requires manual editing to find all details of cluster duplications, such as multigene families, potentially inverted ORF members, more than averagely diverged ORFs, and tRNA genes.

The 17-kb subtelomeric cluster duplication between chromosomes XIV and VI (cluster duplications 14–6) consists entirely of highly conserved ORF pairs (average 96.6% amino-acid identity) and shows stringent synteny. The intergenic regions are also highly conserved, suggesting that the duplication of the six ORFs is a relatively recent event on an evolutionary timescale.

Most of the ORF pairs in the other six cluster duplications are much less conserved, and their promotor and terminator regions lack significant homologies, suggesting that they are ancient duplication events. Five of the highly conserved ORF pairs of these ancient duplications code for ribosomal proteins (average 95.3% amino-acid identity), one for two members of the 70K heat-shock protein family (99.3% amino-acid identity), one for two forms of iso-propyl malate synthase (88.5% amino-acid identity) and one for two forms of citrate synthase (81.4% amino-acid identity) (Table 1). Excluding these ORF pairs, which are apparently under high selection pressure to preserve their sequence information, the average homology of ORF pairs was determined for each of the cluster duplications. ORF pairs in cluster duplications CD14–15B and CD 14–3 (average 56% amino-acid identity) seem to be less diverged than ORF pairs in CD14–15A, CD14–8, CD14–9 (average 47.5% amino-acid identity) and CD14–4 (average 37% amino acid identity). However, there are too few ORF pairs to draw conclusions about different temporal orders for the cluster duplications involving chromosome XIV.

Could the six ancient cluster duplications, at the time of their creation, have looked similar to the recent cluster duplications between chromosomes XIV and VI, with perfect synteny of all ORFs? And could they have been shaped over evolutionary time by base-pair changes, insertions of new ORFs, deletions of some of the originally duplicated ORFs, inversions of single or groups of ORFs, and translocations to yield the present picture of 'relaxed synteny'? This is certainly possible if the now visible arrangements indeed originated from duplications of gene clusters, perhaps by long-range gene conversions or chromosome duplications. However, it remains possible that the evolutionary history of *S. cerevisiae* involved fusion of two ancient forms of yeast cells with smaller genomes already displaying sequence divergencies and some level of relaxed synteny and that, for most of the duplicated ORFs, one copy was lost over time because of a lack of selective advantage for *S. cerevisiae* to keep more than one copy. □

Received 22 July 1996; Accepted 11 March 1997.

1. Mortimer, R. K., Contopoulou, C.R. & King, J.S. *Yeast* 8, 817–902 (1992).
2. http://www.mips.biochem.mpg.de/mips/yeast/
3. Verhasselt, P., Aert, R., Voet, M. & Volckaert, G. *Yeast* 10, 945–951 (1994).
4. Verhasselt, P., Aert, R., Voet, M. & Volckaert, G. *Yeast* 10, 1355–1361 (1994).
5. Jonniaux, J. L., Coster, F., Purnelle, B. & Goffeau, A. *Yeast* 10, 1639–1645 (1994).
6. Kick, C. T., Maurer, J. H., Maurer, U. & Planta, R.J. *Yeast* 11, 1303–1310 (1995).
7. Mallet, L., Bussereau, F. & Jacquet, M. *Yeast* 11, 1195–1209 (1995).
8. Van Dyck, L., Pascual-Ahuir, A., Purnelle, B. & Goffeau, A. *Yeast* 11, 987–991 (1995).
9. Coster, F., van Dyck, L., Jonniaux, J.-L., Purnelle, B. & Goffeau, A. *Yeast* 11, 85–91 (1995).
10. Bergez, P., Doignon, F. & Crouzet, M. *Yeast* 11, 967–974 (1995).
11. Maftahi, M., Nicaud, J.-M., Levesque, H. & Gaillardin, C. *Yeast* 11, 567–572 (1995).
12. Maftahi, M., Nicaud, J.-M., Levesque, H. & Gaillardin, C. *Yeast* 11, 1077–1085 (1995).
13. Maurer, K. C., Urbanus, J. H. & Planta, R. J. *Yeast* 11, 1303–1310 (1995).
14. Soler-Mira, A., Saiz, J. E., Ballesta, J. P. G. & Remacha, M. *Yeast* 12, 485–491 (1996).
15. Levesque, H., Lepingle, A., Nicaud, J.-M. & Gaillardin, C. *Yeast* 12, 289–295 (1996).
16. Sen-Gupta, M., Lyck, R., Fleig, U., Niedenthal, R. K. & Hegemann, J. H. *Yeast* 12, 505–514 (1996).
17. Nasr, F., Bécam, A.-M. & Herbert, C.J. *Yeast* 12, 169–175 (1996).
18. Nasr, F., Bécam, A.-M. & Herbert, C.J. *Yeast* 12, 493–499 (1996).
19. Pöhlmann, R. & Philippsen, P. *Yeast* 12, 391-402 (1996).
20. Saiz, J. E., Buitrago, M. J., Soler-Mira, A., Del Rey, F. & Revuelta, J.L. *Yeast* 12, 403–409 (1996).
21. Garcia-Cantalejo, J. M., Boskovic, J. & Jimenez, A. *Yeast* 12, 599–608 (1996).
22. Pandolfo, D., De Antoni, A., Lanfranchi, G. & Valle, G. *Yeast* 12, 1071–1076 (1996).
23. Kalogeropoulos, A. *Yeast* 11, 555–565 (1995).
24. Logghe, M., Molemans, F., Fiers, W. & Contreras, R. *Yeast* 10, 1093–1100 (1994).
25. Rodriguez-Medina, J. R. & Raymond, B. C. *Mol. Gen. Genet.* 243, 532–539 (1994).
26. Garrels, J. I. *Nucleic Acids Res.* 24, 46–49 (1996).
27. Tugenreich, S., Boguski, M. S., Seldni, M. S. & Hieter, P. *Proc. Natl Acad. Sci. USA* 90, 10031–10035 (1993).
28. Tugenreich, S., Bassett, D. E., McKusick, V. A., Boguski, M. S. & Hieter, P. *Hum. Mol. Genet.* 3, 1509–1517 (1994).
29. Pöhlmann, R. & Philippsen, P. *Yeast* 11, 634 (1995).
30. Steensma, H.Y., de Jonge, P., Kaptein, A. & Kaback, D.B. *Curr. Genet.* 16, 131–137 (1989).
31. Lalo, D., Stettler, S., Mariotte, S., Slonimski, P. P. & Thuriaux, P. *C.R. Acad. Sci. Paris* 316, 367–373 (1993).
32. Johnston, M., *et al. Science* 265, 2077–2082 (1994).
33. Wolfe, K. H. & Lohan, A. J. *Yeast* 10, 41–46 (1994).
34. Melnick, L. & Sherman, F. *J. Mol. Biol.* 233, 372–388 (1993).
35. http://speedy.mips.biochem.mpg.de/programs/GENOME_BROWSER.html
36. Friedman, K. L. *et al. Genes Dev.* 10, 1595–1607 (1996).
37. Thierry, A., Gaillon, L., Galibert, F. & Dujon, B. *Yeast* 11, 121–135 (1995).
38. Stucka, R. & Feldmann, H. in *Molecular Genetics of Yeast* (ed. Johnston, J. R.) 49–64 (IRL Oxford, 1994).
39. Hamberg, K. *PhD-Thesis, Univ. Giessen* (1993).
40. Vollrath, D., Davis, W.D., Cornelly, C. & Hieter, P. *Proc. Natl. Acad. Sci. USA* 85, 6027–6031 (1988).
41. Riles, L. *et al. Genetics* 134, 81–150 (1993).
42. Louis, E.J. & Borts, R.H. *Genetics* 139, 125–136 (1995).
43. Chan, C.S.M. & Tye, B.-K. *Cell* 33, 563–573 (1983).
44. Louis, E.J. *Yeast* 11, 1553–1573 (1995).
45. Stoler, S., Keith, K.C., Curnick, K.E. & Fitzgerald-Hayes, M. *Genes Dev.* 9, 573–586 (1995).

