Complete DNA sequence of yeast chromosome XI

B. Dujon[1], D. Alexandraki[2], B. André[3], W. Ansorge[4], V. Baladron[5], J. P. G. Ballesta[6], A. Banrevi[4]*, P. A. Bolle[7], M. Bolotin-Fukuhara[8], P. Bossier[9], G. Bou[6], J. Boyer[1], M. J. Buitrago[5], G. Chéret[10], L. Colleaux[1]*, B. Dalgnan-Fornier[8], F. del Rey[5], C. Dion[7], H. Domdey[11], A. Düsterhöft[12,13], S. Düsterhus[14], K-D. Entian[14], H. Erfle[4], P. F. Esteban[5], H. Feldmann[15], L. Fernandes[9], G. M. Fobo[16], C. Fritz[17], H. Fukuhara[10], C. Gabel[18], L. Gaillon[1], J. M. Carcia-Cantalejo[6], J. J. Garcia-Ramirez[5], M. E. Gent[19], M. Ghazvini[1]*, A. Goffeau[20,21], A. Gonzaléz[5], D. Grothues[4], P. Guerreiro[9], J. Hegemann[12], N. Hewitt[4], F. Hilger[7], C. P. Hollenberg[17], O. Horaitis[2]*, K. J. Indge[19], A. Jacquier[1], C. M. James[19], J. C. Jauniaux[3,22], A. Jimenez[6], H. Keuchel[17], L. Kirchrath[17], K. Kleine[16], P. Kötter[14], P. Legrain[1], S. Liebl[16], E. J. Louis[23], A. Maia e Silva[9], C. Marck[24], A-L. Monnier[1], D. Möstl[13], S. Müller[18], B. Obermaier[11], S. G. Oliver[19], C. Pallier[8], S. Pascolo[1]*, F. Pfeiffer[16], P. Philippsen[12,25], R. J. Planta[26], F. M. Pohl[27], T. M. Pohl[28], R. Pöhlmann[12,25], D. Portetelle[7], B. Purnelle[20], V. Puzos[8], M. Ramezani Rad[17], S. W. Rasmussen[29], M. Remacha[6], J. L. Revuelta[5], G-F. Richard[1], M. Rieger[18], C. Rodrigues-Pousada[9], M. Rose[14], T. Rupp[4], M. A. Santos[5], C. Schwager[4], C. Sensen[4], J. Skala[20]*, H. Soares[9], F. Sor[10], J. Stegemann[4], H. Tettelin[20], A. Thierry[1], M. Tzermia[2], L. A. Urrestarazu[3], L. van Dyck[20], J. C. van Vliet-Reedijk[26]*, M. Valens[8], M. Vandenbol[7], C. Vilela[9], S. Vissers[3], D. von Wettstein[29], H. Voss[4], S. Wiemann[4], G. Xu[17], J. Zimmermann[4], M. Haasemann[16]*, I. Becker[16] & H. W. Mewes[16]

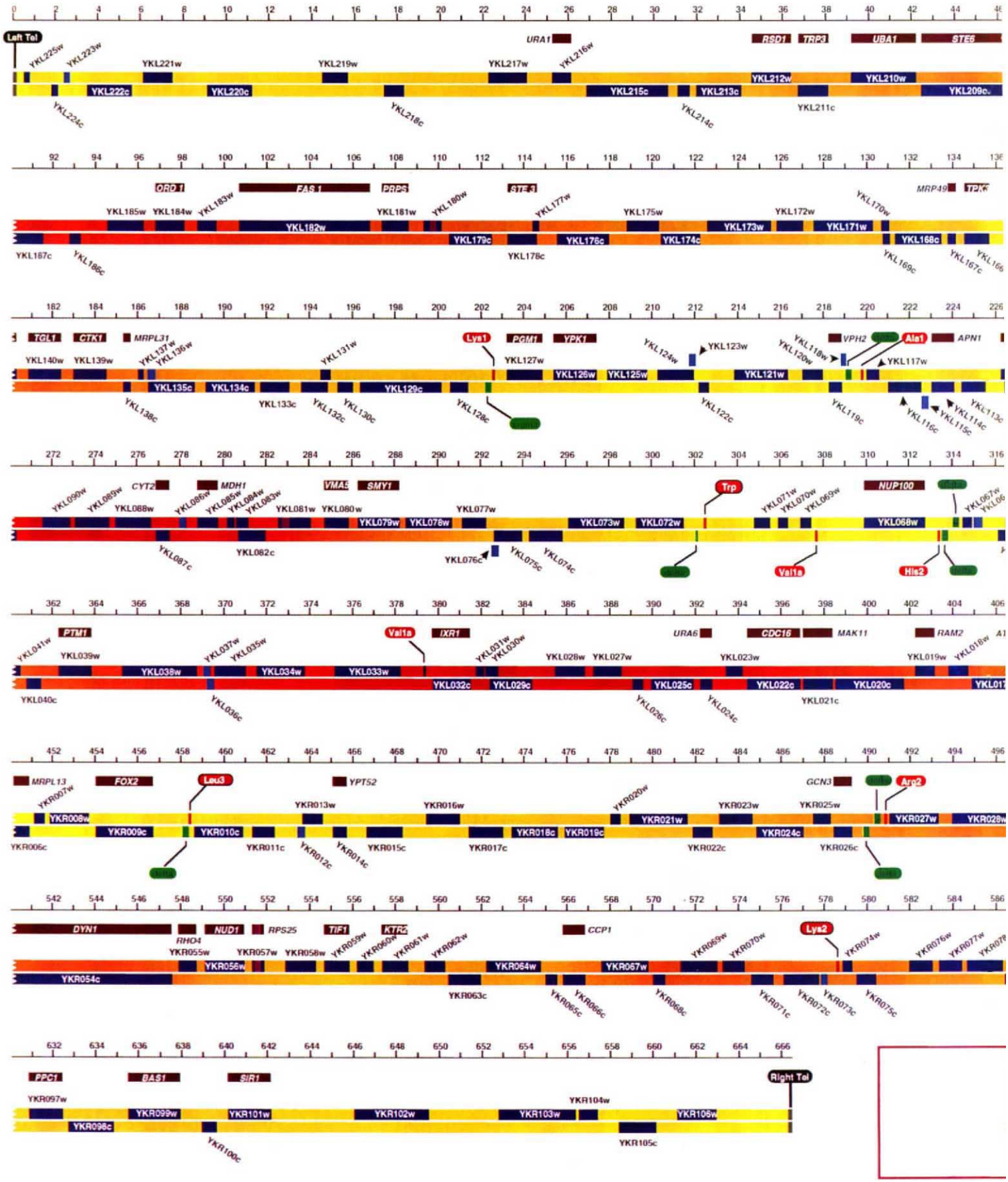
The complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome XI has been determined. In addition to a compact arrangement of potential protein coding sequences, the 666,448-base-pair sequence has revealed general chromosome patterns; in particular, alternating regional variations in average base composition correlate with variations in local gene density along the chromosome. Significant discrepancies with the previously published genetic map demonstrate the need for using independent physical mapping criteria.

