The complete DNA sequence of yeast chromosome III

S. G. Oliver[1], Q. J. M. van der Aart[2], M. L. Agostoni-Carbone[3], M. Aigle[4], L. Alberghina[5], D. Alexandraki[6], G. Antoine[7], R. Anwar[1], J. P. G. Ballesta[8], P. Benit[7], G. Berben[9], E. Bergantino[10], N. Biteau[4], P. A. Bolle[9], M. Bolotin-Fukuhara[11], A. Brown[1], A. J. P. Brown[12], J. M. Buhler[13], C. Carcano[3], G. Carignani[10], H. Cederberg[14], R. Chanet[7], R. Contreras[15], M. Crouzet[4], B. Daignan-Fornier[11], E. Defoor[16], M. Delgado[17], J. Demolder[15], C. Doira[11], E. Dubois[18], B. Dujon[19], A. Dusterhoft[20], D. Erdmann[20], M. Esteban[17], F. Fabre[7], C. Fairhead[19], G. Faye[7], H. Feldmann[21], W. Fiers[15], M. C. Francingues-Gaillard[11], L. Franco[8], L. Frontali[22], H. Fukuhara[7], L. J. Fuller[23], P. Galland[6], M. E. Gent[1], D. Gigot[18], V. Gilliquet[9], N. Glansdorff[18], A. Goffeau[24,37], M. Grenson[25], P. Grisanti[22], L. A. Grivell[26], M. de Haan[26], M. Haasemann[27], D. Hatat[28], J. Hoenicka[8], J. Hegemann[20], C. J. Herbert[29], F. Hilger[9], S. Hohmann[14], C. P. Hollenberg[30], K. Huse[14], F. Iborra[11], K. J. Indge[1], K. Isono[31], C. Jacq[28], M. Jacquet[11], C. M. James[1], J. C. Jauniaux[25], Y. Jia[29], A. Jimenez[8], A. Kelly[32], U. Kleinhans[30], P. Kreisl[27], G. Lanfranchi[10], C. Lewis[23], C. G. van der Linden[33], G. Lucchini[3], K. Lutzenkirchen[30], M. J. Maat[26], L. Mallet[11], G. Mannhaupt[21], E. Martegani[5], A. Mathieu[7], C. T. C. Maurer[33], D. McConnell[32], R. A. McKee[23], F. Messenguy[18], H. W. Mewes[27], F. Molemans[15], M. A. Montague[32], M. Muzi Falconi[3], L. Navas[17], C. S. Newlon[34], D. Noone[32], C. Pallier[11], L. Panzeri[3], B. M. Pearson[23], J. Perea[28], P. Philippsen[20], A. Pierard[18], R. J. Planta[33], P. Plevani[3], B. Poetsch[27], F. Pohl[35], B. Purnelle[24], M. Ramezani Rad[30], S. W. Rasmussen[36], A. Raynal[11], M. Remacha[8], P. Richterich[35], A. B. Roberts[12], F. Rodriguez[5], E. Sanz[8], I. Schaaff-Gerstenschlager[14], B. Scherens[18], B. Schweitzer[20], Y. Shu[28], J. Skala[24], P. P. Slonimski[29], F. Sor[7], C. Soustelle[11], R. Spiegelberg[20], L. I. Stateva[1], H. Y. Steensma[2], S. Steiner[20], A. Thierry[19], G. Thireos[6], M. Tzermia[6], L. A. Urrestarazu[25], G. Valle[10], I. Vetter[21], J. C. van Vliet-Reedijk[33], M. Voet[16], G. Volckaert[16], P. Vreken[33], H. Wang[32], J. R. Warmington[1], D. von Wettstein[36], B. L. Wicksteed[12], C. Wilson[22], H. Wurst[35], G. Xu[30], A. Yoshikawa[31], F. K. Zimmermann[14] & J. G. Sgouros[27]

The entire DNA sequence of chromosome III of the yeast *Saccharomyces cerevisiae* has been determined. This is the first complete sequence analysis of an entire chromosome from any organism. The 315-kilobase sequence reveals 182 open reading frames for proteins longer than 100 amino acids, of which 37 correspond to known genes and 29 more show some similarity to sequences in databases. Of 55 new open reading frames analysed by gene disruption, three are essential genes; of 42 non-essential genes that were tested, 14 show some discernible effect on phenotype and the remaining 28 have no overt function.

THE sequence of the DNA molecule of chromosome III from the budding yeast *S. cerevisiae* has been completed by a consortium of 35 European laboratories within the framework of the European Community's Biotechnology Action Programme. The sequence is 315 kilobases long and contains 182 open reading-frames (ORFs) encoding putative proteins of $\geq$100 amino acids. So far, 55 novel ORFs have been subjected to functional analysis by gene disruption. In addition to the putative protein-encoding genes there are 10 transfer RNA genes, of which four had previously been defined by suppressor mutations. Regions of sequence variation between chromosomes III of different strains of *S. cerevisiae* have been identified, as have the differences between the physical and genetic maps. These data indicate that systematic genome sequencing projects can reveal new functions that have been missed by more traditional approaches and also illuminate the mechanisms of genome evolution. Most importantly, they reveal that our knowledge of molecular genetics is far from complete and that we are ignorant about the biological function of the majority of genes in a eukaryotic genome as small and well-defined as that of yeast.

The bakers' yeast *S. cerevisiae* is one of the most important experimental organisms for studying eukaryotic molecular genetics[1]. This yeast has a very small nuclear genome which, at about 14 megabases (Mb), is less than four times the size of that of the bacterium *Escherichia coli*. Like all eukaryotes, yeast divides its nuclear genome between a number of linear chromosomes. The 16 yeast chromosomes have been defined by both genetic[2] and physical[3] analyses. Each contains a single duplex DNA molecule[4] whose average size is similar to that of the genome of a T-even bacteriophage and thus represents an achievable objective for nucleotide sequencing with current technology[5]. Such a sequence analysis would not only advance yeast molecular genetics but also provide information important to our understanding of the genomes of higher eukaryotes. The pattern of gene expression in yeast is essentially similar to that in higher organisms[1]. Moreover, a number of yeast genes have been identified that share both structural and functional homology with genes of multicellular eukaryotes.

