The complete DNA sequence of yeast chromosome III 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 >100 amino acids. So far, none 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 an eukaryotic genome as small and well-defined as that of yeast.

The baker's yeast *S. cerevisiae* is one of the most important experimental organisms for studying eukaryotic molecular genetics. 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 and physical analyses. Each contains a single duplex DNA molecule whose average size is similar to that of the genome of *T. thermophilus* bacteriophage and thus represents an achievable objective for nucleotide sequencing with current technology. 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. 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. 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, telomeres and replication origins. 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. The availability of the complete nucleotide sequence of a yeast chromosome permits
a detailed comparison of genetic and physical maps and so helps to define local variations in the frequency of recombination. \textit{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 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 where this 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 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 which are available in yeast. Second, high-capacity cloning vectors (yeast artificial chromosomes, or YACs), 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. 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 (62 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 Martinried Institute for Protein Sequences (MIPS), where the sequence was assembled by means of the GCG package and a variety of algorithms were used to analyse ORFs and other sequence elements. MIPS also assembled a database for all extant \textit{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 \textit{S. cerevisiae}; its size was estimated from pulsed-field gel electrophoresis studies to be 300–350 kb. This chromosome has been the subject of intensive study, not least because it contains the three genetic loci involved in mating-type control: \textit{MAT}, \textit{HML} and \textit{HMR}. \textit{MAT}, in the middle of the right arm of the chromosome, is the expression locus that determines mating-type, whereas \textit{HML} and \textit{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 \textit{HML} × \textit{MAT} recombination produces a 63-μm ("small ring") derivative which has been isolated 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 cosmids vectors. These two gene banks were used to generate the sequence covering 280 kb of the 315 kb of chromosome III and were derived from \textit{S. cerevisiae} strains XJ24-24a and AB972 (ref. 21). For the remaining 35 kb, clones derived from strains A364A 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 comprehensive 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). Many strains (but not XJ24-24a) contain another Ty, distal to the RAHS, close to the gene \textit{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, 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 and ends within the X sequence of the right telomere, about 400 bp from the terminus. A single Ty2 element is shown in the LAHS.

**FIG. 1** Location of ORFs and known genetic markers on chromosome III. Royal blue arrow-heads: 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. Nucleotide scale is in kb. Upper blue bars indicate the clone used to obtain the primary sequence and lower black bars indicate the region of the chromosome. Dark turquoise bars: clones JDA [36, 6], D16 [10, 16], B9G [23], A6G [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], D62B-2D [28], H6B [21], K3B [9], H6G [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 D62B-2D (from DCC21: DCC22±62) and H6B (from CNNC, see addresses except for the region XJ24-24a (ref. 19). The BamHI junctions between these EcoRI tags were checked using a 1ChA4A bank derived from strain A364A (ref. 19). Light turquoise bars: clones 712I [12, 32], 3270 [5, 3], 5210 [11, 13], 5329 [26, 33], 5307 [17, 15], 6588 [8], 7260 [21, 20], 3712 [22, 10] and 9189 [33, 26] were from strain AB972 (ref. 20). Clones 712I and 3270 are in the overlap region between the Newlon and Olson banks and are not indicated in the figure. The ends of the chromosome were sequenced using clones (dark turquoise, cross-hatched) P783 [36, 2, 11, 19] derived from strain A364A (ref. 19) and (light turquoise, cross-hatched) H4 [25] from strain DC5 (ref. 22). The Sanger dye-sequencing-technique was used for the entire chromosome with the exception of parts of clone D10B, for which the chemical procedure of Maxam and Gilbert was employed. Full details of the sequencing strategy will be published (S.G.O. et al., manuscript in preparation) and the full SL5, 356-bp sequence has been deposited in the EMBL Data Library (accession number X59720). Two ORFs from the petit18 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, YCR09w.
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 Isono22 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 Isono22 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 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 euchromatic 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 1A. 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 nifS gene product of nitrogen-fixing bacteria33,34. S. cerevisiae does not fix atmospheric nitrogen35; yet this highly conserved gene is essential to its growth36. Recent evidence suggests that this gene is involved in both RNA processing and mitochondrial metabolism (P. Leong-Morgenthaler et al., manuscript submitted). YCL74w shows a high degree of similarity to a copia-like protein from Arabidopsis37, the tobacco transposon (Tn11; ref. 38) and to copia39 itself. Chromosome III contains representatives of all four classes of yeast transposons24,28; 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 1b) 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. 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 143 novel ORFs found in the chromosome III sequences, 55 (38%) have been subjected to gene disruption. 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; Goebel and Petes 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, these 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. 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, 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 has estimated that yeast contains 360 tRNA genes. Four such genes had been identified as suppressors and mapped to chromosome III (SUP53, SUP61, SUP2 and SUP16; ref. 2, and R. K. Mortimer, personal communication). The sequence identifies six more (encoding tRNA^A^, tRNA^G^, tRNA^O^, tRNA^N^, 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 and their possible role in the amplification and spread of such genes has been discussed.

