The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV


The complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome IV has been determined. Apart from chromosome XII, which contains the 1–2 Mb rDNA cluster, chromosome IV is the longest S. cerevisiae chromosome. It was split into three parts, which were sequenced by a consortium from the European Community, the Sanger Centre, and groups from St Louis and Stanford in the United States. The sequence of 1,531,974 base pairs contains 796 predicted or known genes, 318 (39.5%) of which have been previously identified. Of the 478 new genes, 225 (28.3%) have homologous to previously identified genes and 253 (32%) have unknown functions or correspond to spurious open reading frames (ORFs). On average there is one gene approximately every two kilobases. Sequencing on an alternating regional variations in G+C composition, there is a large central domain with a lower G+C content that contains all the yeast transposon (Ty) elements and most of the tRNA genes. Chromosome IV shares with chromosomes II, V, XII and XV some long clustered duplications which partly explain its origin.

The technique of determining the DNA sequence of large genomes has been unchanged for 21 years. Sequencing the yeast genome required considerable organization by the European Union, which initiated the grouping of 35 laboratories to sequence the first yeast chromosome. It was split into three parts, which were sequenced by a consortium from the European Community, the Sanger Centre, and groups from St Louis and Stanford in the United States. The sequence is 1,531,974 base pairs containing 796 predicted or known genes, 318 (39.5%) of which have been previously identified. Of the 478 new genes, 225 (28.3%) have homologous to previously identified genes and 253 (32%) have unknown functions or correspond to spurious open reading frames (ORFs). On average there is one gene approximately every two kilobases. Topology of alternate regional variations in G+C composition, there is a large central domain with a lower G+C content that contains all the yeast transposon (Ty) elements and most of the tRNA genes. Chromosome IV shares with chromosomes II, V, XII and XV some long clustered duplications which partly explain its origin.

The average base composition of chromosome IV is 37.9% G+C, which is lower than for most of the yeast chromosomes (for example, 38.5% for chromosome III (ref. 2) and 38.3% for chromosome II (ref. 5)). Along the 1,513,914 base pairs of the chromosome there are alternating regions about 50 kilobases long of high and low G+C content. The central domain of chromosome IV (coordinates 500,000 to 1,215,000) has a much lower G+C value (37.4%) than the two flanking regions in which no Ty elements are found. A total of 27 tRNA genes are localized on chromosome IV (ref. 8).

The low G+C content of the central domain seems to be correlated with the presence of Ty elements. All nine Ty1 or Ty2 elements, including a truncated form of Ty2, are localized between coordinates 450,000 and 1,190,000 (Fig. 1). Yeast transposons seem to insert into specific chromosomal regions where they are localized preferentially upstream of tRNA genes, as they might interact with the RNA polymerase III machinery. The density of tRNA genes in the central domain of chromosome IV is twice that in the flanking regions in which no Ty elements are found. A total of 27 tRNA genes are localized on each strand of the chromosome, 17 of which are located in the central domain. Of the 27 tRNA genes, 18 are in the vicinity of long terminal repeats (LTRs). Thus most of the tRNA genes, LTRs and Ty elements...
and LTR/Ty clusters

Figure 1 Overall molecular architecture of chromosome IV shows positions of tRNA genes, solo LTR or Ty elements (thin vertical lines), or clusters of them (thick vertical lines), along the chromosome map. Panels show variation of gene density (top) and base composition (bottom) along the sequence-based map of chromosome IV (scale in kilobases from the left telomere). Vertical broken lines indicate the centromere. Gene density is expressed as the probability for each nucleotide to be part of an ORF, together with a lower G+C content, are found in this central domain. In chromosome II the 13 tRNA genes and three Ty elements are in AT-rich regions.

The left telomere of chromosome IV is very similar to other yeast telomeres. Adjacent to the C1 A repeat are the usual STR-A, STR-B, STR-C, STR-D and the core X elements (435–904) shared by most of the telomeres. The left end of chromosome IV shares with the right end of chromosome X a large, nearly identical block of sequence, similarly more than 19 kilobases long. This duplication includes five ORFs, which code for almost identical products. Indeed the sequences are so similar that we needed to exclude the possibility of contamination of the cosmids contig of chromosome IV by DNA sequences from chromosome X. To confirm our data, we established the genomic sequence of the junction between the duplicated sequence and the rest of the chromosome. Such subtelomeric duplications have often been observed in the yeast genome, suggesting either recent or continuous exchange of genetic information. The right telomere has a less conventional structure with an internal TG1 repeat.