Despite their small size, yeast chromosomes resemble their counterparts from higher organisms in structure and in their mechanisms of replication, recombination and segregation[6]. The ease with which the yeast genome can be manipulated by both classical and recombinant DNA techniques has enabled all the functional chromosome components to be isolated: centromeres[7], telomeres[8] and replication origins[9]. These advances in our knowledge of chromosome structure and function, gained by the use of gene cloning techniques, have been paralleled by the study of chromosome recombination using a combination of classical and molecular approaches[10]. The availability of the complete nucleotide sequence of a yeast chromosome permits

[1]Manchester Biotechnology Centre, UMIST, Manchester M60 1QD, UK; [2]Department of Cell Biology Genetics, Leiden University, The Netherlands; [3]Dipartimento di Genetica e di Biologia dei Microrganismi, Universita di Milano, Italy; [4]LBMS, F-33000 Bordeaux, France; [5]Dipartimento di Fisiologia e Biochimica, Universita di Milano, Italy; [6]Instituto Molecular Biology and Biotechnology, PO Box 1527, GR-71110 Heraklio, Crete; [7]Institut Curie, Centre Universitaire, F-91405 Orsay, France; [8]Centro de Biologia Molecular, E-28049 Madrid, Spain; [9]Faculté des Sciences Agronomiques, B-5030 Gembloux, Belgium; [10]Dipartimento di Chimica Biologica, I-35100 Padova, Italy; [11]Institut de Génétique et Microbiologie, Université de Paris-Sud, 91405 Orsay, France; [12]Department of Molecular and Cell Biology, University of Aberdeen, UK; [13]Service de Biochimie CEN Saclay, F-91191, France; [14]Institut für Mikrobiologie, D-6100 Darmstadt, Germany; [15]Lab. voor Moleculaire Biologie, B-9000 Gent, Belgium; [16]Katholieke Universiteit Lab. voor Gentechnologie, B-3001 Leuven, Belgium; [17]La Cruz del Campo S.A., PO Box 53, E-41080 Sevilla, Spain; [18]Research Institute of CERIA-COOVI, B-1070 Bruxelles, Belgium; [19]Institut Pasteur, F-75724 Paris Cedex 15, France; [20]Institut für Mikrobiologie und Molekularbiologie der Universität, 6300 Giessen, Germany; [21]Institut für Physiologische Chemie, Universität München, 8000 München, Germany; [22]University of Rome, Department of Cell and Developmental Biology, I-00185 Roma, Italy; [23]AFRC Institute of Food Research, Norwich Research Park, Colney NR4 7UA, UK; [24]Unité de Biochimie Physiologique, Université de Louvain, B-1348 Louvain-la-Neuve, Belgium; [25]Université Libre de Bruxelles, Lab. Cell Physiology and Yeast Genetics, B-1050 Bruxelles, Belgium; [26]University of Amsterdam, Section for Molecular Biology, NL-1098 SM Amsterdam, The Netherlands; [27]Martinsried Inst. for Protein Sequences, D-8033 Martinsried, Germany; [28]Génétique Moléculaire, École Normale Superieure, F-75005 Paris, France; [29]Centre de Génétique Moleculaire du CNRS, F-91190 Gif-sur-Yvette, France; [30]Institut für Mikrobiologie der Universität Dusseldorf, D-4000 Dusseldorf 1, Germany; [31]Department of Biology, Kobe University, Kobe 657, Japan; [32]Trinity College, Department of Genetics, Dublin 2, Ireland; [33]Department of Biochemistry and Molecular Biology, Vrije Universiteit, NL-1081 HV Amsterdam, The Netherlands; [34]Department of Microbiology and Molecular Genetics, New Jersey Medical School, NJ07103, USA; [35]Fakultät für Biologie der Universität, D-7750 Konstanz, Germany; [36]Carlsberg Lab., DK-2500 Copenhagen Valby, Denmark; [37]Commission of the European Communities, B-1049 Brussels, Belgium

a detailed comparison of genetic and physical maps and so helps to define local variations in the frequency of recombination[11]. *S. cerevisiae* also contains a number of transposons (called Ty elements) which show similarities to retroviruses and other mobile genetic elements found in multicellular eukaryotes[12] and which are a major source of chromosome polymorphisms.

All of this suggests that sequencing the yeast genome should provide a useful paradigm to guide our interpretation of the information gained from the sequences of larger genomes when these become available. Indeed, at 14 Mb (compared with 3,000 Mb for the human genome), the 16 chromosomes of yeast might be regarded as comprising a basic gene-set required for the maintenance of a eukaryotic mode of cell organization and propagation. There are also practical reasons why the determination of the yeast genome sequence should be a useful precursor to larger enterprises such as sequencing the genomes of higher plants or humans. First, the functional analysis of novel genes discovered from the sequence is facilitated by the easy methods for gene disruption and replacement[13] which are available in yeast. Second, high-capacity cloning vectors (yeast artificial chromosomes, or YACs[14]), propagated in yeast, are important for creating clone banks of human or plant genes. Thus, it is useful to establish in the first instance the genome sequence of the vehicle organism. Finally, sequencing the yeast genome can act as a pilot scheme for assessing new techniques designed to speed sequence determination and analysis[15]. The yeast chromosome III sequence reported here has been determined mainly by conventional techniques. In fact, only 25 kb of the 385-kb sequence from which the final 315-kb consensus was derived has been obtained using automated techniques (7%). The following approaches were used to divide the primary clones (indicated by upper bars in Fig. 1) into fragments of a size suitable for sequencing: directed subcloning (19 kb; 5%), shotgun cloning (82 kb; 21%), chromosome walking with synthetic oligonucleotides (137 kb; 36%), and nested deletions (147 kb; 38%).

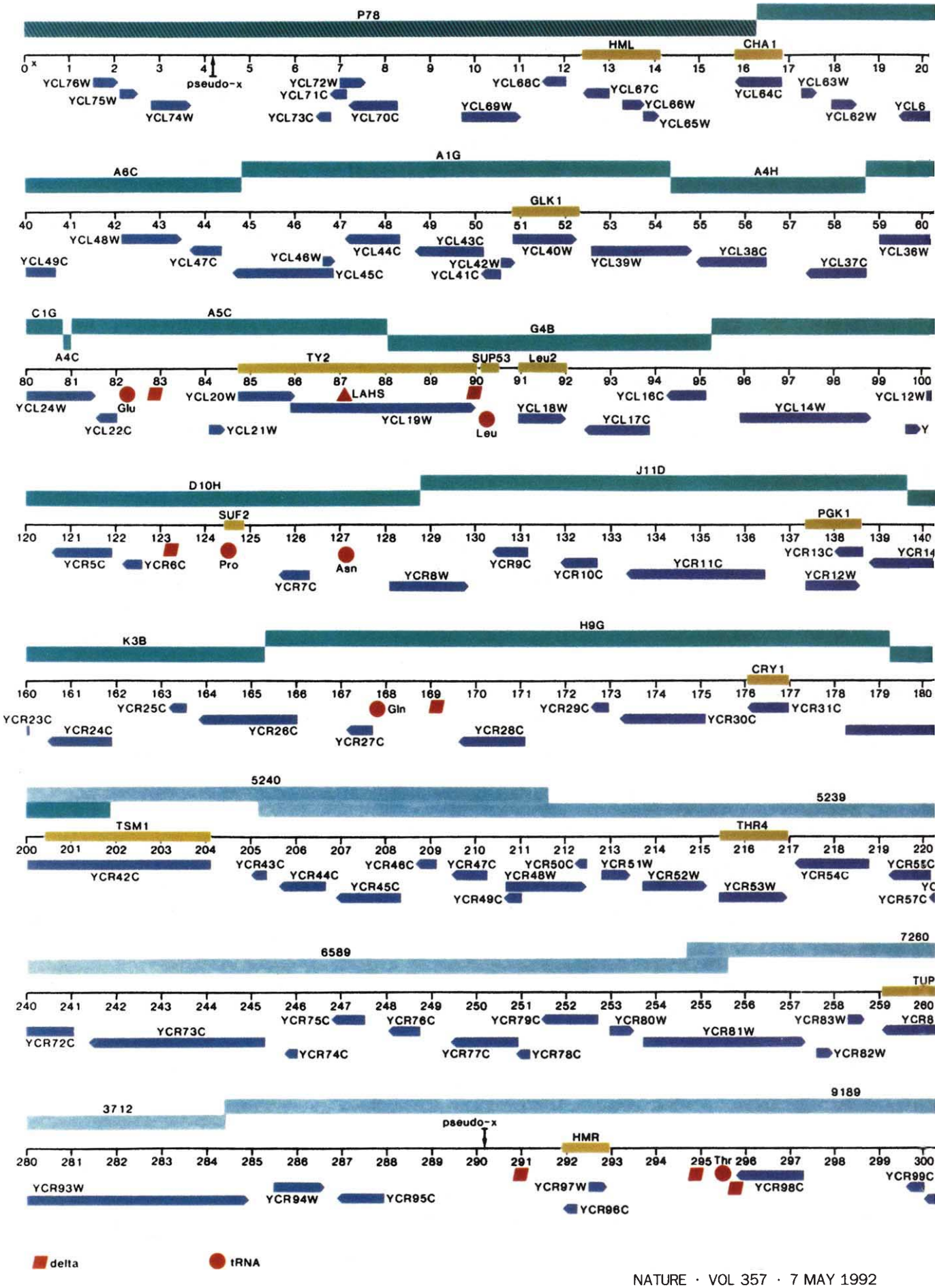
Chromosome III was divided between the laboratories of the Consortium and accredited clones were distributed to each group. Each laboratory was responsible for one or two work units (1 work unit is 8 kb of primary sequencing and 3 kb of overlap sequencing) and for disruptions of novel ORFs (some disruptions were performed instead of re-checking overlaps). The individual laboratories forwarded their data to the Martinsried Institute for Protein Sequences (MIPS), where the sequence was assembled by means of the GCG package[16] and a variety of algorithms were used to analyse ORFs and other sequence elements. MIPS also assembled a database for all extant *S. cerevisiae* DNA sequences, scanning not only the computer databases but also all published results. The project was initiated in January 1989 and the sequence completed by May 1991. Sequencing of entire eukaryotic genomes will probably require worldwide collaboration and our chromosome III sequence project represents an organizational and technological model for such enterprises.

