of which partly overlap another ORF. All have a low codon adaptation index (CAI) of not greater than 0.18, are short (with an average length of 132 codons), and have no known homology with other proteins or are associated with no known phenotype. For four of these ORFs in two pairs (YPL034c and YPL035c, and YPR038w and YPR039w) it is unclear which, if any, are biologically meaningful. There are few apparent pseudogenes but these include YPL276w and YPL275w, which occur together in the genome as a frameshifted pair, both with homology to a formate dehydrogenase. This arrangement has been confirmed by direct sequencing of genomic DNA. Whether this region represents a mutation specific to strain S288C awaits experimental determination.

Functional categories have been compiled for chromosome XVI ORFs: 214 (43%) are classified in some form and 283 remain unclassified. The chromosome is sufficiently large to have a broad representation of all of the predominant functional groups. Detailed global genome classification and statistics for these functional assignments are tabulated and have been discussed elsewhere[7,8].

A robust amino-acid sequence motif is the presence of a predicted transmembrane domain, a region of a protein that spans a lipid bilayer. There are 181 ORFs with one or more predicted transmembrane domains on chromosome XVI (37%)[9], a number close to that seen on chromosomes II and III (refs 3, 10). Of these, 68 are known functionally in some form, and this percentage of known membrane proteins (38%) is similar to that of all known ORFs. Most of these membrane-protein encoding ORFs contain one or two predicted transmembrane domains (99 and 39, respectively); 15 proteins have 7 or more predicted transmembrane domains, and Ypl006p has 13.

An immediate way to use the sequence information is to examine experimentally ORF function, and the Canadian group is compiling systematic transcript and gene-disruption data. For a section encoding 89 ORFs in the region spanned by YPL085w to YPR017c, 61 transcripts were detected in haploid cells grown on rich medium at 30°C. This level of ORF expression is similar to that seen on chromosomes I and III (refs 5, 11). Of the 117 genes on chromosome XVI that have been disrupted, 36

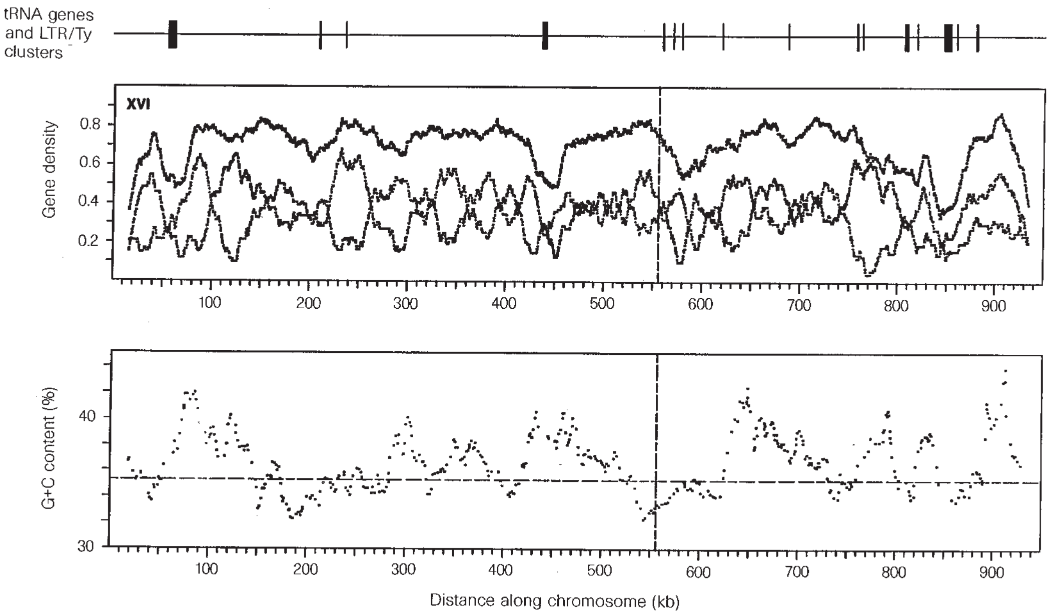

Figure 1 Molecular architecture of chromosome XVI. Top line shows positions of tRNA genes, solo LTR or Ty elements (thin vertical lines) or clusters of them (thick vertical lines) along the chromosome. Panels: variation of gene density (top) and base composition (bottom) along chromosome XVI (scale in kilobases from left telomere). Vertical broken lines indicate position of the centromere. Gene density determined as for chromosome XV is shown for the Watson (medium line) or the Crick (thin line) strands, and the sum of both (thick line). G+C content was calculated as for chromosome XV (horizontal broken line).

