The nucleotide sequence of Saccharomyces cerevisiae chromosome XIV and its evolutionary implications

P. Philippsen, K. Kleine, R. Pöhlmann, A. Düsterhöft, K. Hamberg, J.H. Hegemann, B. Obermaier, L.A. Urrestarazu, R. Aert, K. Albermann, et al.

The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications

P. Philippsen[1,2], K. Kleine[3], R. Pöhlmann[1], A. Düsterhöft[4], K. Hamberg[2], J. H. Hegemann[2], B. Obermaier[5,6], L. A. Urrestarazu[7], R. Aert[8], K. Albermann[3], R. Altmann[1], B. André[7], V. Baladron[9], J. P. G. Ballesta[10], A.-M. Bécam[11], J. Beinhauer[2], J. Boskovic[10], M. J. Buitrago[9], F. Bussereau[12], F. Coster[13], M. Crouzet[14], M. D'Angelo[15], F. Dal Pero[15], A. De Antoni[15], F. Del Rey[9], F. Doignon[14], H. Domdey[5], E. Dubois[16], T. Fiedler[2], U. Fleig[2], M. Floeth[4], C. Fritz[4], C. Gaillardin[17], J. M. Garcia-Cantalejo[10], N. N Glansdorff[16], A. Goffeau[13], U. Gueldener[2], C. Herbert[11], K. Heumann[3], D. Heuss-Neitzel[4], H. Hilbert[4], K. Hinni[1], I. Iraqui Houssaini[7], M. Jacquet[12], A. Jimenez[10], J.-L. Jonniaux[13], L. Karpfinger[3], G. Lanfranchi[15], A. Lepingle[17], H. Levesque[17], R. Lyck[2], M. Maftahi[17], L. Mallet[12], K. C. T. Maurer[18], F. Messenguy[16], H. W. Mewes[3], D. Möstl[4], F. Nasr[11], J.-M. Nicaud[17], R. K. Niedenthal[2], D. Pandolfo[15], A. Piérard[16], E. Piravandi[5], R. J. Planta[18], T. M. Pohl[19], B. Purnelle[13], C. Rebischung[1], M. Remacha[10], J. L. Revuelta[9], M. Rinke[5], J. E. Saiz[9], F. Sartorello[15], B. Scherens[16], M. Sen-Gupta[2], A. Soler-Mira[10], J. H. M. Urbanus[18], G. Valle[15], L. Van Dyck[13], P. Verhasselt[8], F. Vierendeels[16], S. Vissers[7], M. Voet[8], G. Volckaert[8], A. Wach[1], R. Wambutt[20], H. Wedler[20], A. Zollner[3] & J. Hani[3]

[1]Institute for Applied Microbiology, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland
[2]Justus-Liebig-Universität Giessen, Institut für Mikro- und Molekularbiologie, Frankfurter Strasse 107, D-35392 Giessen, Germany
[3]Martinsrieder Institut für Protein Sequenzen, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany
[4]QIAGEN GmbH, Max-Volmer-Strasse 4, D-40724 Hilden, Germany
[5]Laboratorium für Molekulare Biologie, Genzentrum der LMU München, Feodor-Lynen-Strasse 25, D-811377 München, Germany
[6]MediGene GmbH, Lochhamer Strasse 11, D-82152 Martinsried, Germany
[7]Université Libre de Bruxelles, Physiologie Cellulaire et Génétique des Levures, Boulevard du Triompe CP244, B-1050 Bruxelles, Belgium
[8]Katholieke Universiteit Leuven, Laboratory of Gene Technology, Willem de Croylaan 42, B-3001 Leuven, Belgium
[9]Departamento de Microbiologia y Genética, Universidad de Salamanca, Avenida del Campo Charro s/n, E-37007 Salamanca, Spain
[10]Centro de Biologia Molecular, CSIC & UAM, Cantoblanco, E-28049 Madrid, Spain
[11]Centre de Génétique Moléculaire, Laboratoire propre du CNRS associé à l'Université Pierre et Marie Curie, F-91198 Gif-sur-Yvette, France
[12]Université Paris-Sud, Institut de Génétique et Microbiologie, Laboratoire Information Génétique et Développement, Bât. 400, F-91405 Orsay Cedex, France
[13]Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2/20, B-1348 Louvain-la-Neuve, Belgium
[14]LBMS, Université de Bordeaux 2, UPR CNRS 9026, BP 64, 146 rue Léo Saignat, F-33076 Bordeaux Cedex, France
[15]Department of Biology, CRIBI Biotechnology Centre, University of Padova, via Trieste, 75, I-35121 Padova, Italy
[16]CERIA-COOVI, Avenne E. Gryson 1, B-1070 Brussels, Belgium
[17]Institut National Agronomique Paris-Grignon, Laboratoire de Génétique Moléculaire et Cellulaire, Centre de Biotechnologies Agro-Industrielles, F-78850 Thiverval-Grignon, France
[18]Department of Biochemistry and Molecular Biology, IMBW, BioCentrum Amsterdam, Vrije Universiteit de Boelelaan 1083, NL-1081 HV Amsterdam, Netherlands
[19]GATC-Gesellschaft für Analyse-Technik und Consulting mbH, Fritz-Arnold-Strasse 23, D-78467 Konstanz, Germany
[20]AGON GmbH, Glienicker Weg 185, D-12489 Berlin, Germany