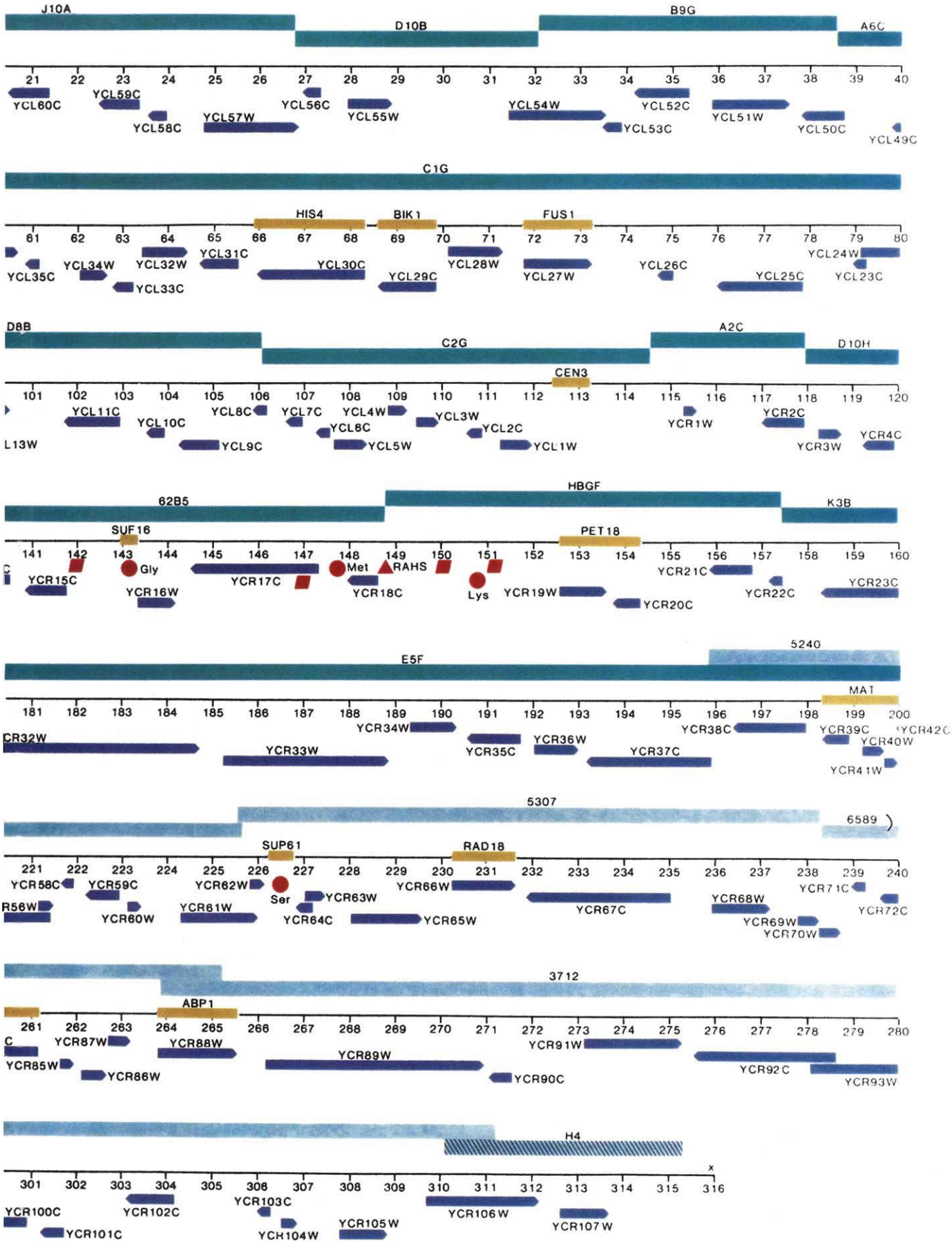
Chromosome III

Chromosome III is the third smallest chromosome in *S. cerevisiae*; its size was estimated from pulsed-field gel electrophoresis studies to be 300–360 kb[17]. This chromosome has been the subject of intensive study, not least because it contains the three genetic loci involved in mating-type control: *MAT*, *HML* and *HMR*[18]. *MAT*, in the middle of the right arm of the chromosome, is the expression locus that determines mating-type, whereas *HML* and *HMR* (located near the left and right ends of the chromosome) represent silent repositories of mating-type information. These three loci share nucleotide sequence homology, which means that intrachromosomal recombination events may generate circular derivatives of the chromosome. An *HML* × *MAT* recombination produces a 63-μm ('small ring') derivative which has been isolated[18] and used to generate an ordered clone bank of the chromosome in the shuttle vector YIp5 (ref. 19). This bank, necessarily, lacks a small part of the left arm and about half of the right arm of the chromosome; the latter was available from a series of clones constructed in lambda and cosmid vectors[20]. These two gene banks were used to generate the sequence covering 280 kb of the 315 kb of chromosome III and were derived from *S. cerevisiae* strains XJ24-24a[18,19] and AB972 (ref. 21). For the remaining 35 kb, clones derived from strains A364A[19] and DC5 (refs 22 and 23) were used. These strains are closely related (XJ24-24a, A364A) or isogenic (AB972, DC5) to the standard yeast laboratory strain S288C (ref. 24). The data are summarized in this article and a full listing of the sequence is available in the EMBL Data Library (accession number X59720) and will also be the subject of a more extensive publication.