(31%) are 'essential' for vegetative growth. Although this fraction is higher than the approximately 12% estimated for the entire genome[12], or found on chromosome I (ref. 5), it still represents only 7% of the ORFs on the chromosome. The final proportion of essential genes on chromosome XVI awaits a complete disruption set. This information is displayed and will be updated (web sites URL http://www.mips.biochem.mpg.de and URL http://genome-www.stanford. edu).

Several ORFs deserve mention, some because of size, some because they were anticipated but not previously found, and others because they occur in metazoans with phenotypes or functions that appear characteristic of the larger multicellular eukaryotes, and whose unexpected presence in yeast affords some insight into function. There are 130 ORFs (27%) with significant similarity to products or predicted products of human genes. Four ORF products (*PAL1, PHO85, ROX1* and *RAD1*) are similar to products of the human genes *ALD, RET, SRY* and *XPF* (*ERCC4*) that, when altered by mutation, lead to X-linked adrenoleukodystrophy, multiple endocrine neoplasia 2A, gonadal dysgenesis, and xeroderma pigmentosum, respectively[13,14]. The largest ORF is YPR117w, specifying a 2,489-codon putative membrane protein of unknown function. Just one other ORF, *SEC18* (YPL085w), has more than 2,000 codons, and only 38 ORFs (8%) are larger than 1,000 codons. Identifying very small ORFs presents special problems[3,15] and just six identified ORFs on chromosome XVI are less than 100 codons. The smallest known functional ORF, *ATP15* (YPL271w), has 62 codons and encodes an F1-ATP synthase epsilon subunit[16]. Other small ORFs undoubtedly exist on the chromosome, and remain to be uncovered by functional analysis.

YPL127c encodes an apparent H1 histone, a protein previously not thought to be present in yeast. Plant and animal H1 histones are known to be involved in the higher-order assembly of nucleosomes but, despite considerable work, the precise role of H1 histone remains unclear[17,18]. Disruption of this single-copy yeast gene indicates that it is not essential for mitotic growth[32]. A detailed analysis of the possible role of this H1 histone in chromosome assembly, stability and the regulation of gene expression can now be explored.

YPR048w encodes a cytochrome P450 protein with similarity to human nitric oxide synthase, NOS. Human NOS is an amino-acid oxidoreductase that oxidizes the terminal N of arginine to give citrulline and nitric oxide. Nitric oxide (NO) in mammals behaves as a hormone, and is involved in many processes including vasodilation and neurological activities[19]. The involvement of NO in yeast metabolism had not previously been imagined. One possibility is that it acts to regulate cell signalling through interaction with small GTP-binding proteins[20].

We found that 74.57% of the chromosomal DNA is involved in the coding of ORFs. These 497 ORFs have an average size of 474 codons, close to the average values seen for other large chromosomes[3,4,15]. Of the 487 ORFs that are not Ty related, 251 are on the Watson strand and 236 are on the Crick strand, with no apparent strand bias[3]. There is, on average, an ORF every 1,908 bp over the chromosome. The positioning of ORFs relative to each other seems random: 111 ORFs are divergent, 118 are convergent, and 195 are tandemly arranged. Of the remaining ORFs, 38 are next to a non-ORF element, and in 12 cases two non-ORF elements are adjacent. With regard to intergenic spacing, the average length between tandemly arranged ORFs is 534 bp, for divergent ORFs it is 569 bp, and for convergent ORFs it is 340 bp. Only one region is apparently devoid of ORFs or other genetic entities for 3 kilobases or more: a subtelomeric 3,863 bp gap between YPL275w and YPL274w. There are no large non-coding regions comparable to those found in the subtelomeric regions of chromosomes I and VIII (refs 5,21).

There are no clustered gene families[5] on chromosome XVI and, as found with other yeast chromosomes, there is little apparent functional clustering of genes. There are two exceptions: a 'syntenic' pair, *CIT3* (YPR001w) and YPR002w, conserving the arrangement of the mmgD and mmgE genes in *Bacillus subtilis* (GenBank accession no. U29084); and two adjacent cyclin genes, *CLB2* (YPR119w) and *CLB5* (YPR120c), an arrangement that is duplicated on chromosome VII as *CLB1* (YGR108w) and *CLB6* (YGR109c), and which forms part of a larger duplication between these two chromosomes (see below).