In 1992 we started assembling an ordered library of cosmid clones from chromosome XIV of the yeast *Saccharomyces cerevisiae*. At that time, only 49 genes were known to be located on this chromosome[1] and we estimated that 80% to 90% of its genes were yet to be discovered. In 1993, a team of 20 European laboratories began the systematic sequence analysis of chromosome XIV. The completed and intensively checked final sequence of 784,328 base pairs was released in April, 1996 (ref. 2). Substantial parts had been published before[3–22] or had previously been made available on request. The sequence contained 419 known or presumptive protein-coding genes, including two pseudogenes and three retrotransposons, 14 tRNA genes, and three small nuclear RNA genes. For 116 (30%) protein-coding sequences, one or more structural homologues were identified elsewhere in the yeast genome. Half of them belong to duplicated groups of 6–14 loosely linked genes, in most cases with conserved gene order and orientation (relaxed interchromosomal synteny). We have considered the possible evolutionary origins of this unexpected feature of yeast genome organization.

Figure 1 shows the map of cosmid, lambda and plasmid clones and of polymerase chain reaction (PCR) fragments from two unclonable regions which were used to determine the sequence of chromosome XIV. The final positions of genes listed in the 1992 map[1] are also presented changing the order of closely linked genes in only three regions. The assembled contig consists of 784,328 bp. The sequence of 180,983 bp (23%) was independently determined twice on both strands. These control sequences included 28 overlapping regions of cosmid and lambda clones (117,891 bp) as well as 108 selected regions, mainly at termini of open reading frames (ORFs), resequenced either on cosmids (54,540 bp) or by genomic PCR (8,552 bp). A total of 27 sequence mistakes were corrected. We estimate that the final sequence carries less than one error in every 10 kilobases, an estimate confirmed by a recent independent control analysis using 83 randomly picked genomic clones of chromosome XIV (G. Valle, unpublished data). Among the 40 kb sequenced, four deviations from our final sequence were noted: three single base-pair changes with neutral effects on coding regions (probably resulting from strain or clone differences), and only one confirmed sequence mistake. The left end of the chromosome carries telomeric repeat sequence (see below) and it is