A comparison of the chromosomes III from these strains, as well as from other laboratory and industrial isolates, permits an assessment of the degree of sequence polymorphism between different representatives of the same chromosome. Ty insertions in the chromosome are mainly confined to two regions that have been dubbed the left-arm and right-arm transposition hot spots (LAHS and RAHS, respectively[25–28]). Many strains (but not XJ24-24a) contain another Ty, distal to the RAHS, close to the gene *CRY1*. Interactions between this distal Ty and the RAHS seem to be a major cause of chromosome length polymorphisms. Some strains have a deletion in this region which causes a number of phenotypic effects, including respiratory deficiency[23], whereas other strains (including the progenitor S288C) have this region duplicated (ref. 19, and B.L.W. *et al.*, manuscript in preparation). We have sequenced a large overlap between the Newlon and Olson clone banks and this has provided detailed information about this polymorphic region as well as permitting the elucidation of the authentic unique sequence of the chromosome. The version of the chromosome presented in Fig. 1 is 315,357 base pairs (bp) long; it starts at the left telomere[29] and ends within the X sequence of the right telomere, about 400 bp from the terminus. A single Ty*2* element is shown in the LAHS,

FIG. 1 Location of ORFs and known genetic markers on chromosome III. ► Royal blue arrow-bars: ORFs (≥100 amino acids) are designated as follows: Y (for yeast) C (the third letter of the alphabet for the third chromosome); L or R (for the left and right arm of the chromosome); a number that designates the relative position of the ORF on its arm of the chromosome (the most centromere-proximal ORF being given the number 1); w or c (for Watson or Crick strand) to indicate the orientation of the ORF. All ORFs shown start with an ATG codon. Number scale is in kb. Upper blue bars indicate the clone used to obtain the primary sequence of that region of the chromosome. Dark turquoise bars: clones J10A [36, 6], D10B [16], B9G [23], A6C [25], A1G [18], A4H [17], C1G [30, 20], A4C [20, 1] A5C (1), G4B [1], D8B [1], C2G [1], A2C [2], D10H [4], J11D [24], 62B5-2D [28], HBGF [21], K3B [9], H9G [35, 14] and E5F [29, 19] were from the YIp5 bank of the small ring form of the chromosome (numbers in square brackets indicate the laboratories responsible for sequencing that clone, see addresses); all, except 62B5-2D (from DC021/DC022≠62) and HBGF (from CN31c), were from strain XJ24-24a (ref. 19). The *Bam*HI junctions between these clones were checked using a λCh4A bank derived from strain A364A (ref. 19). Light turquoise bars: clones 7121 [12, 32], 3270 [5, 3], 5240 [11, 13], 5239 [26, 33], 5307 [7, 15], 6589 [8], 7260 [21, 20], 3712 [22, 10] and 9189 [33, 26, 7] were from strain AB972 (ref. 20). Clones 7121 and 3270 are in the overlap region between the Newlon[19] and Olson[20] banks and are not indicated in the figure. The ends of the chromosome were sequenced using clones (dark turquoise, cross-hatched) p78 [36, 2, 11, 19] derived from strain A364A (ref. 19) and (light turquoise, cross-hatched) H4 [25] from strain DC5 (ref. 22). The Sanger dideoxy-sequencing technique[49] was used for the entire chromosome with the exception of parts of clone D10B, for which the chemical procedure of Maxam and Gilbert[50] was employed. Full details of the sequencing strategy will be published (S.G.O. *et al.*, manuscript in preparation) and the full 315, 356-bp sequence has been deposited in the EMBL Nucleotide Sequence Data Library under the accession number X59720. Two ORFs from the *pet18* complex are <100 amino acids in length and so are not shown in the figure. The following ORFs represent newly discovered essential genes: YCL17c, YCL14c, YCR69w.





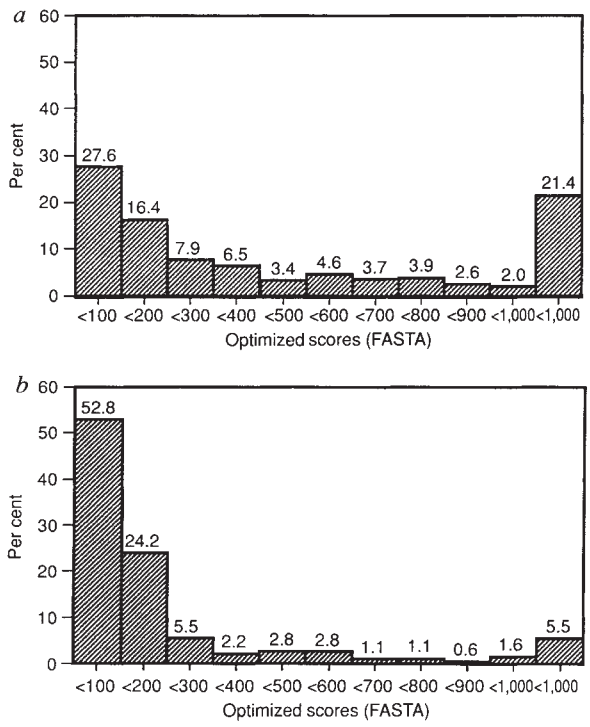


FIG. 2 Distribution of FASTA[51] scores for all yeast proteins (*a*) and for the chromosome III ORFs (*b*). The amino-acid sequence of the 968 protein-encoding yeast genes present in the data bases and the 182 ORFs in chromosome III were subjected to a similarity search against all known protein sequences present in the MIPSX data base using the Pearson and Lipman[51] FASTA algorithm. (MIPSX is a merged database comprising PIR-International, SwissProt and automatic translations of both the EMBL and GenBank nucleotide-sequence data sets.) Optimized scores were used to compile the data. In this analysis, scores due to matches with identical sequences were omitted and the next highest score to a non-identical sequence was taken.

but no transposon is shown in the RAHS; this is because the fused tandem pair of Ty elements found at this site in S288C, AB972, DC5 and XJ24-24a has been produced by a recombination event that deleted part of the unique sequence of the chromosome, including an ORF and a tRNA gene[27]. The form of the chromosome shown, and the sequence entered in the EMBL Data Library, permits the inclusion of these genes. The following lengths of the chromosomes III in the source strains have been calculated from the sequence data: XJ24-24a, 324 kb; A364A, 322 kb; DC5, 336 kb; S288C and AB972, 345 kb.