Acknowledgements. We thank L. Riles, A. Thierry, B. Dujon, R. Stucka, H. Feldmann, E. Louis, K. Friedman and B. Brewer for clones and cosmid libraries; R. Spiegelberg, A. Thierry and D. Fischer for helping to isolate or characterize DNA clones; M. Johnston, T. Donahue, N. Pfanner, B. Winsor and D. Gallwitz for suggestions; and R. Niederhauser for secretarial help. The majority of funding was provided by the Biotech Programs of the European Commission. Additional financial support was contributed by the following national agencies: Groupement de Recherches et d'Etudes sur les Génomes du Ministre de la Recherche, France; Région de Bruxelles-Capital, Belgium; Belgian Federal Services for Science Policy (D.W.T.C.) ; Research Fund of the Katholieke Universiteit Leuven, Belgium; Services Fédéraux des Affaires Scientifiques, Techniques et Culturelles; Pôles d'attraction Inter-universitaire and Région Wallone, Belgium; Fundacion Ramon Areces and Comision Interministerial de Ciencia y Tecnologia, Spain. The participation of scientists from Switzerland was made possible by a grant from the Swiss Federal Agency for Education and Science.

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV

B. Dujon[1], K. Albermann[2], M. Aldea[3], D. Alexandraki[4,5], W. Ansorge[6], J. Arino[7], V. Benes[6], C. Bohn[8], M. Bolotin-Fukuhara[8], R. Bordonné[9], J. Boyer[1], A. Camasses[9], A. Casamayor[7], C. Casas[3], G. Chéret[10], C. Czepluch[11], B. Daignan-Fornier[8], D. V. Dang[8], M. de Haan[12], H. Delius[13], P. Durand[14], C. Fairhead[1], H. Feldmann[15], L. Gaillon[1], F. Galisson[1], F.-J. Gamo[16], C. Gancedo[16], A. Goffeau[17], S. E. Goulding[18], L. A. Grivell[12], B. Habbig[19], N. J. Hand[18], J. Hani[2], U. Hattenhorst[19], U. Hebling[13], Y. Hernando[20], E. Herrero[3], K. Heumann[2], R. Hiesel[21], F. Hilger[14], B. Hofmann[13], C. P. Hollenberg[19], B. Hughes[22], J.-C. Jauniaux[11], A. Kalogeropoulos[8], C. Katsoulou[4], E. Kordes[11], M. J. Lafuente[16], O. Landt[23], E. J. Louis[24], A. C. Maarse[12], A. Madania[9], G. Mannhaupt[15], C. Marck[25], R. P. Martin[9], H. W. Mewes[2], G. Michaux[1], V. Paces[26], A. G. Parle-McDermott[18], B. M. Pearson[20], A. Perrin[1], B. Pettersson[27], O. Poch[9], T. M. Pohl[22], R. Poirey[11], D. Portetelle[14], A. Pujol[11], B. Purnelle[17], M. Ramezani Rad[19], S. Rechmann[6], C. Schwager[6], M. Schweizer[20], F. Sor[10], F. Sterky[27], I. A. Tarassov[9], C. Teodoru[6], H. Tettelin[1,17], A. Thierry[1], E. Tobiasch[11], M. Tzermia[4], M. Uhlen[27], M. Unseld[21], M. Valens[8], M. Vandenbol[14], I. Vetter[15], C. Vlcek[26], M. Voet[28], G. Volckaert[28], H. Voss[6], R. Wambutt[29], H. Wedler[29], S. Wiemann[6], B. Winsor[9], K. H. Wolfe[18], A. Zollner[2], E. Zumstein[20] & K. Kleine[2]