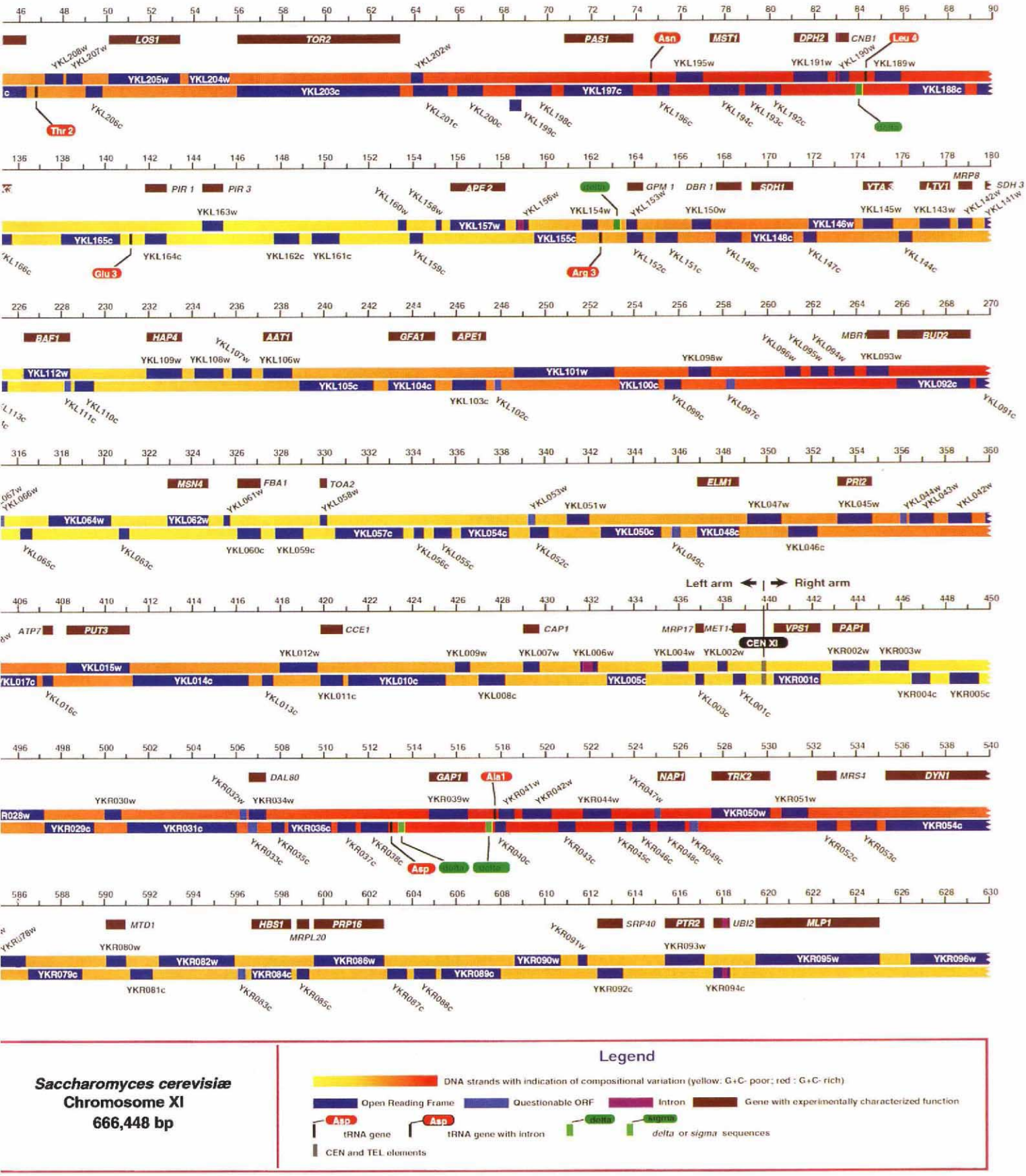
DESPITE recent technical developments[1-4], the availability of complete sequences of complex genomes, such as those of mammals, is still remote. Thus, in an attempt to define functional genes more rapidly, particularly human ones, attention has turned to the sequencing of complementary DNA libraries[5,6]. In this situation, model organisms with small and compact genomes like bacteria[7-9], or with genomes of intermediate sizes such as *Caenorhabditis elegans*[10] or *Arabidopsis thaliana*[11], assume great importance as experimental subjects because their complete genomic sequences can be anticipated using current methodology. Of all model organisms, *Saccharomyces cerevisiae* occupies a unique position as it combines the advantages of being a eukaryote, being susceptible to powerful genetic techniques, and of having a genome size of only 13.5 megabases (Mb) (200 times smaller than that of the human genome). With yeast, the set of genes sufficient to build a simple eukaryotic cell can be deciphered with a relatively limited effort, and beyond its intrinsic biological significance can serve as a reference against which sequences of human, animal or plant genes may be compared.

Two years ago, a consortium of 35 European laboratories published the first complete sequence of a eukaryotic chromosome, that of chromosome III of *S. cerevisiae*[12]. The sequence revealed that almost half of the many new genes discovered had no homologue among previously described genes of either yeast or other organisms. But the experience with chromosome III was also important to help establish precise organizational rules for subsequent phases of the European Union genome-sequencing programmes[13]. During the past three years, our consortium has turned its efforts to the sequencing of yeast chromosomes II (820 kilobases (kb)) and XI. We report here the complete sequence of chromosome XI (666,448 base pairs (bp)), the second eukaryotic chromosome ever entirely sequenced. Beyond the many novel genes it contains, and the more precise description of the yeast genome composition it permits, the larger size of that chromosome, compared to chromosome III (315 kb), is sufficient to reveal new chromosomal organization patterns.

[1]Unité de Génétique Moléculaire des Levures (URA 1149 du CNRS and UFR927 University P.M. Curie), Département de Biologie Moléculaire, Institut Pasteur, F-75724 Paris CEDEX 15, France; [2]Foundation for Research and Technology-Hellas, Institute of Molecular Biology and Biotechnology, PO Box 1527, GR-Herakliion 711 10 Crete, Greece; [3]Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, Campus de la Plaine, CP244, Boulevard du Triomphe, B-1050, Bruxelles, Belgium; [4]European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, Germany; [5]Universidad de Salamanca, Departemento de Microbiologia y Genética, Av. del Campo Charro s/n, E-37007 Salamanca, Spain; [6]Universidad Autonoma de Madrid and Consejo Superior de Investigaciones Cientificas Facultad de Ciencias, Centro de Biologia Molecular, Cantoblanco, E-28049 Madrid, Spain; [7]Faculté des Sciences Agronomiques, Laboratoire de Microbiologie, 6, Avenue Maréchal Juin, B-5030 Gembloux, Belgium; [8]Laboratoire de Génétique Moléculaire, IGM (URA 1354 du CNRS), Université de Paris Sud, Batiment 400, F-91405 Orsay-Cedex, France; [9]Instituto Gulbenkian de Ciencia, Laboratorio de Genética Molecular, Rua da Quinta Grande 6, PL-2781 Oeiras, Portugal; [10]Institut Curie, Section de Biologie, Rue Georges Clémenceau, 15, Bâtiment 110, Centre Universitaire, F-91405 Orsay, France; [11]Laboratorium für Molekulare Biologie, D-82152 Martinsried, Germany; [12]Institute für Mikrobiologie und Molekularbiologie, Universität Giessen, Frankfurterstr. 107, D-35392 Giessen, Germany; [13]QUIAGEN GmbH, Max-Volmer Str. 4, D-40724 Hilden, Germany; [14]Institut für Mikrobiologie der Johann Wolfgang Goethe-Universität Frankfurt am Main, Marie Curie Str. 9, D-60439 Frankfurt-am-Main, Germany; [15]Institut für Physiologische Chemie, Physikalische Biochemie und Zellbiologie der Universität München, Goethestr. 33, D-80336 München, Germany; [16]MIPS, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany; [17]Institut für Mikrobiologie der Universität Düsseldorf Geb. 26.12, Universitätsr. 1, D-40225 Düsseldorf, Germany; [18]Biotechnologische und Molekularbiologische Forschung Angelhofweg 39, D-69259 Wilhelmsfeld, Germany; [19]Manchester Biotechnology Centre, UMIST, PO Box 88, Sackville street, Manchester M60 1QD, UK; [20]Unité de Biochemie Physiologique, Université Catholique de Louvain, Place Croix du Sud, 2-20, B-1348 Louvain-La-Neuve, Belgium; [21]Commission of the European Communities, B-1049 Brussels, Belgium; [22]Angewandte Tumorvirologie and Virologie appliquée à l'oncologie (Unité INSERM 375) Deutsches Krebsforschungszentrum, Abt.0610, P.101949, D-69009 Heidelberg, Germany; [23]Yeast Genetics, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, OX3 9DU, UK; [24]Service de Biochemie et de Génétique Moléculaire, Département de Biologie Cellulaire et Moléculaire, DSV/CEA, CE-Saclay, F-91191 Gif-sur-Yvette Cedex, France; [25]Institute für Angewandte Mikrobiologie, Biozentrum, Klingelbergstr. 70, CH-4056 Basel, Switzerland; [26]Department of Biochemistry and Molecular Biology, Institute for Molecular Biological Sciences, Vrije Universiteit, BioCentrum Amsterdam, de Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands; [27]Fakultät für Biologie der Universität Konstanz, Postfach 55 60, D-78434 Konstanz, Germany; [28]GATC-Gesellschaft für Analyse Technik und Consulting, Fritz-Arnold-Str. 23, D-78467 Konstanz, Germany; [29]Carlsberg Laboratory, Department of Physiology, Gamle Carlsberg Vej, 10, DK-2500 Copenhagen Valby, Denmark.