The G+C periodicity on chromosome XVI varies about an average content of 38.1% (Fig. 1). Gene density also fluctuates along the chromosome, although it shows little correlation with G+C periodicity. Overall the G+C content of coding regions is 39.5%, and for non-coding regions is 33.29%. The centromere lies in a region of low G+C content, and the Ty and tRNA element clusters lie in regions of low gene density; such long-range compositional variations have been discussed elsewhere[22].

The centromere of chromosome XVI spans nucleotide residues 555,952–556,069, between *HAT1* and *CIT3*, making the chromosome close to being metacentric. Both telomeres seem to be typical in structure. There are 47 potential origins of DNA replication that match the 11-bp ARS element consensus, although the actual autonomously replicating sequences (ARS) used in replicating the chromosome remain to be determined experimentally.

Like much of the yeast genome, regions of chromosome XVI are duplicated. There are some large-scale DNA duplications spanning 25 kb or more[23,24]. The largest of these is on a 129-kb section of the right arm of chromosome XVI, nucleotide coordinates 731,001–860,000, where regions are duplicated onto a 129.5-kb section on the right arm of chromosome VII (nucleotide coordinates 648,001–777,500). Although removed from the comparison that identified this duplication, the region on chromosome XVI is rich in repetitive elements and contains three Ty elements, five additional LTRs and six tRNA genes. Such DNA duplications form large regions of partial gene synteny between these two chromosomes. An example is a section from nucleotide 834,000 to 860,000 on chromosome XVI and from 762,422 to 777,500 on chromosome VII. In the chromosome XVI interval from YPR154 to YPR159, four of the six genes, two delta elements and two tRNA genes, are syntenic with their chromosome VII counterparts, YGR136 toYGR143, with the exception of a tandem Ty1 element inserted between YPR158 and YPR159 on chromosome XVI. It has been suggested[25] that the origin of the *KRE6* (YPR159w)/*SKN1*(YGR143w) pair in this region resulted from some event involving duplication through transposition of the retrotransposon or tRNA elements. The origin of these major cluster homology regions remains unclear but probably consisted of more than one event; the subject is discussed elsewhere at a global genome level[7]. Evidence for duplications can also be found at the ORF level; 125 of the ORFs on chromosome XVI have one or more counterpart with significant similarity in the yeast genome. In addition to duplications, chromosome XVI also contains members of larger gene families, with at least 38 ORFs having similarity to two or more yeast ORFs. An example is *KTR6* (YPL053c), a member of a family of nine mannosyltransferase-encoding genes located on eight different chromosomes[26].

Establishment of the genetic and physical maps were critical precursors to obtaining the nucleotide sequence of the chromosomes. The genetic map of chromosome XVI is 251 cM in length, giving an average cM/kb value of 0.26, the smallest seen for a yeast chromosome, and close to that previously reported[27]. It indicates that a chromosome XVI bivalent has, on average, ten crossovers per meiosis. Although the genetic map is generally correct, there are several discrepancies with the positions found on the sequenced chromosome. Of the 46 genes mapped genetically through the phenotypes of alleles, 14 (*spoT16*, *rad53*, *nib1*, *sot1*, *SUF21*, *dna1*, *mak6*, *tsm0120*, *ymc1*, *spoT20*, *SUP15*, *SUP16*, *cdc67* and *rad56*) remain to be identified on the sequenced chromosome.

Determination of the nucleotide sequence of chromosome XVI completes the sequence of the yeast genome, allowing a genome-wide analysis of a small and experimentally amenable eukaryotic organism. Such systematic studies should enhance our knowledge of cell function, and help us to understand the structure and function of eukaryotes with larger genomes. □

Methods

Chromosome XVI was sequenced using a set of overlapping cosmid and lambda clones based on a previous chromosome XVI physical map[28] (L. Riles and M.V. Olson, personal communication) from the S288C-derived strain AB972. Two gaps on the right arm and both telomeres were sequenced using polymerase chain reaction (PCR) products amplified from genomic DNA[29]. Individual cosmids representing the portion mapped by the EU group were sequenced by contracting laboratories using a variety of subcloning, sequencing and assembly methods[3, 4]. Explanations of the cloning, sequencing, assembly and quality control methods used by the other groups have been described[21,30]. The telomeres were sequenced using specially devised procedures[31]. Assembly of the completed chromosome sequence was made in Martinsried or Montreal as described[3–5,15]. Determination of overlapping sequence between groups indicated that seven differences were found and resolved in 81 kb of sequence . This allowed us to estimate the accuracy of the sequence to be conservatively within the three errors per 10 kb average for the yeast genome[7].