Open reading frames

The transcript analysis of Yoshikawa and Isono[22] predicted that chromosome III specifies 160 messenger RNA molecules (≥350 nucleotides) in addition to the transcripts of the Ty elements. This is likely to be an underestimate as the northern blots were run only on poly(A)$^+$ RNA extracted from cells growing in a rich medium. Nonetheless, the agreement between the transcript map and the ORF map is, in general, very high. For example, Yoshikawa and Isono[22] show two transcripts of 6.5 and 4.2 kb in the region corresponding to base pairs 178,250 to 188,900 in the sequence. These transcript sizes agree very well with those of the two ORFs in this region, YCR32w (6,501 bp) and YCR33w (3,678 bp)[30,31]. The complete sequence analysis of the chromosome reveals 182 ORFs of ≥100 amino acids. ORFs of this length have <0.2% probability of occurring by chance in *S. cerevisiae* DNA[32]. All 182 ORFs begin with an ATG; those ORFs whose sequence is entirely contained within another reading-frame have been excluded from the analysis. Seventeen ORFs overlap one another by between 7 and 449 bp; these are all included. Figure 1 shows the distribution of the 182 ORFs in the chromosome sequence. Of the 182 putative protein-encoding genes shown, only 34 appear on the *S. cerevisiae* genetic map (ref. 2, and R. K. Mortimer, personal communication). These numbers illustrate the fact that, even in a genome as small and as intensively studied as that of yeast, only a minor fraction of the genes has been identified by classical means. Therefore one justification for large-scale sequencing projects is the identification of new areas of investigation, as demonstrated by our sequence analysis of a single small eukaryotic chromosome.

The MIPS YEASTPROT database currently contains the sequence of 968 protein-encoding genes from *S. cerevisiae*. Of these 968, a total of 251 (26%) show amino-acid sequence similarity to other proteins encoded by the same species, giving FASTA scores of ≥200. A further 286 (30%) show a similar level of homology to proteins from organisms other than *S. cerevisiae*. A survey of the 145 newly discovered ORFs from chromosome III presents a strikingly different picture. Only ~10% (15/145) show significant similarity (FASTA ≥ 200) to non-*S.cerevisiae* genes already in the databases, 15 more are homologous with other genes from *S. cerevisiae* itself, and 117 (80%) show no significant homology to any previously sequenced genes. (Figure 2 shows the distribution of FASTA scores for the set of 182 chromosome III ORFs and all previously known ORFs from *S. cerevisiae*.)

The 15 newly discovered chromosome III gene products that show significant similarity to non-*S. cerevisiae* proteins are listed in Table 1*a*. Some, for example alcohol dehydrogenase and asparagyl-tRNA synthetase, are routine housekeeping proteins, but many others are unexpected. For instance, one of the highest scores is given by YCL17c, which is homologous to the *nif*S gene product of nitrogen-fixing bacteria[33,34]. *S. cerevisiae* does not fix atmospheric nitrogen[35], yet this highly conserved gene is essential to its growth[36]. Recent evidence suggests that this gene is involved in both tRNA processing and mitochondrial metabolism (P. Leong-Morgenthaler *et al.*, manuscript submitted). YCL74w shows a high degree of similarity to a *copia*-like protein from *Arabidopsis*[37], the tobacco transposon (Tnt*1*; ref. 38) and to *copia*[39] itself. Chromosome III contains representatives of all four classes of yeast transposons[25-28]; YCL74w may indicate the presence of a fifth class. If the set of 15 proteins encoded by the rest of the *S. cerevisiae* genome that show greatest similarity to proteins from other species is examined, then the majority of matches are to proteins from other yeasts and filamentous fungi and all 15 are proteins of well-known function. These findings demonstrate that genome sequencing reveals, at a high frequency, new functions that have not been discovered by classical techniques and suggest that yeast molecular geneticists are working on only a small subset of the problems presented by their experimental organism.

The set of chromosome III ORFs that show significant similar-

ity to previously characterized genes on other yeast chromosomes (Table 1*b*) reveals no particular pattern, except that kinase genes are overrepresented (5/15; YCL24w, YCR8w, YCR91w, YCR73c, YCR38c). This group includes similarities to some genes, like *YKR*, which were expected to have homologues elsewhere in the genome[40]. It may be necessary to perform several inactivations to reveal the role of genes of this type, but the task of assigning functions to novel genes identified by systematic sequencing is still more readily addressed in yeast than in any other eukaryote. Of the 145 novel ORFs found in the chromosome III sequences, 55 (38%) have been subjected to gene disruption[13]. For three genes this was a lethal event. Of the remainder, 42 disruptants have been tested for other effects on phenotype, using a standard battery of tests compiled by P.P.S., and in 14 cases some effect was observed. The phenotypes included heat and cold sensitivity, respiratory deficiency (or an enhanced level of mutation to that phenotype), hypersensitivity to various drugs, sterility and defects in budding, sporulation or meiotic recombination. This leaves 28 of 45 genes tested (62%) for which no phenotype can be assigned; Goebl and Petes[41] had previously suggested, based on a study of random disruptions, that 70% of the yeast genome was 'dispensable'. It is unlikely that these genes make no contribution to the fitness of the organism and this implies that our understanding of yeast physiology and cell biology is lagging behind our molecular genetic analysis.

A very small proportion of genes sequenced from *S. cerevisiae* contain introns (about 2%; ref. 42). On this basis, it might be expected that chromosome III would include three intron-containing genes. Apart from the mating-type loci[18], three have been detected (YCR31c and two others that are less than 100 amino acids long) by searching for the TACTAAC box and the 5′ and 3′ splice junctions[42]. One of these genes, *cry1* (YCR31c), was already known; it encodes the ribosomal protein rps59 (ref. 43). A large proportion of intron-containing genes in yeast encode ribosomal protein subunits[42], but neither the protein-encoding sequences nor the upstream regions of the two small intron-containing genes on the chromosome give any indication that they are ribosomal proteins as well.

Transfer RNA genes

Like most organisms, *S. cerevisiae* contains multiple gene copies encoding iso-accepting species of tRNA. Feldmann[44] has estimated that yeast contains 360 tRNA genes. Four such genes had been identified as suppressors and mapped to chromosome III (*SUP53, SUP61, SUF2* and *SUP16*; ref. 2, and R. K. Mortimer, personal communication). The sequence identifies six more (encoding $tRNA^{Asn}$, $tRNA^{Glu}$, $tRNA^{Gln}$, $tRNA^{Lys}$, $tRNA^{Met}$ and $tRNA^{Thr}$). A total of seven tRNA genes would be expected on a statistical basis; all of the 10 found are within 500 bp of a Ty or delta element. The frequent association of tRNA genes with such elements has been noted previously[45,46] and their possible role in the amplification and spread of such genes has been discussed[27].