**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. Within chromosomes, recombination frequency varies widely and both recombination hot spots and cold spots have been identified. 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 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. An association between recombination and transcription has been suggested previously, 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; fine structure mapping data (permitting estimation of genetic length) are available for only a few genes.
<table>
<thead>
<tr>
<th>ORF</th>
<th>Coordinates</th>
<th>Rational gene name</th>
<th>Similar sequence (database accession no.)</th>
<th>FASTD score</th>
<th>Effect of gene disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCL57w</td>
<td>24,752–26,887</td>
<td>Rat soluble metalloendopeptidase (Genbank: M61142)</td>
<td>1.012</td>
<td>Non-lethal, no detectable phenotype*</td>
<td></td>
</tr>
<tr>
<td>YCR92c</td>
<td>278,640–275,500</td>
<td>Human DLC-1 protein upstream of DHFR (EMBL: J04810)</td>
<td>960</td>
<td>Lethal†</td>
<td></td>
</tr>
<tr>
<td>YOL127c</td>
<td>93,881–92,391</td>
<td>NFS1</td>
<td>778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR2c</td>
<td>117,951–116,986</td>
<td>Mouse DFF6 protein, regulated by gp90(MEL-14) (EMBL: X13303)</td>
<td>660</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR24c</td>
<td>161,932–160,457</td>
<td>E. coli asp-RTA synthetase (EMBL: M33145)</td>
<td>550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCL74w</td>
<td>2,816–3,739</td>
<td>Arabidopsis copia-like protein (EMBL: M53975)</td>
<td>550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR11c</td>
<td>135,474–133,328</td>
<td>Drosophila white pigment protein (EMBL: X51749)</td>
<td>512</td>
<td>Non-lethal, no detectable phenotype‡</td>
<td></td>
</tr>
<tr>
<td>YCR63w</td>
<td>227,040–227,510</td>
<td>Xenopus G10 protein, developmentally regulated (EMBL: X51243)</td>
<td>480</td>
<td>Slow growth§</td>
<td></td>
</tr>
<tr>
<td>YCL43c</td>
<td>50,196–48,631</td>
<td>Protein disulfide isomerase precursor (PIR: JX0182; EMBL: M52815, X52313)</td>
<td>2,501</td>
<td>Lethal</td>
<td></td>
</tr>
<tr>
<td>YCR10w</td>
<td>307,800–308,882</td>
<td>Zymomonas alcohol dehydrogenase I (EMBL: M32100)</td>
<td>366</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR12c</td>
<td>167,715–167,089</td>
<td>Dictyostelium ras-related protein (EMBL: X49291)</td>
<td>309</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR65w</td>
<td>228,031–229,626</td>
<td>Drosophila Ak homoeotic gene (EMBL: M33177)</td>
<td>286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR57c</td>
<td>221,442–220,126</td>
<td>Human transducin β-2 chain, GTP-binding (Gen Bank: M38429)</td>
<td>266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCL11c</td>
<td>102,964–101,684</td>
<td>Xenopus poly(A)-binding protein (PIR: S12000)</td>
<td>227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR127w</td>
<td>312,620–313,708</td>
<td>Tobacco auxin-induced protein (EMBL: X56269)</td>
<td>218</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. kuvum. 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 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>
<thead>
<tr>
<th>ORF</th>
<th>Coordinates</th>
<th>Rational gene name</th>
<th>Similar sequence (database: accession no.)</th>
<th>FASTD score</th>
<th>Effect of gene disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCR75c</td>
<td>247,550–246,771</td>
<td><strong>ERS1</strong></td>
<td>ER defect suppressor, ERS1 protein (EMBL: X52468)</td>
<td>1.455</td>
<td></td>
</tr>
<tr>
<td>YCR5c</td>
<td>121,931–120,552</td>
<td><strong>CIT2</strong></td>
<td>Cytosolic citrate synthase (EMBL: Z11113)</td>
<td>2.241</td>