Received 29 July 1997; accepted 11 March 1997.

1. Hawthorne, D.C. & Mortimer, R.K. *Genetics* 60, 735-742 (1968).
2. Broach, J.R. in *The Molecular Biology of the Yeast* Saccharomyces, *Life Cycle and Inheritance* 653–727 (Cold Spring Harbor Laboratory Press, NY, 1981).
3. Feldmann,H. *et al. EMBO J.* 13, 5795–5809 (1994).
4. Galibert, F. *et al. EMBO J.* 15, 2031–2049 (1996).
5. Bussey, H. *et al. Proc. Natl Acad. Sci. USA* 92, 3809–3813 (1995).
6. Mortimer, R.K, *et al*. http://genome.www.stanford.edu/saccdb/edition12.html (1995).
7. Goffeau, A. *et al. Science* 274, 546 (1996).
8. Goffeau, A. *et al. Science* 274, 563–567 (1996).
9. Klein, P., Kanehisa, M. & Delesi, C. *Biochem. Biophys. Acta* 815, 468–476 (1985).
10. Goffeau, A., Nakai, K., Slonimski, P.P. & Risler, J.L. *FEBS Lett.* 325, 112–117 (1993).
11. Tanaka, S. & Isono K. *Nucleic Acids Res*. 21, 1149–1153 (1992).
12. Goebl, M.E. & Petes, T.D. *Cell* 46, 983–992 (1986).
13. Bassett, D.E., Boguski, M. & Hieter, P. *Nature* 379, 589–590 (1996).
14. van Vuuren, A.J. *et al. EMBO J.* 12, 3693– 3701 (1993).
15. Dujon, B. *et al. Nature* 369, 371–378 (1994).
16. Guelin, E. *et al. J. Biol. Chem*. 268, 161–167 (1993).
17. Zlatanova, J. & Doenecke D. *FASEB J.* 8, 1260–1268, (1994).
18. Sirotkin, A.M.*et al. Proc. Natl Acad. Sci. USA* 92, 6434–6438 (1995).
19. Forstermann, U. *et al. Hypertension* 23, 1121–1131(1994).
20. Lander, H.M. *et al. Nature* 381, 380–381 (1996).
21. Johnston, M. *et al*. *Science* 265, 2077–2082 (1994).
22. Dujon, B. *et al. Trends Genet*. 12, 263–270 (1996).
23. Heumann, K. & Mewes, H.W. *Nature Genet*. (submitted).
24. Heumann, K., Harris, C. & Mewes, H.W. in *Proc. Fourth Int. Conf. Intelligent Systems for Mol. Biol.* (St Louis, MO, 1996).
25. Roemer, T., Fortin, N. & Bussey, H. *Yeast* 10, 1527–1530 (1994).
26. Lussier, M., *et al. Yeast* 13, 267–274 (1997).
27. Mortimer, R.K., Contopoulou, C.R. & King, J.S. *Yeast* 8, 817–902 (1992).
28. Riles, L. *et al. Genetics* 134, 81–150 (1993).
29. Louis, E.J. & Borts, R.H. *Genetics* 139, 125–136 (1995).
30. Wilson, R. *et al. Nature* 368, 32–38 (1994).
31. Louis, E.J. *Biochemica* 3, 25–26 (1995).
32. Ushinsky, S. C. *et al. Yeast* 13, 151–161 (1997).

Acknowledgements. We thank D. Oki for administrative assistance and manuscript preparation; C. Jallet, D. Jauniaux, M. Vanden Barre and A. Wong for technical help; M. Lussier, T. Ketela and B. Baum for input; and B. Dujon for preparation of the G+C periodicity figure, using unpublished software developed in collaboration with C. Marck. This work was supported by the Canadian Genome Analysis and Technology Program, European Yeast Genome Sequencing Project, Human Genome Organization, The Wellcome Trust, Région Wallone, and the Region of Brussels Capital.

Correspondence and requests for materials should be addressed to H.B.
(e-mail: hbussey@ monod.biol.mcgill.ca)