Comparison of genetic and physical maps

The average ratio between genetic and physical map distances in yeast is estimated to be 0.34 centimorgans (cM) per kb (ref. 2). Smaller chromosomes, such as I and VI, have a higher ratio and Kaback *et al.* suggested that this is necessary to ensure at least one crossover occurs at each meiosis so that these small chromosomes can segregate correctly[47]. Within chromosomes, recombination frequency varies widely and both recombination hot spots and cold spots have been identified[11]. The sequence of chromosome III may suggest some molecular basis for these local variations in recombination frequency. A detailed comparison of the physical map of the chromosome (based on the sequence) and the current version of the genetic map (ref. 2; and R. K. Mortimer, personal communication) is unrealistic as the latter is, necessarily, a compromise between the results of several laboratories and also takes advantage of the emerging physical data. Nevertheless, some instructive generalizations may be made. The average ratio of genetic map distance to physical distance for the chromosome is 0.51 cM per kb; this is in line with the estimates for chromosomes I and VI (0.62 and 0.55 cM per kb, respectively) and supports Kaback's[47] postulate. The variation in the cM per kb ratio for different intervals along the chromosome is at least 10-fold; it is lowest close to the centromere and greatest midway down each arm. There is thus an approximate correlation between the pattern of genetic recombination and that of transcription along the chromosome[22]. An association between recombination and transcription has been suggested previously[48], and it may be that both processes require relatively naked DNA within the chromatin. A region where the frequency of recombination is particularly high is the interval between *MAT* and *thr4* (1.12–1.27 cM per kb); this

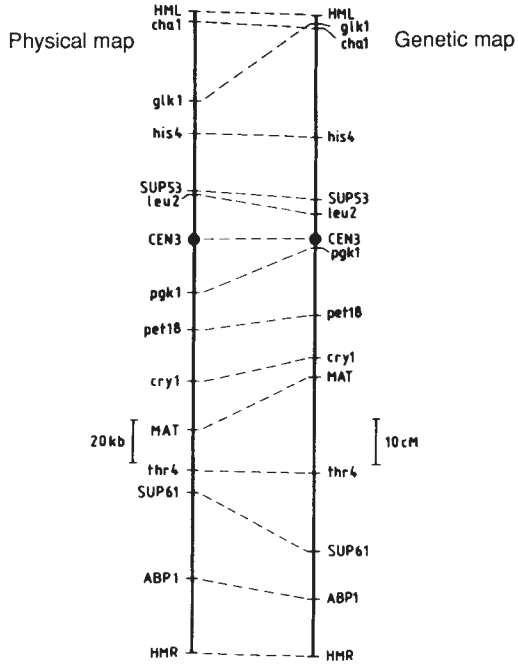


FIG. 3 Comparison of genetic and physical maps. The physical map has been derived from the sequence by assigning each locus to a point in the centre of the appropriate ORF, group of ORFs, or DNA element. Thus each locus, given as a point on this physical map, may represent a region of the chromosome from 0.1 kb to 6 kb in length. The genetic map is derived mainly from the analysis of meiotic recombination events[2]; fine structure mapping data (permitting estimation of genetic length) are available for only a few genes.

TABLE 1 Chromosome III gene products showing significant homology to extant protein sequences

ORF	Coordinates	Rational gene name	Similar sequence (database: accession no.)	FASTD score	Effect of gene disruption
			(*a*) Newly discovered ORFs showing similarity to non-*S. cerevisiae* proteins		
YCL57w	24,752–26,887		Rat soluble metalloendopeptidase (Genbank: M61142)	1,012	
YCR92c	278,640–275,500		Human DUC-1 protein upstream of DHFR (EMBL: J04810)	960	Non-lethal, no detectable phenotype*
YCL17c	93,881–92,391	*NFS1*	*Anabaena nifs* nitrogen-fixation protein (EMBL: J05111; M21840)	778	Lethal†
YCR2c	117,951–116,986		Mouse DIFF6 protein, regulated by gp90(MEL-14) (EMBL: X13303)	660	
YCR24c	161,932–160,457		*E. coli* asp-tRNA synthetase (EMBL: M33145)	550	
YCL74w	2,816–3,739		*Arabidopsis copia*-like protein (EMBL: M53975)	550	
YCR11c	136,474–133,328	*ADP1*	*Drosophila* white pigment protein (EMBL: X51749)	512	Non-lethal, no detectable phenotype‡
YCR63w	227,040–227,510		*Xenopus* G10 protein, developmentally regulated (EMBL: X15243)	480	Slow growth§
YCL43c	50,196–48,631	*PDI1*	Protein disulphide isomerase precursor (PIR: JX0182; EMBL: M62815, X52313)	2,501	Lethal‖
YCR105w	307,800–308,882		*Zymomonas* alcohol dehydrogenase I (EMBL: M32100)	366	
YCR27c	167,715–167,089		*Dictyostelium ras*-related protein (EMBL: X54291)	309	
YCR65w	228,031–229,626		*Drosophila fkh* homeotic gene (EMBL: J03177)	286	
YCR57c	221,442–220,126		Human transducin β-2 chain, GTP-binding (Gen Bank: M36429)	266	
YCL11c	102,964–101,684		*Xenopus* poly(A)-binding protein (PIR: S12000)	227	
YCR107w	312,620–313,708		Tobacco auxin-induced protein (EMBL: X56269)	218	
			(*b*) Newly discovered ORFs showing significant similarity to other *S. cerevisiae* proteins		
YCL25c	77,880–75,982		General amino-acid permease, GAP1 (EMBL: X52633)	1,527	
YCL64c	16,869–15,790		L-serine dehydratase, SDH1 (EMBL: X52657)	984	
YCL48w	42,140–43,528		Sporulation-specific SPS2 protein (EMBL: M13629)	941	
YCR67c	235,046–231,852		Sec12p membrane glycoprotein (Gen Bank: X13161)	889	
YCL24w	79,119–81,566		Sucrose non-fermenting SNF1 protein kinase (EMBL: M13971)	650	
YCR8w	128,072–129,880		Nitrogen permease reactivator, NPR1 (EMBL: X56084)	445	
YCR52w	213,725–215,173		Gene complementing petite type mutation (EMBL: X62430)	434	
YCR45c	208,342–206,870		Proteinase B precursor (EMBL: M18097)	429	
YCR91w	273,138–275,315		Protein kinase, YKR (EMBL: M24929)	397	
YCL35c	61,142–60,813		Glutaredoxin (PIR: A35492)	396	
YCR73c	245,320–241,379		Protein kinase, Ssp31 (PIR: JQ1118)	321	
YCR83w	258,312–258,692		Thioredoxin II (EMBL: M59169)	272	
YCR28c	171,115–169,580		Allantoate permease (EMBL: M24098)	255	
YCR89w	266,168–270,994		A-agglutinin core subunit, AGA1 (EMBL: M28164)	250	
YCR38c	197,967–196,354		Cell division cycle, CDC25 protein (EMBL: M15458)	201	

Full details of similarity search procedures are given in the legend to Fig. 2. For each ORF, the highest optimum FASTA score found is listed. References to the authors of the pre-existing sequence may be found in each database entry, the accession numbers for which are given in the figure. The term 'Yeast' here indicates the conspecific group, *S. cerevisiae*, *S. carlsbergensis* and *S. uvarum*. For the division between *S. cerevisiae* and non-*S. cerevisiae* matches (see text), first priority was given to similarities to other *S. cerevisiae* sequences; thus, if a match with a FASTA score $\geq$200 with another *S. cerevisiae* sequence was found, this was counted even if a match to a non-*S. cerevisiae* sequence produced a higher score. ►