[1]Unité de Génétique Moléculaire des Levures (URA 1149 CNRS and UFR 927 Univ. P.M. Curie), Institut Pasteur, 25 Rue du Dr. Roux, F75724, Paris Cedex 15, France
[2]Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152, Martinsried, Germany
[3]Department of Basic Medical Sciences, Faculty of Medicine, University of Lleida, E-25006, Lleida, Spain
[4]Foundation for Research and Technology-Hellas, IMBB, P.O. Box 1527, Heraklion 711 10 Crete, Greece
[5]Department of Biology, University of Crete, Heraklion 711 10 Crete, Greece
[6]Biochemical Instrumentation Program, EMBL, Meyerhofstrasse 1, D-69117, Heidelberg, Germany
[7]Departamento de Bioquimica y Biologia Molecular, Universidad Autonoma de Barcelona, Bellaterra, E-08193, Spain
[8]Institut de Génétique et Microbiologie, Bâtiment 400, Université Paris-Sud,

F-91405 Orsay Cedex, France
[9]UPR9005 (MMDCD) du CNRS, 15 rue René Descartes, F-67084, Strasbourg, France
[10]Institut Curie, Bâtiment 110, Centre Universitaire, F-91405, Orsay Cedex, France
[11]Tumorvirologie Abteilung 0610 and Virologie appliquée à l'oncologie (INSERM U375), Deutsches Krebsforschungszentrum, P.101949, D-69009, Heidelberg, Germany
[12]Section for Molecular biology, Department of Molecular Cell Biology, University of Amsterdam, Kruislaan 318, NL-1098 SM Amsterdam, The Netherlands
[13]Abteilung für Angewandte Tumorvirologie 0686, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany
[14]Faculté Universitaire des Sciences Agronomiques, Unité de Microbiologie, 6 Avenue Maréchal Juin, B-5030 Gembloux, Belgium
[15]Institut für Physiologische Chemie, Physikalische Biochemie und Zellbiologie der Universität München, Goethestrasse 33, D-80336 München, Germany
[16]Instituto de Investigaciones Biomedicas, CSIC, Arturo Duperier 4, E-28029 Madrid, Spain
[17]Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2-20, B-1348 Louvain-La-Neuve, Belgium
[18]Department of Genetics, Univeristy of Dublin, Trinity College, Dublin 2, Ireland
[19]Institut für Mikrobiologie der Heinrich-Heine-Universität Düsseldorf, Geb. 26.12, Universitätstrasse 1 D-40225 Düsseldorf, Germany
[20]Institute of Food Research, Genetics and Microbiology Department, Norwich Research Park, Colney, Norwich NR4 7UA, UK
[21]Institut für Genbiologische Forschung, Ihnestrasse 63, D-14195 Berlin, Germany
[22]GATC-Gesellschaft für Analyse-Technik und Consulting mbH, Fritz-Arnold-Strasse 23, Konstanz, D-78467, Germany
[23]TIB MolBiol, Tempelhofer Weg 11-12, D-10829 Berlin, Germany
[24]Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK
[25]Service de Biochimie et de Génétique moléculaire, Département de Biologie Cellulaire et Moléculaire, DSV/CEA-Saclay, F-91191 Gif-sur-Yvette, France
[26]Institute of Molecular Genetics, Academy of Sciences, Prague, Czech Republic
[27]Department of Biochemistry and Biotechnology, Royal Institute of Technology (KTH), S10044, Stockholm, Sweden
[28]Laboratory of Gene Technology, Katholieke Universiteit Leuven, Willem de Croylaan, 42, B-3001, Leuven, Belgium
[29]AGON GmbH, Glienicker Weg 185, D-12489 Berlin, Germany

Chromosome XV was one of the last two chromosomes of *Saccharomyces cerevisiae* to be discovered[1]. It is the third-largest yeast chromosome after chromosomes XII and IV, and is very similar in size to chromosome VII. It alone represents 9% of the yeast genome (8% if ribosomal DNA is included). When systematic sequencing of chromosome XV was started, 93 genes or markers were identified, and most of them were mapped[2]. However, very little else was known about chromosome XV which, in contrast to shorter chromosomes, had not been the object of comprehensive genetic or molecular analysis. It was therefore decided to start sequencing chromosome XV only in the third phase of the European Yeast Genome Sequencing Programme, after experience was gained on chromosomes III, XI and II (refs 3–5). The sequence of chromosome XV has been determined from a set of partly overlapping cosmid clones derived from a unique yeast strain, and physically mapped at 3.3-kilobase resolution before sequencing. As well as numerous new open reading frames (ORFs) and genes encoding tRNA or small RNA molecules, the sequence of 1,091,283 base pairs confirms the high proportion of orphan genes and reveals a number of ancestral and successive duplications with other yeast chromosomes.