* Present addresses: Virus department, National Institute of Public Health, Vasi Ut174, Budapest 1393, Hungary (A.B.); Institut de Pédiatrie, BP24, F-13385 Marseille-CEDEX 5, France (L.C.); Unité des Virus Ongogènes, Institut Pasteur, Paris F-75724, France (M.G.); Institut Jacques Monod, Université Paris VII, 2 Place Jussieu, F-75005 Paris-CEDEX 5, France (M.H.); Murdoch Institute for Research into Birth Defects, Royal Children's Hospital, Flemington Road, Parkville, Vic 3052, Australia (O.H.); Unité d'Immunité Cellulaire Antivirale, Institut Pasteur, Paris F-75724, France (S.P.); Institute of Microbiology, Wroclaw University, Przybyszewskiego 63, 51–148 Wroclaw, Poland (J.S.); Unilever Research Laboratorium, Olivier ven Noortlaan 120, Vlaardingen NL-3130 AC, The Netherlands (J.C. v. V.-R.)


Left Tel
URA1
RSP1
TRP3
UBA1
STE6
YKL225w
YKL223w
YKL221w
YKL219w
YKL217w
YKL216w
YKL212w
YKL210w
YKL222c
YKL220c
YKL215c
YKL213c
YKL209c
YKL224c
YKL218c
YKL214c
YKL211c
ORD1
FAS1
PRPS
STE4
MRP49
TPK1
YKL185w
YKL184w
YKL183w
YKL181w
YKL180w
YKL177w
YKL175w
YKL172w
YKL170w
YKL182w
YKL179c
YKL176c
YKL174c
YKL173w
YKL171w
YKL168c
YKL187c
YKL186c
YKL178c
YKL169c
YKL167c
TGL1
CTK1
MRPL31
Lys1
PGM1
YPK1
VPH2
Ala1
APN1
YKL140w
YKL139w
YKL137w
YKL136w
YKL131w
YKL127w
YKL124w
YKL123w
YKL120w
YKL118w
YKL117w
YKL135c
YKL134c
YKL129c
YKL126w
YKL125w
YKL121w
YKL138c
YKL133c
YKL132c
YKL130c
YKL128c
YKL122c
YKL119c
YKL116c
YKL115c
YKL114c
YKL113c
CYT2
MDH1
VMA5
SMY1
Trp
NUP100
YKL090w
YKL089w
YKL088w
YKL086w
YKL085w
YKL084w
YKL083w
YKL081w
YKL080w
YKL077w
YKL079w
YKL078w
YKL073w
YKL072w
YKL071w
YKL070w
YKL069w
YKL068w
YKL087c
YKL082c
YKL076c
YKL075c
YKL074c
Val1a
His2
PTM1
Val1a
IXR1
URA6
GDC16
MAK11
RAM2
YKL041w
YKL039w
YKL037w
YKL035w
YKL031w
YKL030w
YKL028w
YKL027w
YKL023w
YKL019w
YKL018w
YKL038w
YKL034w
YKL033w
YKL032c
YKL029c
YKL025c
YKL022c
YKL020c
YKL040c
YKL036c
YKL026c
YKL024c
YKL021c
MRPL13
FOX2
Leu3
YPT52
GCN3
Arg2
YKR007w
YKR013w
YKR016w
YKR020w
YKR023w
YKR025w
YKR008w
YKR009c
YKR010c
YKR021w
YKR027w
YKR028w
YKR006c
YKR011c
YKR012c
YKR014c
YKR015c
YKR017c
YKR018c
YKR019c
YKR022c
YKR024c
YKR026c
DYN1
NUD1
RPS25
TIF1
KTR5
CCP1
Lys2
RHO4
YKR055w
YKR057w
YKR058w
YKR059w
YKR060w
YKR061w
YKR062w
YKR069w
YKR070w
YKR074w
YKR076w
YKR077w
YKR056w
YKR064w
YKR067w
YKR054c
YKR063c
YKR065c
YKR066c
YKR068c
YKR071c
YKR072c
YKR073c
YKR075c
PPC1
BAS1
SIR1
Right Tel
YKR097w
YKR104w
YKR099w
YKR101w
YKR102w
YKR103w
YKR106w
YKR098c
YKR100c
YKR105c



Chromosome XI map, deduced from complete sequence—see overleaf for details.

Assembly and verification of sequence

The sequence was determined from a set of 29 selected partially overlapping cosmid clones of a purpose-built genomic library (A.T. *et al.*, manuscript in preparation) from *S. cerevisiae* strain FY1679 (a direct derivative of S288C). Cosmids were distributed between the collaborating laboratories according to a scheme presented elsewhere (B.D. *et al.*, manuscript in preparation). Telomeres were physically mapped relative to the terminal-most cosmid inserts using I-*Sce*I chromosome fragmentation and then sequenced from plasmids rescued from integrations into the terminal $(C_{1\text{-}3}A)_n$ repeats of chromosome XI (note that the number of such repeats retained in the present sequence is arbitrary). Sequencing strategies and methods were left to the initiative of each laboratory and were diverse. Sequences were considered as final and entered into the MIPS data library when all bases had been unambiguously determined on both strands and strategies had been examined by the DNA coordinator. Sequences were then compared to the detailed restriction map constructed for this work, further examined for the possible occurrence of frameshifts, and if necessary returned to the sequencing laboratories for additional verification. Independent verification was also used to confirm critical or difficult regions and to estimate the overall sequence accuracy (Table 1).