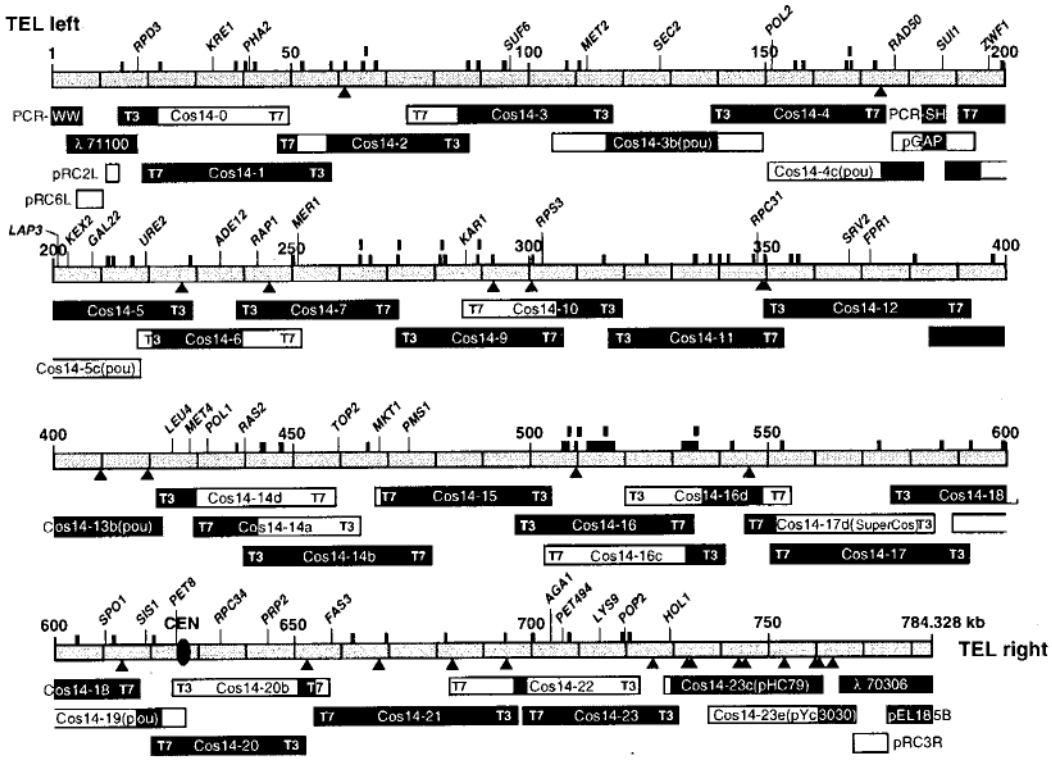

Figure 1 Physical map of subclones of chromosome XIV used for systematic DNA sequence analysis and final locations of originally genetically mapped genes[1]. Position of cosmid clones (cos), lambda clones (λ), plasmid clones (p) and genomic PCR fragments (PCR) are drawn as overlapping bars. Sequenced regions are shown in black. The 108 short regions selected for verification analyses are shown as small bars (resequenced cosmid clones) or triangles (sequenced genomic PCR fragments) along the contig. All clones were derived from S288C strains, except plasmid pGAP (carrying a spontaneus nonsense mutation in the toxic YNL247w), which originates from strain A364a[36]. Most cosmid clones with chromosome XIV DNA were isolated from cosmid libraries provided by B. Dujon[37] and R. Stucka[38], and mapped by a modified chromosome fragmentation approach[39,40]. Several clones extending into or bridging remaining gaps were isolated by colony screening using non-radioactively labelled restriction fragments as hybridization probes. Two cosmid clones (14–17d and 14–23c) and both lambda clones carrying telomere DNA were obtained from L. Riles[41] and the right telomere clone pEL185 was provided by E. Louis[42]. A more detailed description of the mapping strategy will be published elsewhere (K. Hamberg *et al.*, manuscript in preparation). The complete sequence can be retrieved from the EMBL database, accession nos Z71277–Z71692 or from the Martinsrieder website[2]. Different parts were sequenced in different laboratories: 1–6035 (R. Wambutt); 3,203–17,700 (B. Obermeier); 13,990–22,212 (A. Goffeau); 18,699–58,748 (C. Gaillardin); 47,022–51,246 and 57,523–87,525 (R. J. Planta); 85,152–132,424 (N. N. Glansdorf); 130,724–187,891 (J. H. Hegemann); 183,004–187900 (P. Philippsen); 187,809–192,153 (F. Del Rey); 192,154–195,234 (A. Jimenez); 190,506–229,360 (G. Valle); 220,854–239,907 (A. Düsterhöft); 238,582–273,742 (A. Goffeau); 271,932–319,898 (H. Domdey); 317,148–353,960 (C. Herbert); 349,559–393,039 (M. Jacquet); 384,059–421,858 (G. Valle); 421,188–443,100 (J.-L. Revuelta & F. Del Rey); 443,001–456,300 (A. Jimenez); 456,201–479,289 (J. P. G. Ballesta); 468,833–504,727 (P. Philippsen); 496,969–541,433 (M. Crouzet); 536,275–549,131 (C. Herbert); 545,180–592,214 (A. Düsterhöft); 575,858–617,912 (A. Urrestarazu); 617,105–622,324 (M. Crouzet); 620,016–652,539 (G. Volckaert.); 650,830–653,557 (R. J. Planta.); 651,995–654,446 (A. Düsterhöft); 654,389–731,357 (T. M. Pohl); 729,267–768,530 (A. Düsterhöft); 764,973–784,328 (A. Urrestarazu); 774,980–784,145 (C. Gaillardin).