The DNA sequence of 1,091,283 nucleotides contains 560 ORFs, of at least 100 sense codons, that are not entirely included within a larger one (our standard basic definition; see ref. 4). If those corresponding to Ty or Y′ elements are excluded, and intron predictions are considered (see below), 551 different ORFs remain. To these were added eight known genes shorter than 100 codons (*BAT2, CRS5, RPB10, RPS30B, RPS33A, SME1, TOM6* and the *CPA1* leader), and a pseudogene 581 codons long (YOL153c) that contains two in-frame ochre codons. It is considered here because its putative translation product has significant homology with the Gly-X carboxypeptidase encoded by *CPS1*, and because, in another yeast strain, the two stop codons are replaced by two glutamine codons CAA[6]. Note that, in the present sequence, YOR031w (*CRS5*) also contains an in-frame ochre codon instead of the CAA codon found in other strains (the reality of the stop codon was verified by direct sequencing on yeast DNA). Other interesting but more complex cases of pseudogenes found in chromosome XV will be described elsewhere. Also note that YOL040c has been considered here (coordinates 253,147– 253,575) instead of the larger antisense ORF within which it is entirely included because it corresponds to a known gene (*RPS21*).

The sequence also reveals 526 short ORFs (from 50 to 99 sense codons) not entirely included within larger ones, six of which are already considered above. In this size range, it is obviously difficult to distinguish actual genes from random occurrences. Using high codon bias (CAI > 0.2) and absence of a partial overlap with larger ORFs as predictive criteria, only six candidates remained. One of these shows very significant homology with the 60 codon-long gene *HOR7* of chromosome XIII. This ORF, named YOL052ca, has been added to the above list, bringing the grand total to 561. Of these ORFs 33 (5.9%) are 'questionable', based on their short size and low CAI (see ref. 4), and 18 of them partly overlap other ORFs, increasing their suspicious character.

The longest ORF of chromosome XV is YOL081w (the *IRA2* gene) with 3,079 codons. The two shortest ORFs, if the 25 amino-acid leader peptide of the *CPA1* gene is ignored, are YOR045w (*TOM6*) and YOL052ca, with 61 codons each. The average size of chromosome XV ORFs is 457 codons, very close to the figure observed for the entire yeast genome[7]. In total, 33 pairs and 5 trios of partly overlapping ORFs are found; 25 pairs are antiparallel, excluding the possibility of sequencing errors. Among the parallel pairs, YOR012w and YOR013w are suggestive of a frameshift error, or a pseudogene, as their products share homology with the amino terminus and the carboxy terminus of the product of YDR391c, respectively.

The overall density for protein-coding genes is 70.6%, slightly lower than the average for the whole yeast genome subtracted from rDNA[7]. Variations are observed along the chromosome, with two short regions showing gene density above 85% (centred around ~200 kb and ~950 kb, respectively; see Fig. 1). These regions also correspond to areas of high G+C content. As is generally observed in yeast, the two subtelomeric regions show low gene density.

The entire chromosome shows no significant strand coding bias, but important local variations are observed, with seven short regions (of ~30–50 kb each) showing a clearcut excess of ORFs on the Watson strand; eight others have an excess on the Crick strand (Fig. 1). Orientation of neighboring ORFs is random for the whole chromosome, with 150 diverging pairs, 149 converging pairs, and 261 tandemly arranged pairs (123 on the Watson strand, and 138 on the Crick strand). The longest tandem array contains 11 successive ORFs (YOR104w to YOR114w, coordinates 517,639–538,451).

A total of 13 introns have been identified, most of which occur in short ORFs with high codon bias. Chromosome XV introns are short, as is typical for yeast, ranging in size from 135 to 527 nucleotides (average 334 nucleotides). They have an average G+C content of 34.9%, significantly lower than that of the entire chromosome (38.2%) or its ORFs, (39.5%). Note that the possible occurence of introns in the 5′ untranslated region of the pre-mRNA molecules has not been examined systematically for lack of discriminative criteria among the numerous occurrences of the intron consensus 5′-GYMHGH-N1-TACTAAC-N2-YAG-3′ in the sequence (294 occurences if N1 and N2 are set shorter than or equal to 400 and 50 nucleotides, respectively, or more than 600 for limits of 1,000 and 180 nucleotides).

The 20 tRNA genes recognized correspond to 12 different amino acids and 22 different codons. Six tRNA genes contain introns. All tRNA genes are significantly richer in G+C content (47–63%) than the average yeast genome sequence. The frequent duplication of tRNA genes in yeast is