TABLE 1 Quality control and results of estimated overall sequence accuracy

Method of verification	Total no. of fragments	Total bp verified	Error % detected
Original overlaps between cosmids	28	63,424	0.02
Resequencing of selected segments (3–5 kb long)	21	72,270	0.03
Resequencing of random segments (~300 bp long)	71	18,778	0.05
Resequencing of suspected segments (~300 bp long) from designed oligonucleotide pairs	60	17,035	0.03
Total	180	171,507	
Overall sequence accuracy, 99.97%			

All sequences submitted by collaborating laboratories to the MIPS data library were subjected to quality control. This included, initially, extensive re-examination of the restriction map, tracking down putative frameshifts, comparison with overlapping sequences from other laboratories and, eventually, the resequencing of selected or random segments on an anonymous basis. Selected segments were either 3–5-kb fragments that were entirely resequenced using the same methods, strategies and criteria as the original sequences, or short segments (~300 bp) chosen from suspected or difficult zones that were re-sequenced directly from cosmids using designated pairs of oligonucleotides as primers. Random segments were inserts of ~500 bp issued from shotgun cloning of the entire chromosome in the pBluescript SK+ vector. The overall sequence accuracy has been estimated from extrapolation of the number of actual errors found in the original sequences after re-examination of each divergence with the verification sequences.

Definition of ORFs and other elements

A total of 331 open reading frames (ORFs) were identified in the entire chromosome using principles explained in Fig. 1 legend. Seven of these ORFs are interrupted by introns. This list includes 43 partially overlapping ORFs (20 pairs and 1 triplet), all but four representing antiparallel overlaps. Six pairs each include a gene whose function is known, whereas nine other pairs and the triplet each include an ORF whose predicted product has a homologue in the databases, suggesting that it corresponds to a real gene (for details, see B.D. *et al.*, manuscript in preparation). In all such cases, the partially overlapping partner ORF is shorter, suggesting that it may not correspond to a real gene. The reality of ORFs as functional genes has been systematically examined using, as criteria, their codon adaptation index (CAI; ref. 14) in conjunction with their size (we have verified that there is no general correlation between CAI and ORF size; data not shown). Although there exist functionally defined genes with CAI <0.110, the average CAI of the subset of 93 chromosome XI genes with known functions (Fig. 1, and see below) is 0.211 ($\sigma = 0.148$, range 0.101–0.868), significantly higher than that of the entire set of 331 ORFs (average 0.170; $\sigma = 0.112$, range 0.045–0.868). To test the possibility that some predicted ORFs may occur by chance, a random sequence of the same size and composition as chromosome XI (both mononucleotide and dinucleotide frequencies are respected) has been generated. This sequence shows 37 ORFs longer than 100 codons (all are shorter than 150 codons; average is 115, with $\sigma = 11.7$). Most of them have a low CAI (average 0.107; $\sigma = 0.023$, range 0.053–0.168). For this reason, ORFs that are both shorter than 150 codons and have CAI < 0.110 are considered as 'questionable' (Fig. 1). Consistent with this set containing many unreal genes, 11 out of the 23 questionable ORFs are partially overlapping, whereas only four of them have homologues (compared to 67% for all ORFs). The reality of ORFs as functional genes may also be challenged by the possible existence of pseudogenes. Although very few are known in the yeast genome, this might be the case for YKR103w and YKR104w, which are in the same frame but separated by a single stop codon (verified by sequencing yeast genomic DNA itself) and whose expected products are homologous to the N- and C-terminal parts of the same protein, respectively. From these considerations, we estimate the total number of chromosome XI ORFs corresponding to real genes to be ~310. To this number could be added a few ORFs, not shown in Fig. 1, that are shorter than 100 codons but have a very high CAI (B.D. *et al.*, manuscript in preparation).

A total of 16 transfer RNA genes, three of them containing short introns, and eleven δ and one σ sequences have also been

FIG. 1 The map on pages 372–373 represents chromosome XI of S. *cerevisiae* as deduced from the complete sequence. The map is drawn to scale from the sequence (coordinates are in kb). The two DNA strands are materialized as colour gradient bars to represent compositional variations (Fig. 3). The top strand (designated 'Watson' strand) is oriented 5′ to 3′ from left to right. The sequence has been interpreted using the following principles: (1) all intron splice-site/branch-point pairs were listed; (2) all ORFs containing at least 100 contiguous sense codons (including the first ATG) and not entirely contained within a longer ORF on either DNA strand were listed (this includes partially overlapping ORFs); (3) the two lists were merged and all splice-site/branch-point pairs occurring inside an ORF but in opposite orientation were disregarded. The possible occurrence of reading frames flanking putative splice sites was then examined for all remaining pairs (namely those occurring inside ORFs in direct orientation or in inter-ORF regions). In this case, no lower size limit applied, but the existence (after splicing) of an ATG codon in-frame with the stop codon was essential. Finally, tRNA gene (red boxes), centromere, telomeres (grey boxes), δ, σ and τ elements (green boxes) were sought by comparison with a previously characterized dataset of such elements. (H. Feldmann, unpublished results). This procedure identified 331 ORFs (blue boxes), which have been numbered in increasing order from the centromere and designated YKL for the left arm and YKR for the right arm (w/c suffix indicates the Watson/Crick coding strand). Twenty-two ORFs shorter than 150 codons and having a CAI <0.110 were considered as questionable (light-blue boxes). The same has been applied to YKL118w (CAI 0.136) as it overlaps a δ sequence. Predicted ORF translation products have been compared to sequence data libraries using the FastA, and BlastX algorithms as well as protein pattern search regimes: 93 ORFs correspond to genes whose functions have been identified (brown boxes above the chromosome map); 93 other ORFs have homologues of known function (for details, see B.D. *et al.*, manuscript in preparation). Parts of this sequence have been published[17,29–52] during the course of this programme (note that there are some differences as the published sequences had not been submitted to final quality controls of the entire chromosome).

identified (Fig. 1). All δ sequences occur less than 300 nucleotides upstream of a tRNA gene (or less than 300 nucleotides upstream of the first δ in cases of double δs). The σ sequence partially overlaps the 5′ end of the tRNA Lys1 gene. These elements represent the long terminal repeats of yeast retroposons (Ty) but, in contrast to chromosome III (ref. 15), no complete Ty was found.

Analysis of predicted protein products

Comparison of the present sequence with public databases revealed that 93 of the 331 ORFs (28%) correspond either to previously known protein-encoding genes or to genes whose functions have been determined during this work (Fig. 1 and B.D. *et al.*, manuscript in preparation). All other ORFs, 72% of the total, represent novel putative yeast genes whose functions need to be experimentally determined. But 93 of them (another 28% of the total) have homologues among gene products from yeast or other organisms whose functions are known, whereas 37 others (11% of the total) have homologues that are themselves of unknown function. The remaining 108 ORFs (33% of the total) either have no homologues in data libraries or show levels of similarity of uncertain significance (note that this last set includes 18 questionable ORFs). Overall, about 40–44% of the genes of chromosome XI are thus of unpredicted function, a figure similar to that of chromosome III (ref. 16).