<td></td>
</tr>
<tr>
<td>(c) Genes previously sequenced but not assigned to chromosome III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCL50c</td>
<td>38,775–37,813</td>
<td><strong>DTP1</strong></td>
<td>Diadenosine tetraphosphate phosphorylase, DTP1 (EMBL: M35204)</td>
<td>1.602</td>
<td></td>
</tr>
<tr>
<td>(d) Genes previously sequenced and assigned to chromosome III but not mapped</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCR53w</td>
<td>215,428–216,969</td>
<td><strong>THR4</strong></td>
<td>Corynebacterium threonine synthase (Gen Bank: X56037)</td>
<td>759</td>
<td></td>
</tr>
<tr>
<td>YOR9c</td>
<td>131,144–130,350</td>
<td><strong>RVS161</strong></td>
<td>Reduced viability on starvation, RVS161 protein (Pfam: Y01077)</td>
<td>1,280</td>
<td></td>
</tr>
<tr>
<td>(e) Genes previously mapped on chromosome III but not sequenced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YCL18w</td>
<td>90,935–92,026</td>
<td><strong>LEU2</strong></td>
<td>β-isopropyl-malate dehydrogenase (EMBL: X03840)</td>
<td>1,688</td>
<td></td>
</tr>
<tr>
<td>YCL27w</td>
<td>71,768–73,303</td>
<td><strong>RUS1</strong></td>
<td>Cell fusion protein, Fus 1 (EMBL: M16717)</td>
<td>2,425</td>
<td></td>
</tr>
<tr>
<td>YCL29c</td>
<td>68,568–69,867</td>
<td><strong>Bik1</strong></td>
<td>Nuclear fusion protein, Bik1 (EMBL: M16717)</td>
<td>1,304</td>
<td></td>
</tr>
<tr>
<td>YCL30c</td>
<td>68,299–65,903</td>
<td><strong>HIS4</strong></td>
<td>Histidinol dehydrogenase (EMBL: V01310)</td>
<td>3,716</td>
<td></td>
</tr>
<tr>
<td>YCL40w</td>
<td>50,810–52,309</td>
<td><strong>GLK1</strong></td>
<td>Glucokinase (EMBL: M24077)</td>
<td>2,422</td>
<td></td>
</tr>
<tr>
<td>YCL66w</td>
<td>13,271–13,795</td>
<td><strong>HML</strong></td>
<td>Hmlα1 protein</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>YCL67c</td>
<td>13,007–12,378</td>
<td></td>
<td>Hmlα2 protein</td>
<td>979</td>
<td></td>
</tr>
<tr>
<td>YOR12w</td>
<td>137,347–138,594</td>
<td><strong>PGK1</strong></td>
<td>Phosphoglycerate kinase (EMBL: J01342)</td>
<td>1,911</td>
<td></td>
</tr>
<tr>
<td>YCR19w</td>
<td>152,544–153,632</td>
<td><strong>PET18</strong></td>
<td>Maintenance of killer, MAK32 protein (Pfam: Y01077)</td>
<td>1,807</td>
<td></td>
</tr>
<tr>
<td>YCR20c</td>
<td>154,366–153,722</td>
<td><strong>PET18</strong></td>
<td>MAK31 protein (Pfam: Y01077)</td>
<td>1,155</td>
<td></td>
</tr>
<tr>
<td>YOR31c</td>
<td>176,968–176,128</td>
<td><strong>CRY1</strong></td>
<td>Ribosomal protein s59 (EMBL: M16126)</td>
<td>617</td>
<td></td>
</tr>
<tr>
<td>YCR40w</td>
<td>198,173–198,697</td>
<td><strong>MAT</strong></td>
<td>Matα1 protein</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>YCR39c</td>
<td>198,909–198,280</td>
<td></td>
<td>Matα2 protein (EMBL: V01315)</td>
<td>979</td>
<td></td>
</tr>
<tr>
<td>YCR42c</td>
<td>204,128–199,908</td>
<td><strong>TSM1</strong></td>
<td>Temperature-sensitive lethal TSM1 protein (EMBL: M60486)</td>
<td>6,911</td>
<td></td>
</tr>
<tr>
<td>YCR66w</td>
<td>230,224–231,684</td>
<td><strong>RAD18</strong></td>
<td>DNA repair protein, Rad18 (EMBL: X125880)</td>
<td>2,315</td>
<td></td>
</tr>
<tr>
<td>YCR84c</td>
<td>261,186–259,048</td>
<td><strong>TPL1</strong></td>
<td>Repressor protein (EMBL: M35661)</td>
<td>3,156</td>
<td></td>
</tr>
<tr>
<td>YCR88w</td>
<td>263,402–265,577</td>
<td><strong>ABP1</strong></td>
<td>Actin-binding protein (EMBL: X51780)</td>
<td>2,674</td>
<td></td>
</tr>
<tr>
<td>YCR97w</td>
<td>292,568–293,051</td>
<td><strong>HMR</strong></td>
<td>Matα1 protein (EMBL: V01315)</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>YCR96c</td>
<td>292,271–291,915</td>
<td></td>
<td>Matα2 protein (EMBL: V01315)</td>
<td>979</td>
<td></td>
</tr>
</tbody>
</table>