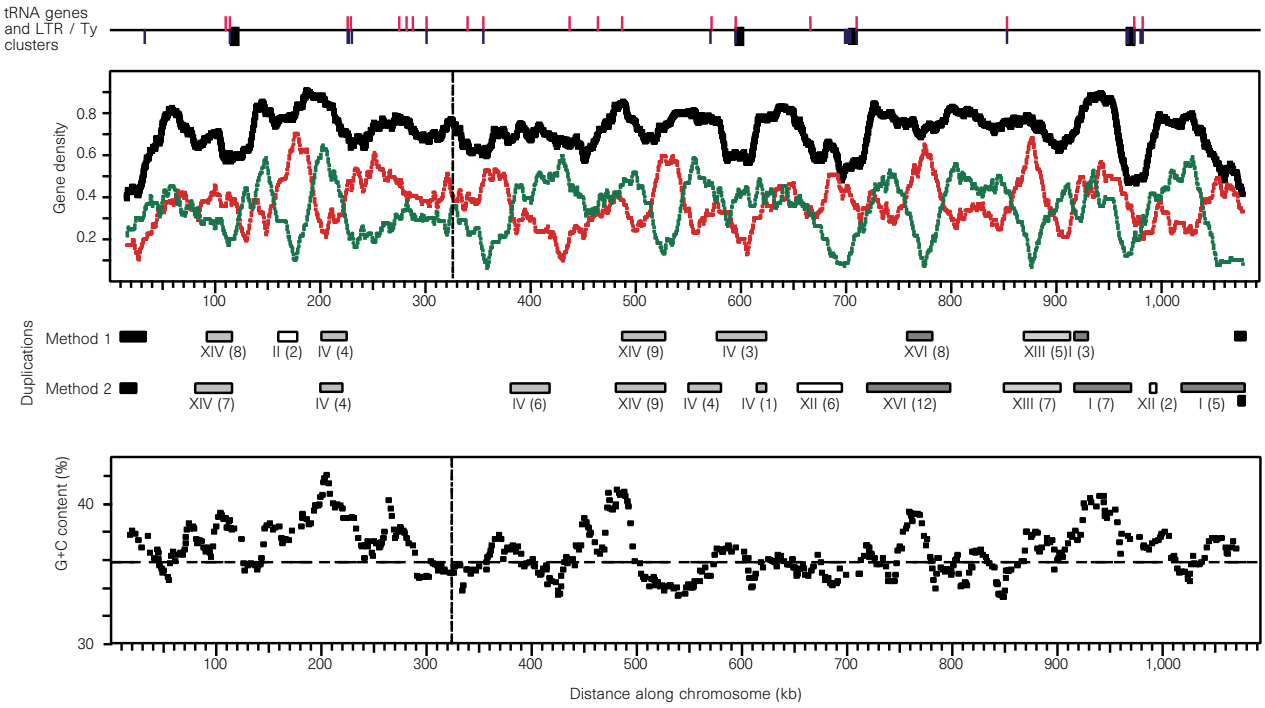

Figure 1 Variation in gene density (top) and base composition (bottom) along the sequence of chromosome XV (scale in kilobases from left telomere). Vertical broken lines indicate position of the centromere. Gene density is expressed as the probability that each nucleotide is part of an ORF, calculated using sliding windows of 30 kb (steps of 0.5 kb) for the Watson strand alone (red line), the Crick strand alone (green line), and their sum (black line). The Watson strand is oriented 5′ to 3′ from left to right on the chromosome map[4]. G+C content was calculated from silent codon positions using a sliding window of 13 consecutive ORFs (horizontal broken line indicates average G+C content at these positions; 35.7%). Top line, positions of tRNA genes (thin bars above line), solo LTR (thin bars below line), and complete Ty elements (thick bars below line). Middle, positions of major 'ancient' chromosomal duplications. Blocks represent extent of clusters (method 1), or arrays of ORF pairs (method 2), as defined in text, with indication of the matching chromosome and the number of ORFs involved (in parentheses). Note that several blocks corresponding to a given chromosome are often intermingled with those corresponding to another. Blocks in subtelomeric position (filled) match with several distinct chromosomes.

noticeable on chromosome XV alone, with four gly-tRNA genes (*tG(GGY)OL1* and *tG(GGY)OL2* are identical in sequences, and *tG(GGG)OR1* is similar to the previous two), three pro-tRNA genes (identical in sequences except for their introns), two thr-tRNA genes, and two asn-tRNA genes, respectively identical in sequences; there are also two met-tRNA genes, but they differ in sequence. Duplicated tRNA genes are always found within different sequence environments (including different associated long terminal repeats (LTRs); see below). The tRNA genes are distributed throughout the chromosome (Fig. 1), as is generally the case for yeast which, unlike other organisms, does not show large clusters of tRNA genes in its genome.

Seven genes encoding small known RNA molecules were recognized from the sequence. One of these, *snR17A*, which encodes the U3 snRNA, contains an intron and is duplicated elsewhere in the yeast genome[8]. The downstream part of the gene encoding snR35 partly overlaps the downstream part of an ORF of unknown function, YOR222w, the product of which shares similarity with ADP-ATP carrier proteins.

Four complete Ty elements (two Ty1 and two Ty2), 20 solo LTRs or remnants of them (12 delta, 6 sigma and 2 tau elements), and one Y′ element were found. Solo LTRs, or complete Ty elements, are almost always located immediately upstream of tRNA genes (Fig. 1). Only one solo LTR element and one Ty1 stand alone. Consistent with general trends[9], the two Ty2 elements are found at 'old' sites occupied by several solo LTRs associated with tRNA genes; the same is true for only one of the two Ty1 elements. Conversely, only 11 of the 20 tRNA genes have one or more LTR element upstream of their 5′ end, the closest element always being within 200 base pairs of the tRNA gene (sigma elements are only 16–18 nucleotides upstream of tRNA genes).

In total, RNA-coding genes and transposons occupy only 1% of the chromosome XV sequence.

There remain 309,627 base pairs (28.4% of the chromosome) that we will call here 'intergenic regions'. Such regions contain, for a small part, structural chromosomal elements such as the centromere (coordinates 326,592–326,706), the telomeric $(C_{1-3}A)_n$ repeats (coordinates 1–113 and 1,091,264–1,091,283) and two subtelomeric core X elements and their associated repeated elements (STR) (coordinates 114–847 and 1,083,914–1,084,611). But intergenic regions primarily contain promoters, terminators and transcriptional regulatory elements of the protein-coding genes, most of which have not yet been identified. Intergenic regions believed to contain promoter elements based on the orientation of flanking ORFs are noticeably longer (791 nucleotides on average between diverging ORFs) and are richer in G+C (36.2%) than intergenic regions containing putative terminators (421 nucleotides and 28.7% G+C). The presence of flanking RNA-coding genes does not alter this trend. The sequence also reveals 63 ARS consensus elements (5′-WTTTAYRTTTW-3′) the activity of which remains to be examined; 39 of these occur in intergenic regions.