Table 1 *S. cerevisiae* chromosome XIV ORFs and structurally homologous ORFs of other chromosomes

ORF Chr.XIV* 6–17.3 kb†	ORF Chr.VI* 6–15.5 kb	% identity/ stretch of amino acids	Biochemical or biological function (gene name) Chr. XIV ORF	Homologue in cluster duplication
YNL336w	YFL062w	94.2% overall	unknown	unknown
YNL335w	YFL061w	100 % overall	fungal cyanamide hydratase homologue	fungal cyanamide hydratase homologue
YNL334c	YFL060c	99.1% overall	unknown, probable membrane protein	unknown, probable membrane protein
YNL333w	YFL059w	99.7% overall	unknown	unknown
YNL332w	YFL058w	99.7% overall	thiamine regulated protein homologue	thiamine regulated protein (THI5)
YNL331c	YFL057–56c	87.0%/226 aa‡	probable aryl-alcohol reductase	probable aryl-alcohol reductase
ORF Chr.XIV 38.7–106.7 kb	ORF Chr.XV (A) 25.3–142.6 kb	% identity/ stretch of amino acids		
YNL318c	YOL156w	38.4%/510 aa	hexose transporter (HXT14)	glucose transporter (LGT3)
YNL307c	YOL128c	41.6%/316 aa	Ser/Thr/Tyr protein kinase (MCK1)	probable Ser/Thr protein kinase
YNL302c	YOL121c	99.3% overall	ribosomal protein (RPS16A)	ribosomal protein (RPS16B)
YNL301c	YOL120c	100% overall	ribosomal protein (RP28B)	ribosomal protein (RP28A)
YNL299w	YOL115w	54.1%/556 aa	topoismerase I related protein (TRF5)	topoisomeras I related protein (TRF4)
YNL298w	YOL113w	61.2%/317 aa	Ser/Thr protein kinase (CLA4)	probable Ser/Thr protein kinase
YNL293w	YOL112w	53.2%/417 aa	unknown	unknown
YNL290w	YOL094c	34.4%/317 aa	replication factor C subunit (RFC3)	replication factor C subunit (RFC4)
YNL283c	YOL105c	43.4%/302 aa	similarity to yeast chitinase	unknown
ORF Chr.XIV 252.1–307 kb	ORF Chr.IV 44.1–80.4 kb	% identity/ stretch of amino acids		
YNL209w	YDL229w	99.3% overall	heat shock protein (SSB2)	heat shock protein (SSB1)
YNL204c	YDL226c	30.5%/177 aa	sporul.spec.zinc finger protein (SPS18)	prolif.spec.zinc finger protein (GCS1)
YNL197c	YDL224c	36.5%/581 aa	regulator of cell size (WHI3)	unknown
YNL194c	YDL222c	52.0% overall	unknown, probable membrane protein	unknown, probable membrane protein
YNL183c	YDL214c	42.6%/479 aa	Ser/Thr protein kinase (NPR1)	probable Ser/Thr protein kinase
YNL176c	YDL211c	24.3%/292 aa	unknown, probable membrane protein	unknown, probable membrane protein
ORF Chr.XIV 309–410 kb	ORF Chr.VIII 390.3–341.4 kb	% identity/ stretch of amino acids		
YNL173c	YHR146w	27.4%/351 aa	pheromone-response G protein	unknown, probable G protein
YNL162w	YHR141c	100% overall	ribosomal protein (RPL41A)	ribosomal protein (RPL41A)
YNL160w	YHR139c	45.0%/307 aa	secreted glycoprotein (YGP1)	sporulat.spec. wall maturation (SPS100)
YNL156c	YHR133c	40.5%/205 aa	unknown	unknown
YNL154c	YHR135c	70.9%/499 aa	casein kinase I isoform (YCK2)	casein kinase I (YCK1)
YNL144c	YHR131c	36.2%/464 aa	unknown	unknown
YNL130c	YHR123w	53.8% overall	diacylglyc.choline-P transferase (CPT1)	ethanolamin P-tranferase (EPT1)
YNL121c	YHR117w	49.3%/651 aa§	import recept.mito outer memb.(TOM70)	mitochondrial outer membrane protein
YNL116w	YHR115c	55.4%/424 aa	unknown	unknown
ORF Chr.XIV 419–466 kb	ORF Chr.XV (B) 529–486.8 kb	% identity/ stretch of amino acids		
YNL108c	YOR110w	65.0%/273 aa	unknown	unknown
YNL106c	YOR109w	58.5%/979 aa	inositol phosphatase homologue	probable phosphatase
YNL104c	YOR108w	88.5%/601 aa	2-isopropyl malate synthase (LEU4)	2-isopropyl malate synthase homologue