* Region between CiaI site at 278,001 to the XbaI 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 BglII site at 134,370.
§ URA3 inserted into XbaI site at 227,218.
|| Region between BstEII site at 48,846 to the BstEII 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).
### TABLE 1 Chromosome III gene products showing significant homology to extant protein sequences

<table>
<thead>
<tr>
<th>ORF</th>
<th>Coordinates</th>
<th>Rational gene name</th>
<th>Similar sequence (database: accession no.)</th>
<th>FASTD score</th>
<th>Effect of gene disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCL57w</td>
<td>24,752-26,688</td>
<td>Rat soluble metalloendopeptidase (GenBank: M61142)</td>
<td></td>
<td>1.012</td>
<td></td>
</tr>
<tr>
<td>YCR92c</td>
<td>278,640-275,500</td>
<td>Human DUC-1 protein upstream of DHFR (EMBL: X04810)</td>
<td></td>
<td>960</td>
<td>Non-lethal, no detectable phenotype*</td>
</tr>
<tr>
<td>YCL17c</td>
<td>93,881-92,391</td>
<td>Aratanae rust nitrogen-fixation protein (EMBL: J05111, M22840)</td>
<td></td>
<td>778</td>
<td>Lethal*</td>
</tr>
<tr>
<td>YCR2c</td>
<td>117,951-116,986</td>
<td>Mouse DIFF6 protein, regulated by baro9(MEL-14) (EMBL: X13303)</td>
<td></td>
<td>660</td>
<td></td>
</tr>
<tr>
<td>YCR24c</td>
<td>161,932-160,457</td>
<td>E. coli asp-tRNA synthetase (EMBL: M33145)</td>
<td></td>
<td>550</td>
<td></td>
</tr>
<tr>
<td>YCL74w</td>
<td>2,816-3,739</td>
<td>Arabidopsis cox1-like protein (EMBL: M53975)</td>
<td></td>
<td>550</td>
<td></td>
</tr>
<tr>
<td>YCR11c</td>
<td>136,474-133,328</td>
<td>Drosophila white pigment protein (EMBL: X51749)</td>
<td></td>
<td>512</td>
<td>Non-lethal, no detectable phenotype*</td>
</tr>
<tr>
<td>YCR63w</td>
<td>227,040-227,510</td>
<td>Xenopus G10 protein, developmentally regulated (EMBL: X15243)</td>
<td></td>
<td>480</td>
<td>Slow growth$</td>
</tr>
<tr>
<td>YCL43c</td>
<td>50,196-48,631</td>
<td>Protein disulfide isomerase precursor (PIR: JX0182, EMBL: M62815, X52313)</td>
<td></td>
<td>2,501</td>
<td>Lethal</td>
</tr>
<tr>
<td>YCT105w</td>
<td>307,800-308,882</td>
<td>Zymomonas alcohol dehydrogenase I (EMBL: M32100)</td>
<td></td>
<td>356</td>
<td></td>
</tr>
<tr>
<td>YCR27c</td>
<td>167,715-167,089</td>
<td>Dicytostelium ras-related protein (EMBL: X54291)</td>
<td></td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>YCR65w</td>
<td>228,031-229,626</td>
<td>Drosophila 58d homeotic gene (EMBL: J03177)</td>
<td></td>
<td>286</td>
<td></td>
</tr>
<tr>
<td>YCR57c</td>
<td>221,442-220,126</td>
<td>Human transducin beta 2 chain, GTP-binding (GenBank: M36429)</td>
<td></td>
<td>266</td>
<td></td>
</tr>
<tr>
<td>YCL11c</td>
<td>102,964-101,684</td>
<td>Xenopus poly(A)-binding protein (PIR: S12000)</td>
<td></td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>YCR107w</td>
<td>312,620-313,708</td>
<td>Tobacco auxin-induced protein (EMBL: X56269)</td>
<td></td>
<td>218</td>
<td></td>
</tr>
</tbody>
</table>