Chromosome XV contains few simple sequence iterations. The longest dinucleotide repeat is an alternating poly(AT) stretch of 20-mers (coordinates 45,691–45,730) within the intergenic region that separates the converging ORFs YOL149w (*DCP1*) and YOL148c (*SPT20*). Only 19 other cases exist of either dinucleotide repeats of at least 10-mers (all are alternating poly(AT) or mononucleotide repeats of at least 20-mers (all poly(A) or poly(T)). Similarly, few trinucleotide repeats are found, the longest being a 20-mer of the triplet CAA occurring within an ORF of unknown function, YOR267c, and encoding a poly-glutamine stretch. Six

Table 1 Assembly of the chromosome XV sequence from individual submissions

Cosmids or DNA	Coordinates on final chromosome	Overlap	Strategy*	Reference
telomeric plasmid / pEOA363	1–32687		S, A	
pEOA179 / pEOA461	24486–97824	8202	N, A, M	6, 14, 16, 17, 18, 20
pEOA417	96924–140942	901	S, M	26
pEOA228	139031–178337	1912	S, A	28,29
pEOA1044	177014–210234	1324	N, W, A	in the press
pUOA1217	209185–235991	1050	SS, M	22
pUOA1344	222408–256233	13584	S, W, A	
pEOA321	253576–287613	2658	N, A	in the press
pEOA215	286637–323078	977	S, R	
pEOA156	321732–352202	1347	S, P, A	30
pEOA303 / pEOA270	350740–408356	1463	S, A	
pEOA272	391560–427841	16799	S, A	
pEOA213 / pEOA217	415169–477887	12673	SS, M	in the press
11 cosmids†	476475–606002	1413	S, W, A	27 and in the press
pEOA477 / pUOA1258	604545–660867	1458	SS, W, M	15, 21
pUOA533 / pEOA378 / pEOA241	655864–741096	5004	S, R	
pEOA423 / pEOA048	739215–799188	1882	S, A	8, 11
pUOA1302	795488–832262	3701	SS, W, M	in the press
pUOA1337	823760–861431	8503	S, A	19
pEOA487	859329–895527	2103	SS, W, M	in the press
pEOA284	892246–927955	3282	SS, A	24
pUOA502 / pEOA232	927738–957182	218	S, A	23
pEOA138	955761–996055	1422	S, M	25
5 cosmids‡	992145–1081258	3911	SS, W, M	
Right PCR	1080794–1091283	465	P	

*S, shotgun of cosmid or part thereof (SS); N, nested deletions; W, walking primer; P, PCR fragments; M, manual gels; A, automatic fluorescent; R, direct membrane blotting

†pEOA347, pUOA522, pEOA246, pEOA264, pEOA273, pEOA306, pEOA265, pEOA106, pEOA338, pEOA986 and pEOA1081.

‡pEOA387, pEOA360, pEOA343, pEOA434 and pEOA390.

other cases of trinucleotide repeats of at least 10-mers are found. Polymorphic variations in trinucleotide repeats have been described in yeast[10]. An example of such variations is given by two ORFs of unknown function, YOR229w and YOR230w, that represent an ancient and diverged tandem duplication (67% sequence identity), with an insertion of a long imperfect trinucleotide repeat in YOR229w that is absent from YOR230w[11].

Iterations of a few longer sequence motifs are also present. The clearest example is probably the near-perfect repeat of the 39 nucleotide-long unique sequence 5′-GAGCCTGATTCTGTGGCAGAAGATGAACCGGA-GACTGAT-3′, which occurs nine times between positions 30,935 and 31,309, at the beginning of YOL155c, an ORF of unknown function shows similarity to glucan-1,4-α-glucosidase. The repeats determine a serine-rich amino-acid sequence. At the end of the same ORF (coordinates 28,905–29,279), another near perfect repeat of a unique long sequence, 5′-CAGTAGTGTATGWYTTNGGRGAARCASTRGTTKCKG-3′, occurs four times. In both cases, degenerated copies of the unique sequence are also found beyond the main repeats.

Of the 561 identified ORFs, 212 (37.8%) correspond to known and functionally characterized genes; all of the others are new. By sequence comparison of their products with general databases, 34 of these (6.1% of the total) show significant homology to proteins of known biochemical and/or physiological function of either yeast or other organisms, and 69 others (12.3% of the total) show weak homology. There are 246 ORFs with products that have either no significant homologue (187 cases or 33.3% of total, among which 28 are questionable) or are homologous to proteins that are themselves of unknown function (59 cases or 10.5% of the total).

Using previously defined criteria[12], 239 ORF products (42.6% of total) are predicted to contain at least one transmembrane span, 170 of which are of unknown function. Two proteins have 12 predicted spans (SCM2, the product of YOL020w, and ALG8, the product of YOR067c), and 43 others have five predicted transmembrane spans or more (only 15 of them are functionally characterized).