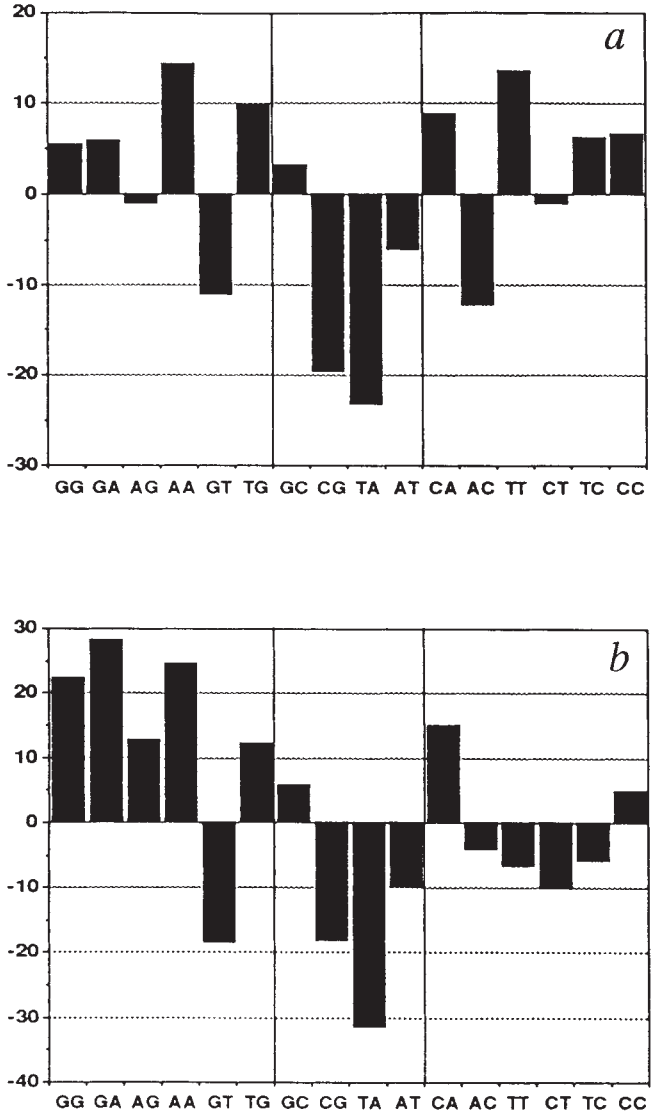


FIG. 2 Compositional symmetry/asymmetry of the chromosome and of its constitutive elements. Relative deviations of dinucleotide frequencies ((observed minus expected)/expected) are shown as vertical bars (expected frequencies are calculated from mononucleotide frequencies). Complementary dinucleotide pairs have been arranged in mirror image to help visualize compositional symmetry or asymmetry. Self-complementary dinucleotides are at the centre. *a*, Data for the entire chromosome, calculated from the Watson strand; *b*, data for ORFs only, calculated from the coding strand.

Organization of the chromosome

The very high gene density previously found with chromosome III is confirmed: ORFs occupy on average 72% of the sequence of chromosome XI. The average ORF size is 488 codons (1,464 bp): the longest ORF is YKR054c (4,092 codons), which encodes dynein[17,18], and the second longest is YKL203c (2,473 codons), which is the *TOR2* gene[19]. Only 25 other ORFs are more than 1,000 codons long. The mean sizes of inter-ORF distances are 804 bp for 'divergent promoters' and only 381 bp for 'convergent terminators', with 'promoter–terminator combinations' being of intermediate length (orientation of ORFs, and hence that of transcription units, is statistically random and approximately equal numbers occur on each DNA strand). There are only six inter-ORF regions that are longer than 3,000 bp; five of these are located close to the telomeres, the sixth is found half-way along the left arm and contains a tRNA gene.

The size of chromosome XI (~5% of the total yeast genome) is sufficient to examine general features of chromosome organization which, combined with previous chromosome III data, should be informative about the yeast genome in general. The average base composition is 38.1% G+C (38.5% for chromosome III), with significant variations between ORFs (average of 40.2%), 'divergent promoters' (36.3%) and 'convergent terminators' (29.8%). Again, 'promoter–terminator' combinations are of intermediate value. The corresponding value for introns is 32.3%. Average base composition is symmetrical over the entire chromosome (the symmetry being even more apparent with dinucleotide frequencies; Fig. 2*a*), but this only reflects the almost equal numbers of ORFs encoded on each DNA strand, the base composition of ORFs themselves showing a significant excess of homopurine pairs on the coding strand (Fig. 2*b*). A regional variation of base composition was noted along chromosome III, with a major (G+C)-rich peak in each arm[20,21]. Variations of similar amplitudes are found here but, owing to the larger size of chromosome XI, a much more orderly pattern appears, with an almost regular periodicity of the G+C content (Fig. 3). Four major (G+C)-rich peaks occur on the left arm, separated from each other by ~90–100 kb, and one major peak (plus a minor one) occurs on the right arm (which is shorter). Re-examination of chromosome III indicates that spacing is similar. Thus a yeast chromosome appears as a succession of (G+C)-rich and (G+C)-poor segments of ~50 kb each (see Fig. 1 for an overall view of the phenomenon). Interestingly, the compositional periodicity correlates with local gene density (Fig. 4), as is the case in more complex genomes in which isochores of composition are, however, much larger[22].

The present sequence also reveals interchromosomal subtelomeric duplications. For example, two left subtelomeric segments of the present sequence (coordinates 348–1,005 and 1,666–2,850) are almost identical with two segments of the right subtelomeric region of chromosome III. Conversely, the right subtelomeric part of the present sequence (from 658 kb to the end) matches almost perfectly the left subtelomeric region of chromosome III over more than 8 kb, the only exception being that a Ty5-1 element is inserted into chromosome III (ref. 23). Such duplications include three ORFs on the left arm and two ORFs on the right arm (Fig. 1). Subtelomeric regions of chromosome XI do not contain Y′ elements (although, intriguingly, the 36-bp repeat internal to Y′ ORF2 sequences[24] is found in a short

ORF located in the middle of the left arm; B.D. *et al.*, manuscript in preparation).

Physical and genetic maps compared

The last edition of the genetic map of *S. cerevisiae*[25] assigned 50 genes or markers to chromosome XI, of which 43 were positioned on a single linear array and seven remained unmapped. Comparison of this map with that deduced from the complete sequence is shown in Fig. 4, which reveals major discrepancies in the gene order. The entire *URA1–STE6* segment, located next to the left telomere (positions ~25–46 kb of the present sequence), was translocated and inverted in the genetic map. *URA6*, which lies in the left arm, was erroneously mapped to the right arm, and the entire centromeric region, including the 3 centromere-linked genes *MET14*, *VPS1* and *PAP1*, was inverted. *GAP1*, which lies between *DAL80* and *TRK2*, was mapped distal to *TIF1*. Finally, a small inversion can also be noted between the two closely linked genes *CDC16* and *MAK11*.