YNL098c	YOR101w	55.8%/303 aa	GTP-binding protein (RAS2)	GTP-binding protein (RAS1)
YNL096c	YOR096w	87.9% overall	ribosomal protein S7 homologue	ribosomal protein (RP30)
YNL095c	YOR092w	55.3%/445 aa	unknown, probable membrane protein	unknown, probable membrane protein
YNL093w	YOR089c	56.9%/209 aa	GTP-binding protein (YPT53)	GTP-binding protein (VPS21)
YNL090w	YOR089c	56.9%/209 aa	GTP-binding protein (RHO2)	GTP-binding protein (VPS21)
YNL087w	YOR086c	54.5% overall	unknown, probable membrane protein	unknown, probable membrane protein
ORF Chr.XIV 478.6–597.6 kb	ORF Chr.IX 89.3–202.1 kb	% identity/ stretch of amino acids		
YNL079c	YIL138c	54.1%/159 aa	tropomyosin (TPM1)	tropomyosin (TPM2)
YNL074c	YIL135c	23.2%/375 aa	unknown	unknown
YNL069c	YIL133c	90.3% overall	ribosomal protein (RP23)	ribosomal protein (RP22)
YNL068c	YIL131c	52.1%/190 aa	unknown, fork head domain (FKH2)	unknown, fork head domain (FKH1)
YNL066w	YIL123w	62.9%/415 aa	ß-glucosidase homologue (SUN4)	homologue of aging gene UTH1
YNL065w	YIL121w	47.7%/342 aa	cycloheximid resist.protein homologue	antibiotic resitance protein homologue
YNL065w	YIL120w	43.0%/351 aa	cycloheximid resist.protein homologue	antibiotic resitance protein homologue
YNL058c	YIL117c	37.3%/126 aa	unknown	unknown
YNL055c	YIL114c	49.5% overall	outer mito membrane porin (OMP2)	OMP2 homologue
YNL053w	YIL113w	48.8%/162 aa	protein phosphatase (MSG5)	protein-Tyr phosphatase homologue
YNL052w	YIL111w	63.6% overall	cytochrome c oxidase (COX5A)	cytochrome c oxidase (COX5B)
YNL049c	YIL109c	61.8%/566 aa	unknown	unknown
YNL047c	YIL105c	54.3%/639 aa	unknown	unknown
YNL037c	YIL094c	35.8%/296 aa	isocitrate dehydrogenase (IDH1)	isopropyl malate dehydrog. homologue
YNL029c	YIL085c	55.7%/476 aa	mannosyl transferase homologue	mannosyl transferase homologue
YNL020c	YIL095w	41.2%/636 aa	probable Ser/Thr protein kinase	probable Ser/Thr protein kinase
ORF Chr.XIV 623.4–753.7 kb	ORF Chr.III 101.7–301.8 kb	% identity/ stretch of amino acids		
YNL004w	YCL011c	39.4%/409 aa	poly(A)bdg.protein homologue (TOM34)	probable TEL associated protein (GBP2)
YNR001c	YCR005c	81.4%/441 aa	citrate synthase (CIT1)	peroxysomal citrate synthase (CIT2)
YNR002c	YCR010c	77.7% overall	unknown, probable membrane protein	unknown, probable membrane protein
YNR013c	YCR037c	48.1%/489 aa	unknown, probable membrane protein	probable phophate transporter (PHO87)
YNR019w	YCR048w	52.5%/459 aa	sterol acyltransferase (SAT1)	cholesterol acyltransferase (ARE1)
YNR023w	YCR052w	29.5%/353 aa	unknown	unknown
YNR026c	YCR067c	45.4%/388 aa	GTP-GDP exchange factor (SEC12)	ER protein (SED4)
YNR028w	YCR069w	33.2%/304 aa	peptidyl-prolyl isomerase homologue	peptidyl-prolyl cis-trans isom.(SCC3)
YNR031c	YCR073c	53.6%/1172 aa	MAPKKK high osm.sign.transd. (SSK2)	MAP kinase kinase kinase (SSK22)
YNR034w	YCR073w-a	76.9% overall	multicopy sup of los1-1 (SOL1)	GlcN-6-P deaminase homologue (SOL2)
YNR047w	YCR091w	72.4%/424 aa	probable Ser/Thr protein kinase	probable Ser/Thr protein kinase (KIN82)
YNR048w	YCR094w	65.4% overall	unknown	unknown
YNR065-66c	YCR099-101c	64.8%/637 aa¶	peptidase Y sorting protein (pseudogene)	peptidase Y sorting (PEP1 homologue)

Degrees of homology were extracted from pairwise FASTA alignments of deduced protein sequences and are listed as percentage identity per stretch of amino acids.