Duplications are frequent in the yeast genome, and take several different forms that suggest distinct mechanisms of formation. Comparison of the chromosome XV sequence with the entire yeast genome (including chromosome XV), reveals 12 'clusters' of duplicated sequences that may represent ancestral chromosomal duplications subsequentely modified. Such clusters range in size from 12 kb to 49.5 kb, and contain two to nine ORFs, making a total of 297.5 kb of the chromosome XV sequence (27% of total) that is duplicated on at least eight other chromosomes. Duplications with chromosomes IV and XIV are each represented by two clusters intermingled along the chromosome XV map, suggesting successive events of chromosome duplication or rearrangements. Among the four clusters that contain tRNA genes, there are two cases where homologous tRNA genes are conserved at equivalent positions on the other chromosome, further supporting, in such cases, the hypothesis of ancestral chromosomal duplications.

For chromosome XV, duplications were also examined using a second approach based on the systematic comparison of predicted translation products of all yeast ORFs against all others, followed by sequence alignments and estimation of their significance compared to randomizations, and to the overall distribution of similarity values for the entire yeast genome. A total of 193 different ORFs of chromosome XV (34.4%) were found to have at least one significant homologue in the yeast genome (including 28 pairs on the chromosome XV itself). Half of these have several homologues, forming gene families with various degrees of divergence. Among these, 35 ORFs, nearly all in a subtelomeric location, are members of large families with five partners or more (one of them has up to 26 partners). Large gene families include *HXT* genes, *PAU* genes and *RAS* genes, with more than 15 members each in the entire yeast genome. The chromosome XV ORF homologues include 426 different ORFs from the other chromosomes (7.8% of the yeast genome). Distribution of duplicated ORFs along chromosome XV using this method gives roughly the same results as the first method (Fig. 1).

Chromosome XV contains several local ORF duplications in tandem or inverted orientations. Those showing the highest degree of sequence

conservation are the tandem YOR229w and YOR230w already mentionned (63% amino-acid identity), and the inverted repeat YOR010c (*TIR2*) and YOR009w (62% identity). Other local duplications show significantly greater sequence divergence, suggesting more ancient events; they are represented by the tandems YOL083w and YOL082w (35% identity), YOR285w and YOR286w (37% identity), and YOL048c and YOL047c (38% identity, the latter ORF containing an intron). There also exists some 'local' duplications including pairs of ORFs, such as YOR162c and YOR172w (44% identity), or YOR381w and YOR384w (37% identity), that are separated by a few unrelated ORFs. They have also diverged, and may correspond to ancient local duplications that subsequentely received intervening DNA.

When this work started, 81 genes or markers were genetically mapped to chromosome XV, and 12 others were assigned to it but unmapped[2]. Seven of the unmapped genes and 55 of the mapped genes could be unambiguously assigned to ORFs or tRNA genes of the present sequence on the basis of previous partial sequence data, use of probes or gene function. Two genes, *ts26* and *PTP1*, originally mapped to chromosome XV, belong to chromosomes XII and IV (YLR268w and YDL230w, respectively). One gene, *TIR2* which corresponds to YOR010c, was originally named *SRP1*, creating confusion with YNL189w on chromosome XIV. Other than that, the original genetic map agrees fairly well with the present data, except for some local inversions of gene order, mostly around the centromere (*TOP1* and *SIN3* are on the left arm, *PEP12* is on the right) and in subtelomeric regions (*MEK1*, *PHR1* and *RAD17*), as is also observed for other chromosomes.

The chromosome XV sequence, like that of other yeast chromosomes, has been interpreted using criteria that are essentially predictive for ORFs, but are comparative with previously described sequences for other genetic features such as RNA-coding genes, Ty elements and the various chromosomal elements. It follows that the number of predicted ORFs is probably overestimated by a few percent compared with the number of actual protein-coding genes, whereas the identification of the other features should be considered as a minimum. Clear-cut identification of the questionable ORFs is not possible without independent experimental evidence, but it is suggested that they represent ~6% of all predicted ORFs. The *a posteriori* comparison of the predicted ORF products with general protein database entries forms the basis for the notion of 'orphan' genes[7]. Orphans are those protein-coding genes, predicted from the genomic sequence, that fail to show significant homology (at the chosen threshold value) when their translation products are compared to gene product sequences translated from all other genomic sequences present in public databases, whether they correspond to *S. cerevisiae* or any other organism. As is the case for other yeast chromosomes, a large fraction (33%) of chromosome XV ORFs are orphans without any significant structural homologue; while another 10.5% are orphans with structural homology to one or several other *S. cerevisiae* orphan(s). It is not surprising that, as with other genes, orphans can be duplicated in the yeast genome, or form diverged gene families. But what is more interesting is the significant deficit of these compared with the overall number of gene families in yeast. Of the 193 chromosome XV ORFs that are duplicated or parts of gene families, only 50 are orphans, compared with 83 predicted (43% of the 561 ORFs). The deficit of orphans among gene families (and the correlated excess of functionally characterized genes) is exactly opposite to the classical expectation from standard genetic screenings based on negative mutant phenotypes, which should tend to ignore isofunctional duplicated genes. One possible explanation for the bias observed may be that molecular methods, in contrast to classical genetic screenings, tend to facilitate the isolation of gene families that are structurally but not functionally related. Systematic functional analysis of the yeast genome, which is expected to follow the completion of the genomic sequence, should help to solve this important question. □

Methods

The sequence of chromosome XV was assembled from 46 cosmids covering the entire length of the chromosome except its left and right telomeres, which were sequenced from a rescued plasmid and a polymerase chain reaction (PCR) genomic product, respectively (Table 1). The cosmid map will be described separately. Because chromosome XV comigrates in PFGE with chromosome VII (which is only 348 bp shorter), the technique of chromosome fragmentation based on the insertion of unique artificial I-*Sce*I sites[13] played a key role in the physical mapping of these two chromosomes.