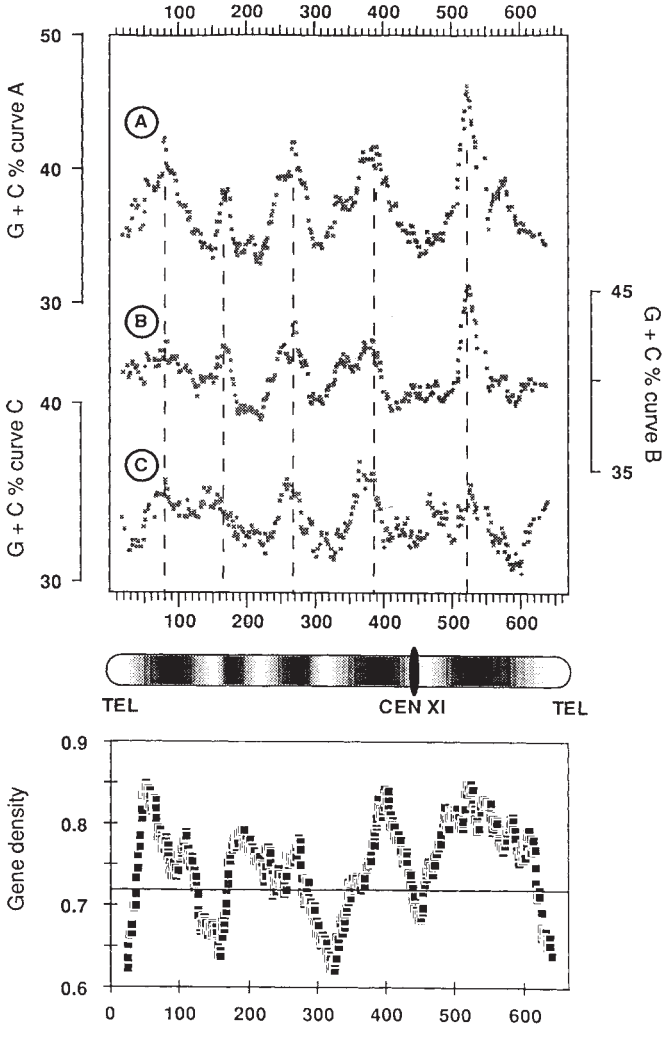


FIG. 3 Compositional variation and gene density distribution along chromosome XI. Top, Compositional variation along chromosome XI calculated as in ref. 20. Each point represents the average G + C composition of 15 consecutive elements but similar results were obtained for averages of 10–30 elements (not shown). Curve A: G + C composition calculated from the silent positions of codons only; curve B: G + C composition calculated from entire ORFs; curve C: G + C composition calculated from inter-ORF regions. Centre, Regional compositional variation along chromosome XI is shown as graduated shading: (white, G + C-poor; dark, G + C-rich). Bottom, Gene density along chromosome XI. Gene density is expressed as the fraction of nucleotides within ORFs versus the total number of nucleotides in sliding windows of 50 kb (steps are 1 kb). Similar results were obtained for sliding windows of 30–70 kb (not shown). Horizontal line represents average gene density for the entire chromosome (0.72).

Discussion

Our 'network' approach to systematic genome sequencing started with chromosome III (refs 12, 13, 26) and has been improved here by the construction, before sequencing, of a high-quality cosmid library and a fine-resolution physical map of the chromosome and by the implementation of novel quality controls. The important discrepancies between the physical and genetic maps were a surprise. The physical map has been constructed without any reference to the genetic map from a complete set of overlapping cosmid clones derived from a unique strain, which have been individually verified for their colinearity with genomic yeast DNA. It is entirely consistent with direct physical measurements of the yeast chromosome after chromosome fragmentation using I-*Sce*I (ref. 27) and was eventually confirmed by the final sequence. The genetic map, on the contrary, is a compromise between results obtained by several laboratories working independently and on different strains. It is possible that some strains, used to establish particular linkages, showed inversions or translocations with respect to the strain used for sequencing. But, in any case, this experience with yeast, where genetics are precise and easy, demonstrates the need for independent physical and genetic mapping data for all genome projects.

The rationale behind our effort on quality control was that if systematic genome sequencing of yeast is to be useful and significant, then sequence accuracy must permit correct interpretations. With the gene density and ORF size distribution of yeast, even relatively rare sequencing errors (many of which are missing nucleotides) result in a large fraction of the protein-coding genes being affected by frameshifts. A simple calculation predicts that with an accuracy of 99%, virtually all predicted genes contain errors, and that at 99.9% accuracy (a figure commonly obtained with good sequencing practices and regarded as satisfactory), two-thirds of the genes still contain at least one error (most often a frameshift). A higher level of accuracy (99.97%) has been achieved in this project only after costly effort; this figure is unlikely to be bettered with current technology, but still implies that about one-third of predicted genes will contain sequencing errors that will affect their interpretation.

In this work, the present sequence has been determined in its entirety, irrespective of the existence of previous sequence data, making *a posteriori* comparisons meaningful. From 107 different entries, totalling 288,252 bp, which correspond to parts of the present sequence but were determined independently by others (EMBL data library, release 73), we found only 26 (24%) of them to be identical to the present sequence. The others (76%) show divergences that range from 0.01% to >4%; the average being 0.3%, or 10 times the error-rate estimated for this study (we have excluded from our calculation those errors found at the extremities of published sequences which correspond to vector contaminations or weaker gel readings). Divergences with published sequences are of several types, but only some of them can plausibly be explained by strain differences. In the cases of the partial sequences published independently by the authors of the present work (a total of 23 entries totalling 309,467 bp; see Fig. 1 legend for refs), only 6 base substitutions and 13 additions/deletions exist compared to the present sequence, consistent with our estimated error rate of 0.03% (remember that these sequences had not yet been submitted to independent verification when published).

Much is still to be learned from large-scale sequencing programmes regarding genome organization and evolution, but the almost periodic variation of base composition and gene density found for chromosome XI is a strong indication that there exist rules in a eukaryotic chromosome that individual genes must

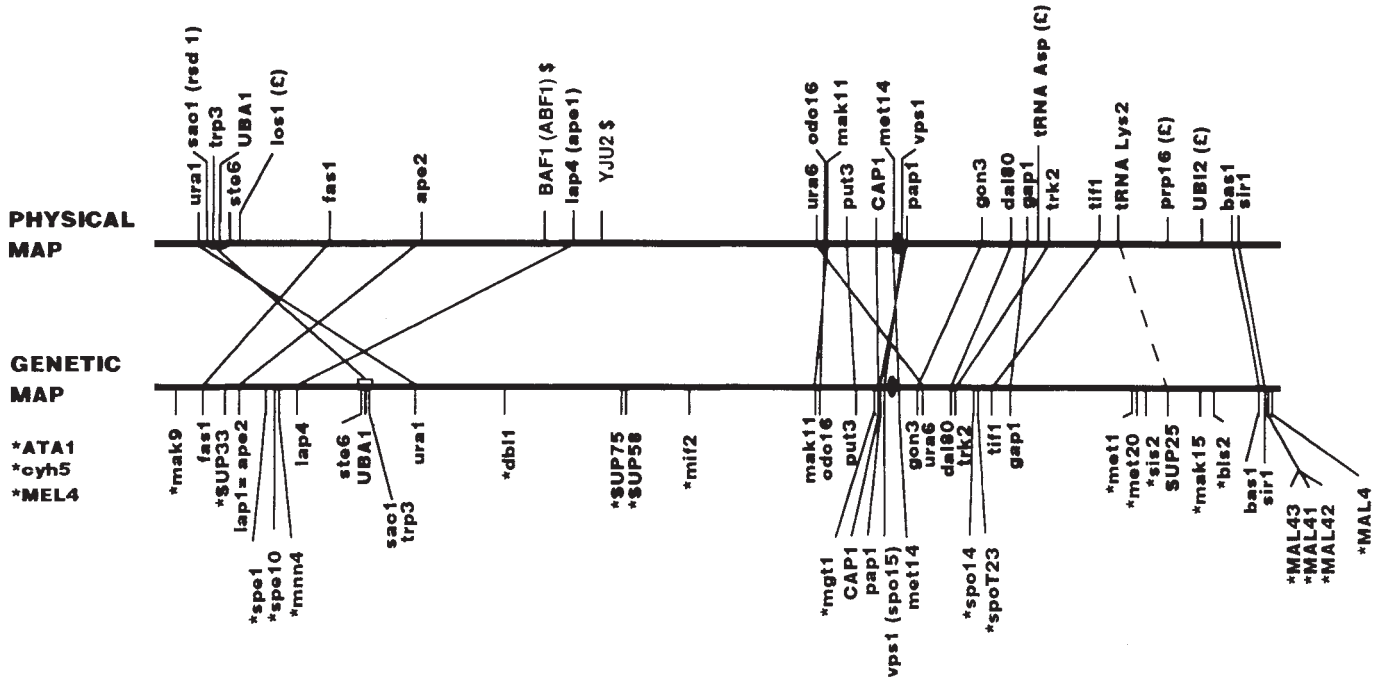


FIG. 4 Comparison of the genetic and physical maps of chromosome XI. The genetic map is redrawn from ref. 25. The physical map deduced from this work has been drawn to the same scale. Filled oval represents the centromere. The genetic map assigned 50 genes or markers to chromosome XI, of which 43 were positioned on a single linear array and 7 remained unmapped. Twenty-four of the 43 mapped genes and 4 of the 7 unmapped genes (*LOS1, PRP16, UB12* and tRNA Asp) could be unambiguously assigned to an ORF or a tRNA gene of the present sequence on the basis of previous partial sequence data, use of probes or gene function (the remaining 22 genes or markers are indicated by an asterisk). The two genes *BAF1 (ABF1)* and *YJU2* indicated by a dollar sign have previously been erroneously mapped to chromosomes V and X, respectively, on the basis of imprecise electrophoretic karyotype analysis[53,54]. Many other genes described here were not previously assigned to a chromosome (compare Figs 2 and 4). The assignation of tRNA Lys2 to SUP25 is only tentative.