TABLE 1—continued

ORF	Coordinates	Rational gene name	Similar sequence (database: accession no.)	FASTD score	Effect of gene disruption
			(*c*) Genes previously sequenced but not assigned to chromosome III		
YCR75c	247,550–246,771	*ERS1*	ER defect suppressor, ERS1 protein (EMBL: X52468)	1,455	
YCR5c	121,931–120,552	*CIT2*	Cytosolic citrate synthase (EMBL: Z11113)	2,241	
			(*d*) Genes previously sequenced and assigned to chromosome III but not mapped		
YCL50c	38,775–37,813	*DTP1*	Diadenosine tetraphosphate phosphorylase, DTP1 (EMBL: M35204)	1,602	
			(*e*) Genes previously mapped on chromosome III but not sequenced		
YCR53w	215,428–216,969	*THR4*	*Corynebacterium* threonine synthase (Gen Bank: X56037)	759	
YCR9c	131,144–130,350	*RVS161*	Reduced viability on starvation, RVS161 protein (PIR: Y01077)	1,280	
			(*f*) Genes previously sequenced and mapped on chromosome III		
YCL18w	90,935–92,026	*LEU2*	β-isopropyl-malate dehydrogenase (EMBL: X03840)	1,688	
YCL27w	71,768–73,303	*FUS1*	Cell fusion protein, Fus 1 (EMBL: M16717)	2,425	
YCL29c	68,568–69,867	*BIK1*	Nuclear fusion protein, Bik1 (EMBL: M16717)	1,304	
YCL30c	68,299–65,903	*HIS4*	Histidinol dehydrogenase (EMBL: V01310)	3,716	
YCL40w	50,810–52,309	*GLK1*	Glucokinase (EMBL: M24077)	2,422	
YCL66w	13,271–13,795	*HML*	Hmlα1 protein	896	
YCL67c	13,007–12,378		Hmlα2 protein (EMBL: V01315)	979	
YCR12w	137,347–138,594	*PGK1*	Phosphoglycerate kinase (EMBL: J01342)	1,911	
YCR19w	152,544–153,632	*PET18*	Maintenance of killer, MAK32 protein (PIR: 507695)	1,807	
YCRC20c	154,366–153,722	*PET18*	MAK31 protein (PIR: 507695)	1,155	
YCR31c	176,968–176,128	*CRY1*	Ribosomal protein s59 (EMBL: M16126)	617	
YCR40w	199,173–199,697	*MAT*	Matα1 protein	896	
YCR39c	198,909–198,280		Matα2 protein (EMBL: V01315)	979	
YCR42c	204,128–199,908	*TSM1*	Temperature-sensitive lethal TSM1 protein (EMBL: M60486)	6,911	
YCR66w	230,224–231,684	*RAD18*	DNA repair protein, Rad18 (EMBL: X125880	2,315	
YCR84c	261,186–259,048	*TUP1*	Repressor protein (EMBL: M35861)	3,156	
YCR88w	263,802–265,577	*ABP1*	Actin-binding (EMBL: X51780)	2,674	
YCR97w	292,568–293,051	*HMR*	Matα1 protein	896	
YCR96c	292,271–291,915		Matα2 protein (EMBL: V01315)	979	

* Region between *Cla*I site at 278,001 to the *Xba*I site at 277,004 replaced with *URA3*.

† Gene disruption experiments for this ORF are detailed in ref. 36, and in P. Leong-Morgenthaler *et al.*, manuscript submitted.

‡ *URA3* inserted into *Bgl*II site at 134,370.

§ *URA3* inserted into *Xba*I site at 227,218.

‖ Region between *Bst*EII site at 48,846 to the *Bst*EII site at 50,887 replaced with *URA3*. The techniques used for the other 51 gene disruptions may be summarized as: 23 insertion mutations (9, *URA3*; 6, *HIS3*; 5, MiniMu; 2, *TRP1*; 1, *HIS4*) and 28 deletion/replacement mutations (15, *URA3*; 6, *HIS3*; 5, *LEU2*; 1, *TRP1*; 1, *LYS2*).

partly explains the apparent paucity of markers in this segment of the genetic map. There is only a single discrepancy in the gene order derived from the sequence and that in the genetic map; the order of *glk1* and *cha1*, on the far left arm of the chromosome, is reversed. These data are summarized in Fig. 3.

The sequence data also provide evidence of past recombination events. A 245-bp sequence between residues 290,656 and 290,901 on the right arm (Fig. 1) shows a high degree of homology with the X element found in all yeast telomeres[29]. A similar sequence, 252 bp long, is found between residues 4,065 and 4,317 at the left end of the chromosome. It may be that chromosome III was once even shorter than it is now and that it has grown by recombination events involving its telomeres. Both the instability of linear plasmids and small artificial chromosomes[6] in yeast, as well as the Kaback rule[47], suggest that there may be selection against short chromosomes in *S. cerevisiae*. Moreover, a region near the middle of the left arm of chromosome XI has been shown by our analysis to contain sequences homologous to both the X and Y′ telomere elements (B.D. *et al.*, unpublished results).

Discussion

In sequencing the entire unique 315 kb of yeast chromosome III, we have generated 385 kb of total sequence. Thus 22% of the chromosome has been independently sequenced by at least two laboratories. The level of agreement between these sequence determinations was very high, the mismatch frequency being 0.4 per kb for clones from the same strain and 6.2 per kb for clones from different strains. In the latter case, 65% of the mismatches in ORFs occur in the third base of the codons. A re-check of the mismatch regions by the different pairs of laboratories reduced the discrepancies to zero, where clones came from the same strain, but had no significant impact on the number of differences between strains. In an independent check of the sequence, the restriction map predicted from the sequence has been compared with that derived by experiment (I. Collins and C.S.N., unpublished results). Of 504 6-bp restriction sites sampled, only four showed a discrepancy between the predicted and experimental maps (a putative error rate of 1.3 per kb). All four discrepancies represent predicted sites that could not be identified by experiment, therefore these sites may not be recognized by the appropriate restriction enzyme owing to context effects or DNA modifications. The chromosome III sequence is the largest data set available that has been subjected to such independent checking and so indicates the level of accuracy that larger genome projects may achieve using current technology.

The chromosome III sequence has revealed 145 novel protein-encoding genes and a start has been made on their functional analysis. The results so far indicate that there are vast areas of yeast genetics of which we are completely ignorant and emphasize the need for molecular genetics and physiological studies to proceed hand-in-hand. The data also call for a radical re-appraisal of our view of the yeast genetic map. The availability of the sequence establishes unequivocally the locations of the different genes on the chromosome. In consequence, the genetic map acquires a different emphasis; it becomes much more a tool with which to study recombination and the dynamics of chromosome evolution. The goal of sequencing the entire yeast genome is achievable with present technology and this sequence will prove at least as important to the future development of eukaryotic molecular biology as the classical *S. cerevisiae* genetic map has in the past. The complete sequence of the yeast genome will open up new areas of molecular genetics and establish a foundation for the interpretation of sequence data from higher organisms. □

Received 30 December 1991; accepted 27 March 1992.