*'Y' is included in Fig. 2 in the cluster duplication but is not listed in this table.

†Coordinates of clusters.

‡Overall homology between YNL331c and the sum of YFL056c and YFL057c (pseudogene in chromosomeVI?).

§Gaps introduced by the alignment algorithm may result in homology stretches slightly longer than the protein sequences.

¶Overall homology of the pseudogene (YNR065–YNR066c) to the sum of YCR099c, YCR100c and YCR101.

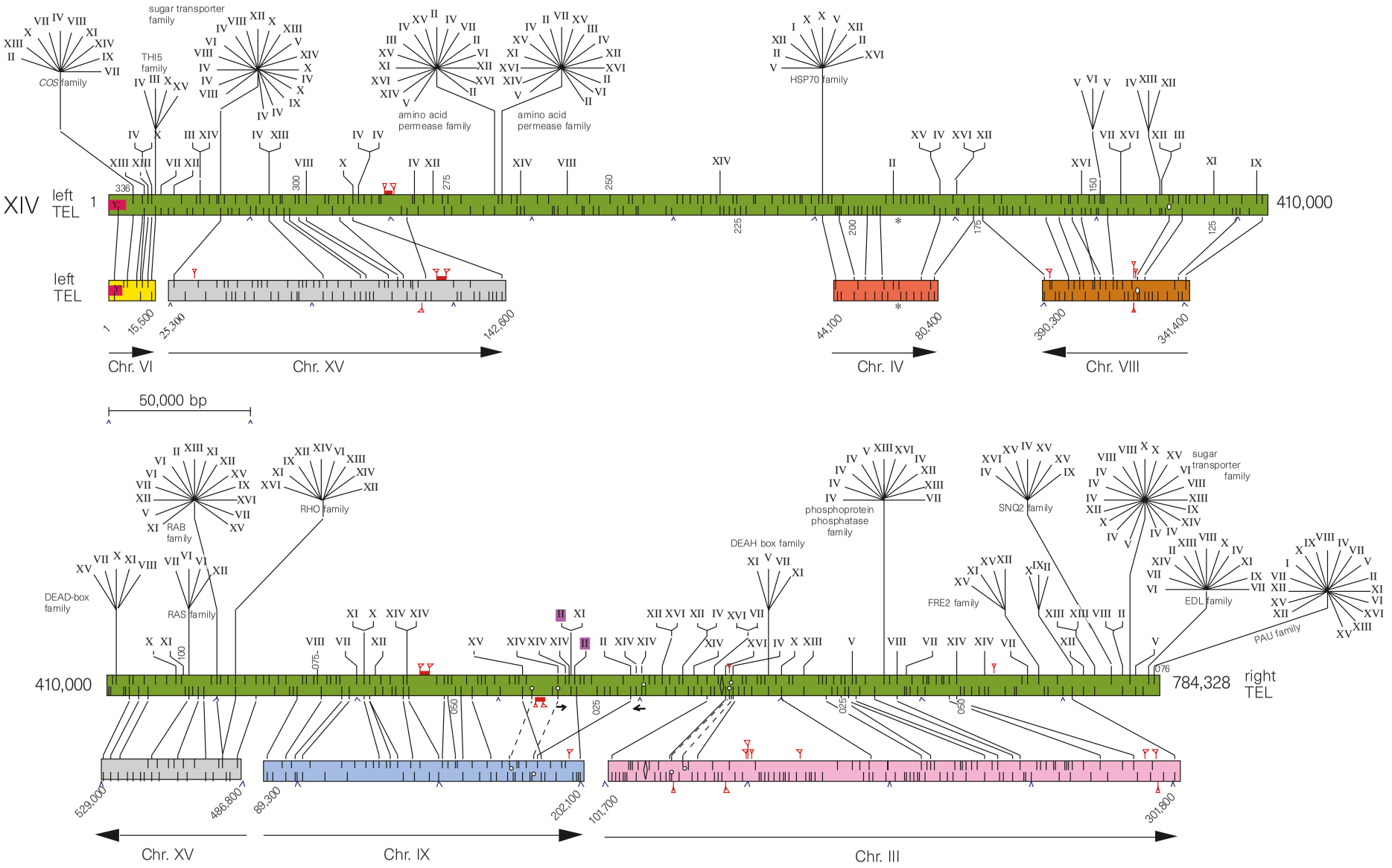
XIV
left TEL
1
410,000
COS family
THI5 family
sugar transporter family
amino acid permease family
amino acid permease family
HSP70 family
left TEL
Chr. VI
Chr. XV
Chr. IV
Chr. VIII
50,000 bp
RAB family
RHO family
phosphoprotein phosphatase family
DEAH box family
SNQ2 family
sugar transporter family
DEAD-box family
RAS family
FRE2 family
EDL family
PAU family
410,000
784,328
right TEL
Chr. XV
Chr. IX
Chr. III