obey, and which are location-dependent. Such rules may influence the expression of genes, their stability towards mutagenic forces or their recombination properties. The reason for the fairly regular succession of (G+C)-rich and (G+C)-poor segments along yeast chromosomes is unclear. A correlation with the location of replication origins or their order of firing during S phase might be imagined. However, the map of functional ARS elements for chromosome III (ref. 28) does not easily fit the distribution of its compositional peaks (functional ARS elements have yet to be defined for chromosome XI). In addition, many ARS are dispensable. Alternatively, the compositional periodicity of a yeast chromosome could reflect the folding of that chromosome, its attachment to the nuclear matrix or structural elements involved in chromosome segregation, or in the 'homology search' that precedes synapsis in the early meiotic prophase.

Insight into eukaryotic genome organization and evolution can also be gained from the degree of internal genetic redundancy. Determination of this value is, however, very imprecise at this stage of the yeast sequencing programme. If we extrapolate to the entire yeast genome, the fact that 15 of the chromosome XI ORFs of unknown function have homologues among ORFs also of unknown function and lying on other systematically sequenced chromosomes or on chromosome XI itself, we obtain the surprising figure that genetic redundancy in the yeast genome must be high, at least among the genes of unknown function.

The availability of the complete sequences of yeast chromosomes III and XI offered a chance to search for interesting or unexpected genes. Opinions may vary as to what constitutes an interesting gene, but among those are certainly the homologues to genes that perform differentiated functions in multicellular organisms (such as the *Drosophila* white-pigment gene on chromosome III; ref. 12), or are involved in human pathologies such as xeroderma pigmentosum (YKL113c) or adrenoleukodistrophy (YKL188c) on chromosome XI. The existence of such genes in yeast, where their role remains to be clarified, may offer a powerful experimental system to identify their function or to develop new therapeutic agents. About a quarter of all predicted ORF products of chromosome XI have homologues in cDNA databases (including the human expressed sequence tags sequences), enabling genes common to all eukaryotes to be directly identified. Interestingly, this fraction is much higher (42%) among the 186 ORFs of known or predictable functions and much lower (<7%) among the remaining 145 ORFs of unknown functions. It looks as if many of these novel functions only required transient or low-level transcription in complex organisms or were primarily phylum-specific. Such questions may quickly be answered as systematic sequencing of the yeast genome progresses (chromosomes I, II and VI are nearly completed and all others are being sequenced at present), and new routes are explored to interpret the wealth of information. □

Received 1 March; accepted 27 April 1994.

1. Smith, L. M. *et al. Nature* **321,** 674–679 (1986).
2. Ansorge, W., Sproat, B. S., Stegemann, J. & Schwager, C. J. *Biochem. biophys. Meth.* **13,** 315–323 (1986).
3. Pohl, F. M. & Beck, S. *Meth. Enzym.* **155,** 250–259 (1987).
4. Church, G. M. & Kieffer-Higgins, S. *Science* **240,** 185–188 (1988).
5. McCombie, W. R. *et al. Nature Genet.* **1,** 124–131 (1992).
6. Okubo, K. *et al. Nature Genet.* **2,** 173–179 (1992).
7. Danieis, D. L., Plunkett, G., Burland, V. & Blattner, F. R. *Science* **257,** 771–778 (1992).
8. Kunst, F. & Devine, K. *Res. Microbiol.* **142,** 905–912 (1991).
9. Honore, N. *et al. Molec. Microbiol.* **7,** 207–214 (1993).
10. Wilson, R. *et al. Nature* **368,** 32–38 (1994).
11. Meyerowitz, E. M. & Pruitt, R. E. *Science* **229,** 1214–1218 (1985).
12. Oliver, S. G. *et al. Nature* **357,** 38–46 (1992).
13. Vassarotti, A., Dujon, B., Feldmann, H., Mewes, H. W. & Goffeau, A. *J. Biotech.* (in the press).
14. Sharp, P. M. & Li, W. H. *Nucleic Acids Res.* **15,** 1281–1295 (1987).
15. Wicksteed, B. L. *et al. Yeast* **10,** 39–57 (1994).
16. Koonin, E. V., Bork, P. & Sander, C. *EMBO J.* **13,** 493–503 (1994).
17. Eshel, D. *et al. Proc. natn. Acad. Sci. U.S.A.* **90,** 11172–11176 (1993).
18. Li, Y. Y., Yeh, E., Haus, T. & Bioom, K. *Proc. natn. Acad. Sci. U.S.A.* **90,** 10096–10100 (1993).
19. Kunz, J. *et al. Cell* **73,** 585–596 (1993).
20. Sharp, P. & Lloyd, A. *Nucleic Acids Res.* **21,** 179–183 (1993).
21. Karlin, S. *et al. Nucleic Acids Res.* **21,** 703–711 (1993).

22. Bernardi, G. *Gene* **135,** 57–66 (1993).
23. Voytas, D. F. & Boeke, J. D. *Nature* **358,** 717 (1992).
24. Louis, E. J. & Haber, J. E. *Genetics* **133,** 559–574 (1992).
25. Mortimer, R. K., Contopoulou, R. & King, J. S. *Yeast* **8,** 817–902 (1992).
26. Vassarotti, A. & Goffeau, A. *Trends Biotech.* **10,** 15–18 (1992).
27. Thierry, A. & Dujon, B. *Nucleic Acids Res.* **20,** 5625–5631 (1992).
28. Dershowitz, A. & Newlon, C. S. *Molec. cell. Biol.* **13,** 391–398 (1993).
29. Alexandraki, D. & Tzermia, M. *Yeast* (in the press).
30. Bossier, P., Fernandes, P., Villela, L. & Rodrigues-Pousada, C. *Yeast* (in the press).
31. Bou, G. *et al. Yeast* **9,** 1349–1354 (1993).
32. Boyer, J., Pascolo, S., Richard, G-F. & Dujon, B. *Yeast* **9,** 279–287 (1993).
33. Chéret, G. C. *et al. Yeast* **9,** 1259–1265 (1993).
34. Colleaux, L., Richard, G. F., Thierry, A. & Dujon, B. *Yeast* **8,** 325–336 (1992).
35. Düsterhöft, A. & Philippsen, P. *Yeast* **8,** 749–759 (1992).
36. Garcia-Cantalejo, J. *et al. Yeast* **10,** 221–245 (1994).
37. Jacquier, A., Legrain, P. & Dujon, B. *Yeast* **8,** 121–132 (1992).
38. James, C. M., Gent, M. E. & Oliver, S. G. *Yeast* **10,** 257–264 (1994).
39. James, C. M., Gent, M. E., Indge, K. J. & Oliver, S. G. *Yeast* **10,** 247–255 (1994).
40. Pallier, C. *et al. Yeast* **9,** 1149–1155 (1993).
41. Pascolo, S. *et al. Yeast* **8,** 987–995 (1992).
42. Purnelle, B., Skala, J., Van Dyck, L. & Goffeau, A. *Yeast* **8,** 977–986 (1992).
43. Purnelle, B., Skala, J., Van Dyck, L. & Goffeau, A. *Yeast* **10,** 125–130 (1994).
44. Purnelle, B. *et al. Yeast* **9,** 1379–1384 (1993).
45. Rasmussen, S. W. *Yeast* (in the press).
46. Rasmussen, S. W. *Yeast* (in the press).
47. Singer-Krüger, B. *et al. J. Cell Biol.* (in the press).
48. Tzermia, M., Horaitis, O. & Alexandraki, D. *Yeast* (in the press).
49. van Vliet-Reedijk, J. C. & Planta, R. J. *Yeast* **9,** 1139–1147 (1993).
50. Vandenbol, M. *et al. Yeast* (in the press).
51. Vandenbol, M. *et al. Yeast* (in the press).
52. Wiemann, S. *et al. Yeast* **9,** 1343–1348 (1993).
53. Rhode, P. R., Sweder, K. S., Oegema, K. F. & Campbell, J. *Genes Dev.* **3,** 1926–1939 (1989).
54. Forrova, H., Kolarov, J., Ghislain, M. & Goffeau, A. *Yeast* **8,** 419–422 (1992).