(b) Newly discovered ORFs showing significant similarity to other *S. cerevisiae* proteins

<table>
<thead>
<tr>
<th>ORF</th>
<th>Coordinates</th>
<th>Rational gene name</th>
<th>Similar sequence (database: accession no.)</th>
<th>FASTD score</th>
<th>Effect of gene disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCL25c</td>
<td>77,880-75,982</td>
<td>General amino-acid permease, GAP1 (EMBL: X52633)</td>
<td></td>
<td>1,527</td>
<td></td>
</tr>
<tr>
<td>YCL64c</td>
<td>16,899-15,790</td>
<td>L-serine dehydratase, SDH1 (EMBL: X52657)</td>
<td></td>
<td>984</td>
<td></td>
</tr>
<tr>
<td>YCL48c</td>
<td>42,140-43,528</td>
<td>Sporulation-specific SPS2 protein (EMBL: M13629)</td>
<td></td>
<td>941</td>
<td></td>
</tr>
<tr>
<td>YCR67c</td>
<td>238,046-231,852</td>
<td>Sec12p membrane glycoprotein (GenBank: X13161)</td>
<td></td>
<td>889</td>
<td></td>
</tr>
<tr>
<td>YCL24w</td>
<td>79,119-81,566</td>
<td>Sucrose non-fermenting SNF1 protein kinase (EMBL: M13971)</td>
<td></td>
<td>650</td>
<td></td>
</tr>
<tr>
<td>YCR8w</td>
<td>128,072-129,880</td>
<td>Nitrogen permease reactivator, NPR1</td>
<td></td>
<td>445</td>
<td></td>
</tr>
<tr>
<td>YCR52w</td>
<td>213,725-215,173</td>
<td>Gene complementing petite type mutation</td>
<td></td>
<td>434</td>
<td></td>
</tr>
<tr>
<td>YCR1c</td>
<td>208,342-206,870</td>
<td>Proteinase A precursor (EMBL: M18097)</td>
<td></td>
<td>429</td>
<td></td>
</tr>
<tr>
<td>YCR91w</td>
<td>273,138-275,315</td>
<td>Protein kinase, YKR (EMBL: M24929)</td>
<td></td>
<td>397</td>
<td></td>
</tr>
<tr>
<td>YCL35c</td>
<td>61,142-60,813</td>
<td>Glutaredoxin (PIR: A35492)</td>
<td></td>
<td>396</td>
<td></td>
</tr>
<tr>
<td>YCR73c</td>
<td>245,320-241,379</td>
<td>Protein kinase, Ssp31 (PIR: J11118)</td>
<td></td>
<td>321</td>
<td></td>
</tr>
<tr>
<td>YCR83w</td>
<td>258,312-258,692</td>
<td>Thiodoxinedoxin II (EMBL: M59169)</td>
<td></td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>YCR28c</td>
<td>171,116-169,580</td>
<td>Allantocitrinate permease (EMBL: M24098)</td>
<td></td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>YCR93w</td>
<td>256,168-270,994</td>
<td>A allantocitrinate core subunit, AGA1 (EMBL: M28164)</td>
<td></td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>YCR38c</td>
<td>197,967-196,354</td>
<td>Cell division cycle, Cdc25 protein (EMBL: M15458)</td>
<td></td>
<td>201</td>
<td></td>
</tr>
</tbody>
</table>

Full details of similarity search procedures are given in the legend to Fig. 2. For each ORF, the highest optimum FASTD 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 FASTD score >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.

44
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 the bif and clst gene, 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 chromosomes30 in yeast, as well as the Kaback rule,31 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 determinations was very high, the mismatch frequency being 0.4 kb for clones from the same strain and 6.2 kb 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 (J. Collins and C.S.N., unpublished results). Of 504 6-bp restriction sites sampled, our map 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 genome 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 other higher organisms.

References