Figure 2 Map of chromosome XIV ORFs that are members of either multigene families or of pairs, triplets or quadruplets of structurally related *S. cerevisiae* ORFs. The green bar represents both strands of chromosome XIV, with centres of all ORFs (excluding Ty and four short telomere ORFs) drawn as vertical lines; 215 are coded by the upper strand and 195 by the lower strand. Vertical lines with open circles mark selected tRNA genes. Three-digit numbers beneath or above the green bar refer to the systematic ORF nomenclature starting with 1 at either side of the centromere (white dot at 628 kb). Lines above or below the green bar indicate ORFs with structural homologues elsewhere in the genome (at least 30% identity in 150 amino acids, or, in a few cases, 25% in 300 amino acids). Lines extending into branches mark multigene families, with roman numbers indicating members on different chromosomes (order of decreasing homology to the chromosomes XIV ORF from left to right or clockwise, respectively). Coloured bars below chromosome XIV display seven syntenic or partly syntenic segments of other chromosomes with accumulations of ORFs structurally related to and arranged similarly to chromosome XIV ORFs. Broken lines in three of these clusters connect positions of pairs of functionally identical tRNA genes. The star at 280 kb indicates a functional ARS element on chromosome XIV (Ref. 36) which seems to be positionally conserved on chromosome IV. Further details of these cluster duplications are given in Table 1. The red bar of the left telomere represents the ubiquitous Y′ element found at many ends of *S. cerevisiae* chromosomes[43,44]. Red triangles mark positions of solo delta sequences (remnants of Ty elements) and red bars flanked by triangles indicate Ty elements. The two black arrows at 570 kb and 600 kb indicate an intrachromosomal highly conserved inverted repeat, involving in each repeat element one $tRNA^{Ile}$ gene and two new ORFs (YNL034w–YNL035w and YNL019c–YNL018c, respectively). The two marked chromosome II homologues and the corresponding chromosome XIV ORFs at 575 kb represent the two copies of the duplicated histone H3–H4 gene pair. As indicated there is an additional homologue to histone H3 on chromosome XI (*CSE4*), probably the yeast homologue of the human *CENP-A* gene[45]

possble that this end is a few hundred base pairs longer than indicated in Figure 1.

A systematic search of the chromosome XIV contig revealed 414 ORFs with 100 and more codons, including overlapping ORFs but excluding ORFs located within longer ORFs on either the same or the complementary strand. Chromosome XIV also has at least seven ORFs with less than 100 codons, of which four are known genes (*MFA2*, *TOM7*, *ATX1* and *PBI2*) and three show significant homology to known genes. A systematic nomenclature was given to all ORFs (excluding the six ORFs of the three Ty retrotransposons), indicating the organism (Y), the chromosome (N), the chromosome arm (L or R), the coding strand (Watson, w or Crick, c), and increasing numbers starting at the centromere; examples include YNL001w and YNR001c.

A simultaneous search for introns (using the EXPLORA program[23]) revealed 16 intron-carrying genes, in two of which, (YNL066w and YNL065w) the introns are located in the non-translated 5′ region[24,25]. EXPLORA failed to locate an additional, experimentally verified intron in ORF YNL044w, because it has an unusual 5′ splice sequence (EMBL database, accession nos X97400 and X97401).