1. Strathern, J. N., Jones, E. W. & Broach, J. R. (eds) *The Molecular Biology of the Yeast* Saccharomyces: *Metabolism and Gene Expression* (Cold Spring Harbor Laboratory, New York, 1982).
2. Mortimer, R. K. *et al. Yeast* **5,** 321–403 (1989).
3. Carle, G. F. & Olson, M. V. *Nucleic Acids Res.* **12,** 5647–5664 (1984).
4. Petes, T. D. *et al. Cold Spring Harb. Symp. quant. Biol.* **38,** 9–16 (1973).
5. Baer, R. *et al. Nature* **310,** 207–211 (1984).
6. Newlon, C. S. *Microbiol. Revs* **52,** 568–601 (1988).
7. Clarke, L. & Carbon, J. *Nature* **287,** 504–509 (1980).
8. Szostak, J. W. & Blackburn, E. H. *Cell* **29,** 245–255 (1982).
9. Stinchcomb, D. T., Struhl, J. & Davis, R. W. *Nature* **282,** 39–43 (1979).
10. Klar, A. & Strathern, J. N. (eds) *Mechanisms of Yeast Recombination* (Cold Spring Harbor Laboratory, New York, 1986).
11. Petes, T. D., Malone, R. E. & Symington, L. S. in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis and Energetics* (eds Broach, J. R., Pringle, J. R. & Jones, E. W.) 407–521 (Cold Spring Harbor Laboratory, New York, 1991).
12. Boeke, J. D. in *Mobile DNA* (eds Berg, D. E. & Howe, M. M.) 335–374 (ASM, Washington DC, 1989).
13. Rothstein, R. J. *Meth. Enzym.* **101,** 202–211 (1983).
14. Burke, D. T., Carle, G. F. & Olson, M. V. *Science* **236,** 806–812 (1987).
15. Martin, W. J. *Genome* **31,** 1073–1080 (1989).
16. Devereux, J., Haeberli, P. & Smithies, O. *Nucleic Acids Res.* **12,** 387–395 (1984).
17. de Jonge, P. *et al. Yeast* **2,** 193–204 (1986).
18. Strathern, J. N., Newlon, C. S., Herskowitz, I. & Hicks, J. B. *Cell* **18,** 309–319 (1979).
19. Newlon, C. S. *et al. Genetics* **129,** 343–357 (1991).
20. Olson, M. V. *et al. Proc. natn. Acad. Sci. U.S.A.* **83,** 7826–7830 (1986).
21. Link, A. J. & Olson, M. V. *Genetics* **127,** 681–698 (1991).
22. Yoshikawa, A. & Isono, K. *Yeast* **6,** 383–401 (1990)
23. Toh-e, A. & Sahashi, Y. *Yeast* **1,** 159–172 (1985).
24. Mortimer, R. K. & Johnston, J. R. *Genetics* **113,** 35–43 (1986).
25. Warmington, J. R. *et al. Nucleic Acids Res.* **13,** 6679–6693 (1986).
26. Pederson, M. B. *Carlsberg Res. Commun.* **51,** 163–183 (1986).
27. Warmington, J. R. *et al. Nucleic Acids Res.* **15,** 8963–8982 (1987).
28. Stucka, R., Lochmuller, H. & Feldmann, H. *Nucleic Acids Res.* **17,** 4993–5001 (1989).
29. Button, L. L. & Astell, C. R. *Molec. cell. Biol.* **6,** 1352–1356 (1986).
30. Jia, Y., Slonimski, P. P. & Herbert, C. J. *Yeast* **7,** 413–424 (1991).
31. Rodriguez, F. *et al. Yeast* **7,** 631–641 (1991).
32. Sharp, P. M. & Cowe, E. *Yeast* **7,** 657–678 (1991).
33. Mulligan, M. E. & Haselkorn, R. *J. biol. Chem.* **264,** 19200–19207 (1989).
34. Arnold, W. *et al. J. molec. Biol.* **203,** 715–738 (1988).
35. Cook, A. H. *The Chemistry and Biology of Yeasts* (Academic, New York, 1958).
36. Symington, L. S. & Petes, T. D. *Molec. cell. Biol.* **8,** 595–604 (1988).
37. Konieczny, A. *et al. Genetics* **127,** 801–809 (1991).
38. Grandbastien, M.-A., Spielman, A. & Caboche, M. *Nature* **337,** 376–380 (1989).
39. Mount, S. M. & Rubin, G. M. *Molec. cell. Biol.* **5,** 1630–1638 (1985).
40. Ohno, S. *et al. FEBS Lett.* **222,** 279–285 (1987).
41. Goebl, M. E. & Petes, T. D. *Cell* **46,** 983–992 (1986).
42. Oliver, S. G. & Warmington, J. R. in *The Yeasts* Vol. 3 (eds Rose, A. H. & Harrison, J. S.) 117–160 (Academic, London, 1989).
43. Larkin, J. C., Thompson, J. R. & Woolford, J. R. *Molec. cell. Biol.* **7,** 1764–1775 (1987).
44. Feldmann, H. *Nucleic Acids Res.* **3,** 2379–2386 (1976).
45. Eigel, A. & Feldmann, H. *EMBO J.* **1,** 1245–1250 (1982)
46. Genbauffe, F. S., Chisholm, G. E. & Cooper, T. G. *J. biol. Chem.* **259,** 10518–19525 (1984).
47. Kaback, D. B., Steensma, H. Y. & de Jonge, P. *Proc. natn. Acad. Sci. U.S.A.* **86,** 3694–3698 (1989).
48. Keil, R. L. & Roeder, G. S. *Cell* **39,** 377–386 (1984).
49. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
50. Maxam, A. & Gilbert, W. *Meth. Enzym.* **65,** 499–559 (1980).
51. Pearson, W. R. & Lipman, D. J. *Proc. natn. Acad. Sci. U.S.A.* **85,** 2444–2448 (1988).

ACKNOWLEDGEMENTS. The Consortium of 35 European laboratories was aided by the provision of clones from laboratories in the USA and Japan and was established by the Biotechnology Division of the Biology Directorate of the CEC funded by their Biotechnology Action Programme (BAP). The overall organizer and leader of the Consortium was A. Goffeau (Université Catholique de Louvain, Belgium), assisted by S. Oliver (UMIST, UK) as DNA Coordinator, and J. Sgouros and W. Mewes (MIPS, Germany) as Informatics coordinators. We thank M. Olson and L. Riles for AB972 clones and for advice, and R. Mortimer for communicating genetic map data prior to publication. The graphic for Fig. 1 was prepared by Edeltraud Hanesch (Klinikum, Grosshadern) and that for Fig. 2 by Gertrude Schitteck (Universität Greifswald). This study represents the first phase of a major yeast genome sequencing project initiated by the EC and was carried out within the framework of the Biotechnology Action Programme (BAP) of the CEC under the general coordination of A. Vassarotti (Biotechnology Directorate) and the Université Catholique de Louvain. In addition, the following national agencies provided support:- Belgium: Service of the Prime Minister-Science Policy Office; Germany: Bundesminister für Forschung und Technologie; Greece: Ministry of Industry, Research and Technology; Italy: Consiglio Nazionale delle Ricerche, Comitato per le Biotecnologie e la Biologia Moleculare; Spain: Comision Interministerial de Ciencia y Tecnologia.