ACKNOWLEDGEMENTS. The Laboratory Consortium operating under contracts with the European Commission was initiated and organized by A. Goffeau. This study is part of the second 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, and under the scientific responsibility of B. Dujon, as DNA coordinator, and H. W. Mewes, as Informatics coordinator. We thank P. Mordant for accounting; F. Winston for the yeast strains FY23 and FY73; F. Foury and G. Thireos for advice and administration; P. Jordan for computing at MIPS; E. Sonnhammer for preparation of AScDB database files; M. Rambaud for administrative assistance; and our colleagues for help and for discussion. This work was supported by the EC under the BRIDGE Programme and by the Région Wallonne, the Fond National de la Recherche Scientifique and La région de Bruxelles Capitale; The Bundesminister für Forschung und Technologie and the Fonds der Chemischen Industrie; the Comision Interministerial de Ciencia y Tecnologia and the Fundacion Ramon Areces; the Ministère de l'Education Nationale and the Ministère de la Recherche et de l'Espace, Institut Pasteur and Institut Curie; the Greek Ministry of Industry, Energy and Technology; the Fundaçao Calouste Gulbenkian and the Junta Nacional de Investigaçao Cientifica e Technologica; and The Wellcome Trust. Nucleic acid and protein sequences including annotations are available through anonymous ftp retrieval in different standard database formats and in a form accessible to the AscDB database program from the following computer nodes: ehpmic.mips.biochem.mpg.de; ftp.embl-heidelberg.de; ftp.pasteur.fr; ftp.sanger.ac.uk; genome-ftp.stanford.edu. A CD-ROM containing a database of yeast sequences including the complete chromosomes III and XI sequences will be available shortly from MIPS.

LETTERS TO NATURE

Is the ring around SN1987A a protostellar disk?

Richard McCray* & Douglas N. C. Lin†

* Joint Institute for Laboratory Astrophysics, University of Colorado and National Institute of Standards and Technology, Boulder, Colorado 80309-0440, USA
† Lick Observatory, University of California at Santa Cruz, Santa Cruz, California 95064, USA

ACCORDING to conventional wisdom[1–4], the ring around supernova 1987A is a product of winds from the progenitor star, which should have produced a thin, dense, spherical shell[1]. It was accordingly a surprise when images obtained by the Hubble Space Telescope[5–8] revealed that the gas is in fact disposed in a thin ring, with a radial velocity[9] much smaller than that predicted by theory. This could be explained by an asymmetry in the red giant wind[10–12], or by rotational flattening[13,14], but these explanations seem to us to be *ad hoc*, and have associated problems. Here we propose, instead, that the ring is the inner rim of a disk of gas that is left over from the time of formation of the progenitor star. The centre of the disk was evaporated by the ionizing radiation of the progenitor over its lifetime of about ten million years, leaving a ring-like structure. Our hypothesis naturally explains the ring's physical properties, and leads to the prediction that we should see the rest of the disk shortly after the supernova ejecta hit the ring in AD 1999 ± 3.

During its red giant stage, the progenitor of SN1987A should have expelled relatively dense gas in a low velocity ($v_R \approx 20$ km s^{-1}) stellar wind. Then, several thousand years ago, the progenitor became a blue giant and the high-pressure bubble due to its fast ($v_B \approx 10^3$ km s^{-1}) stellar wind began to pack the inner part of the relic red giant wind into a thin, dense spherical shell[1]. Therefore, when evidence for circumstellar gas near SN1987A appeared in the form of narrow ultraviolet and optical emission lines[2,3], it was natural to assume that the line-emitting gas must have been expelled by the supernova progenitor and was most likely the spherical shell resulting from the interacting winds. The overabundance of nitrogen inferred from the ultraviolet emission-line strengths was persuasive evidence in favour of the ejection hypothesis[2,4]. Other observations pose problems for this simple picture, however. In addition to the images[5–8] showing that the gas is in a thin ellipse (apparently the projection of a circular ring inclined at ~45°), the radial velocity of the ring inferred from the velocity gradient along the minor axis[9] is ~10 km s^{-1}, much less than the expansion velocity ~50 km s^{-1} predicted by the interacting winds model.

In an attempt to explain why the circumstellar gas resides in a ring, several authors[10–12] proposed a modified model in which the red giant progenitor ejected its envelope in a wind that was much denser at the equator than at the poles. Then, when the progenitor became a blue giant, the pressure due to the blue giant wind would have moulded the inner boundary of the red giant wind into an hourglass-shaped shell. According to this interpretation, the ring is the waist of the hourglass and the outer bipolar nebulosity consists of those parts of the lobes of the hourglass that have been illuminated by the ionizing flash of the supernova.

There are reasons to be dissatisfied with this model, however. Numerical simulations[12] show that large flux anisotropy in the red giant wind is needed to produce a significantly flattened structure. But even with an equator/pole density ratio ~(5–20)/1, the resulting ring shape appears to be thicker than observed (thickness/radius ratio ≈0.1). A thinner ring is possible, but requires even greater anisotropy in the red giant wind. There is no dynamical account for such a high degree of anisotropy. Furthermore, because the red giant wind is assumed to be accelerated by the blue giant wind, the expansion velocity should be greater than that of the red giant wind. Yet the observed radial velocity (10 km s^{-1}) is significantly less than the expansion velocity (~15–20 km s^{-1}) of most red supergiant winds[15]. To account for the low radial velocity of the ring, the numerical simulations require a red giant wind with an extremely high equatorial mass loss rate and a very low (~5 km s^{-1}) radial velocity.

It has been suggested[13,14] that the small aspect ratio of the ring may be caused by rotational flattening if the ejection is accelerated by a hypothetical companion[16]. If the red giant wind has a velocity of ~10 km s^{-1}, only companions with binary separation comparable to, or smaller than, a few astronomical units, can provide sufficient acceleration. But in this case, the maximum specific angular momentum (per unit mass) that can be attained by the wind would fall short of the value required for rotational flattening at ~1 light yr by two orders of magnitude.

Indeed, we are so uncomfortable with the ejection hypothesis that we think it worthwhile to advance an entirely different