Two pairs of adjacent ORFs (YNR065c and YNR066c; and YNR068c and YNR069c) were separated only by a stop codon; this was confirmed in both cases by genomic PCR. These pairs are rare examples of yeast pseudogenes, as highly conserved copies lacking internal stop codons are present on other chromosomes. Like their functional homologues these pseudogenes should be considered as single ORFs. Taking this into account, 419 ORFs are located on chromosome XIV, including six Ty-elements and 23 questionable ORFs (short ORFs overlapping longer ones). The ORF density, not counting questionable ORFs, is one ORF per 1.98 kb (a total of 396 ORFs in 784 kb), and the average ORF size is 1.5 kb. These numbers are very similar to corresponding numbers obtained with other *S. cerevisiae* chromosomes. The ORF density (the ratio of ORF nucleotides to total nucleotides) fluctuates between 0.6 and 0.9. These fluctuations do not correlate with fluctuations in G+C content; five of eight ORF density peaks coincide with regions of highest G+C content (39.4–40.0%) and the other three with regions of lowest G+C content (36.6–37.7%).

How many of the 396 non-questionable ORFs are new? Presently, functions are known, at least partly, for 149 ORFs (38%), based on detailed experiments or very high sequence homology to known genes[2,26] . Most of these are involved in metabolism, cell growth, cell division, translation, transcription and intracellular transport, with a few involved in energy production, metabolite transport, protein modification, signal transduction, and stress response. Of the 247 new ORFs, some functional predictions can be made for 43 (11%), owing to homologies to characterized genes in *S. cerevisiae* or other organisms. Presumptive products coded by these 43 ORFs include: a human breast cancer-associated autoantigen homologue; a genetically linked cluster of three proteins (transporter, epimerase and reductase) for potential utilization of an unidentified mono- or oligosaccharide; four proteins with homology to prokaryotic ribosomal proteins; three protein kinases; three GTP-binding proteins; two protein phosphatases; two translation factors; two drug- resistance proteins; one actin homologue; one zinc finger; one peptidyl-prolyl isomerase; and ten with presumptive metabolic activities, such as cyanamide hydratase, mannosyl transferase, isocitrate dehydrogenase and inositol phosphatase. Further details can be found on the Martinsrieder website[2].

The functions of the other 204 ORFs (51%) cannot yet be predicted. One third of these code for presumptive membrane proteins, and more than four transmembrane domains are predicted for 18. Of the 204 new ORFs, 12 have homology to human expressed sequence tags (EST)[27, 28] with FASTA scores of 200–760. Remarkably, two of the 23 questionable ORFs (YNL228w and YNL114c) also have significant homology to human EST sequences.

We used FASTA comparisons of all chromosome XIV ORFs (except the highly repetitious Y′ and Ty ORFs) to all *S. cerevisiae* ORFs in order to establish the extent of gene duplications, and found that 116 ORFs shared structural homology with one or more ORFs elsewhere in the genome. For this search, structural homology was defined as over 30% identity in a stretch of 150 amino acids (in some cases, 25% identity in a stretch of at least 300 amino acids). Of these 116 ORFs, 67 belong to pairs of homologues, 32 to groups consisting of three or four homologues, and 17 are members of multigene families. ORFs from all chromosomes contribute to this picture of sequence homology (Fig. 2). The list of homologies based on FASTA analyses also revealed several regions of chromosome XIV with accumulations of homologous ORFs originating from distinct regions of six other chromosomes, and showing, with only a few exceptions, conserved gene orders and gene orientations. One of these apparently ancient duplications, involving ORFs of the left arms of chromosomes IX and XIV, respectively, had previously been reported[19,29]. Duplications involving several genes had been described up to that time, mainly for relatively short subtelomeric and centromeric regions[30–34].

The extent of these types of duplications became apparent after the complete sequence information of the *S. cerevisiae* reference strain S288C was released[2] . With respect to chromosome XIV, so-called gene cluster duplications were found in seven regions of 17 kb to 130 kb. The precise locations of the 67 pairs of ORF homologues in these seven cluster duplications are shown in Fig. 2, together with all other chromosome XIV ORFs for which structural homologues were found; five pairs of positionally conserved duplicated tRNA genes are also indicated. Probably half of these structural homologies among different chromosomes would have remained undetected in classical DNA hybridization experiments.

Complementary to the graphical display of the seven cluster duplications, we have determined the degree of homologies for each ORF pair and, if known or predictable, their functions (Table 1). ORFs displayed from left to right in Fig. 2 are listed from top to bottom in the table. An automated means of finding and displaying structurally homologous segments in genomes several million base pairs long involves the screening of sliding windows of 500 bases between pairs of chromosomes[35]. This very efficient method was also applied to chromosome XIV, and most of the ORF pairs participating in cluster duplications were detected (K. Heumann, unpublished data). However, this automated approach still

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.

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.