Each segment of the chromosome was sequenced on both strands using the methods and strategies indicated in Table 1. Overlaps between sequences submitted by different laboratories range from 218 bp to 16,799 bp (average 3,765 bp). In nearly all cases, overlapping sequences were also entirely determined on both strands by each laboratory and were found to be 100% identical. Only three differences were found in a total of 90,360 bp. After re-examination of the sequences, only one real divergence remained (an A to G transition), probably resulting from a mutation in one of the sequenced cosmids. The present chromosome sequence includes an A at position 400,735 but a G is equally probable. This uncertainty affects YOR036w (*PEP12*), by changing a codon CAG (Gln) to CGG (Arg).

After assembly, the entire sequence was verified as follows. A total of 201 short segments (259 bp–400 bp long) were selected after examination of the sequence using, as criteria, the possible occurrence of frameshifts, compressions (particularly in G+C-rich regions), and the presence of oligomeric stretches of mono- or dinucleotide repeats. Selected segments were attributed anonymously to four different laboratories and resequenced following the protocol of G. V. (unpublished). In total, 64,370 bp were verified revealing 21 original errors (13 nucleotides omitted, 3 nucleotides in excess, and 5 substitutions). Taken together with overlaps between different laboratories, 14.2% of the chromosome (154,730 bp) has thus been sequenced twice independently. The average sequence accuracy is 99.98%. This figure is probably an underestimate, however, as verifications were directed to suspicious regions. Parts of the present sequence were published independently by the sequencers before assembly of the contig and application of final quality controls[6,8,11,14–30]. Several other manuscripts are also in the press.

Received 23 July 1996; accepted 11 March 1997.

1. Hawthorne, D. C. & Mortimer, R. K. *Genetics* 60, 735–742 (1968).
2. Mortimer, R. K., Contopoulou, R. & King, J. S. *Yeast* 8, 817–902 (1992).
3. Oliver, S. G. *et al. Nature* 357, 38–46 (1992).
4. Dujon, B. *et al. Nature* 396, 371–378 (1994).
5. Feldmann, H. *et al. EMBO J.* 13, 5795–5809 (1994).
6. Gamo, F.-J. *et al. Yeast* 12, 709–714 (1996).
7. Dujon, B. *Trends Genet.* 12, 263–270 (1996).
8. Boyer, J. *et al. Yeast* 12, 1575–1586 (1996).
9. Goffeau A. *et al. Science* 274, 546–567 (1996).
10. Richard, G. F. & Dujon, B. *Gene* 174, 165–174 (1996).
11. Galisson, F. & Dujon, B. *Yeast* 12, 877–885 (1996).
12. Goffeau A. *et al. FEBS Lett.* 325, 112–117 (1993).
13. Thierry, A. & Dujon, B. *Nucleic Acids Res.* 20, 5625–5631 (1992).
14. Aldea, M. *et al. Yeast* 12, 1053–1058 (1996).
15. Bordonné, R. *et al. Yeast* 13, 73–83 (1997).
16. Casamayor, A. *et al. Yeast* 11, 1281–1288 (1995).
17. Casamayor, A. *et al. Yeast* 12, 1013–1020 (1996).
18. Casas, C. *et al. Yeast* 11, 1061–1068 (1995).
19. Chéret, G., Bernardi, A. & Sor, F. *Yeast* 12, 1059–1064 (1996).
20. Lafuente, M. J., Gamo, F-J. & Gancedo, C. *Yeast* 12, 1041–1045 (1996).
21. Madania, A. *et al. Yeast* 12, 1563–1573 (1996).
22. Mannhaupt G. *et al. Yeast* 12, 67–76 (1996).
23. Parle-McDermott, A. G. *et al. Yeast* 12, 999–1004 (1996).
24. Pearson, B. M. *et al. Yeast* 12, 1021–1031 (1996).
25. Purnelle, B. & Goffeau, A. *Yeast* 12, 1475–1481 (1996).
26. Vandenbol, M., Durand, P. Portetelle, D. & Hilger, F. *Yeast* 11, 1069–1075 (1995).
27. Wiemann, S. *et al. Yeast* 12, 281–288 (1996).
28. Zumstein, E., Griffin, H. & Schweizer, M. *Yeast* 10, 1383–1387 (1994).
29. Zumstein, E., Pearson, B. M., Kalogeropoulos, A. & Schweizer, M. *Yeast* 11, 975–986 (1995).
30. Sterky, F. *et al. Yeast* 12, 1091–1095 (1996).

Acknowledgements. This work is part of the third phase of the European Yeast Genome Sequencing Project carried out under the administrative coordination of A. Vassarotti (DG-XII) and the Université Catholique de Louvain. We thank P. Mordant for accounting and help; P. Jordan and S. Liebl for computer support; and S. Oliver and colleagues in the various laboratories for discussions. This work was supported by the European Commission under the Biotech II Program with additional contributions from national or local sources: Services Fédéraux des Affaires Scientifiques, Techniques et Culturelles, Pôles d'attraction inter-universitaire, Région Wallone and Katholieke Universiteit Leuven (Belgium); GREG, CNRS, Institut Pasteur and Institut Curie (France); DGICYT, CICYT and PFPI (Spain) and the Wellcome Trust (UK).

Correspondence and requests for materials should be addressed to B. D. (e-mail: bdujon@pasteur.fr).
Details of the sequence data set are available at http://speedy.mips.biochem.mpg.de/mips/yeast/