Using the classical definition of ORFs (one ATG codon followed by at least 99 sense codons), 776 ORFs were recorded in the chromosome; there are also 20 ORFs shorter than 99 amino acids long, making a total of 796 ORFs. Small ORFs of between 25 and 99 codons were extracted and analysed for different properties (codon usage, homologies and ATG environment) to determine their function. Although some ORFs are thought to be ‘questionable’ or ‘orphan’ or ‘orphan’ sequence similarities, roughly 30% of the 796 ORFs of chromosome IV are proteins of the actin family, three for proteins involved in the ubiquitin-dependent protein degradation system, and the rest are distributed between genes that do not necessarily have a high CAI.

...OrF was the first to be described, but many other candidates for common regulation have been revealed by the systematic genome sequencing...
that are very similar in both coding and non-coding regions\(^\text{13}\). Finally, the DNA strand distribution of the ORFs is different in the central region when compared with that of the flanking regions. The ORF arrangement of this region might result from a greater genetic plasticity.

Analysis of structural relationships inside the yeast genome might provide an insight into eukaryotic genome organization and evolution. Redundancy is one of the most salient features of the yeast genome structure\(^\text{20}\), and the DNA sequence of the whole yeast genome reveals several types of redundancy, probably originating from different biological processes. The most common form of redundancy involves individual genes that have a homologue in the genome; about 20% of the genes of chromosome IV are in this class. Second, there are clusters of very similar copies of a gene, often arranged in tandem; for example, there are five copies of the ENA1 (or PM R2) gene on chromosome IV. Third, subtelomeric duplications are frequent and involve large regions of chromosomes that are very similar in both coding and non-coding regions\(^\text{21}\). Finally, clustered duplications are characterized by clusters of homologous genes in the same order, usually in the same orientation, and interspersed by long DNA fragments. Such paralogous regions have already been described between chromosomes III and XIV (ref. \text{25}) and between chromosomes V and X (ref. \text{26}). Only in the case of the duplication between chromosomes III and XIV is the gene order conserved. It is 15 kilobases long and contains four genes. The clustered duplications on chromosome IV are made up of at least 336 kilobase pairs, including 49 pairs of homologous genes. Chromosome IV shares large ordered clusters of homologous genes with chromosomes II, V, VIII, XII and XIII (ref. \text{27}). A careful analysis of these duplications will no doubt tell us a great deal about the evolution of the yeast genome. In the largest interchromosomal clustered duplications, involving chromosome IV (coordinates 449,752–569,763) and chromosome II (238,164–407,122), the 18 gene pairs are all transcribed in the same direction. When known, most of the genes from a pair code for proteins with homologous but not identical functions (for example GAL1 and GAL3)\(^\text{28}\). Homologous genes from a clustered duplication can also be completely identical or totally different in their function. An extreme case of divergence involves YDR037w, which codes for a lysyl tRNA synthetase, and YBR060w\(_A\), which is part of chromosome II, in which many stop codons interrupt an ORF, of which parts are homologous to YDR037w. Such a pseudogene could not be detected by searching the DNA sequence of chromosome II, as it has very short ORFs and no ATG codon. To our knowledge, this is the most degenerated yeast pseudogene yet discovered. This observation suggests that similar degenerated pseudogenes may have escaped previous analyses, and hence that the total number of pseudogenes may be underestimated.

A pair of genes from a clustered duplication can also differ in the presence of an intron. Both YDR055w and YBR078w are homologous to the gene SP52, but only YBR078w has an intron, and the CAI of the two genes differs from 0.27 (YDR055w) to 0.61 (YBR078w), suggesting an unusual evolutionary process. The compared analysis of the interspersed DNA fragments is also very informative. For example, chromosome II has a Ty element where an LTR element is present at the equivalent position on chromosome IV, suggesting that the Ty element was lost from chromosome IV after the duplication process.

The greatly different degrees of similarities between the different gene pairs composing a duplicated region indicate that at least some of the duplications have evolved at very different rates, suggesting in some cases that gene conversion processes\(^\text{29}\) have interfered with slower evolutionary processes. A careful quantitative analysis of the relative evolution rates of the different elements will be required to establish a chronological order of the different events. Nevertheless, evidence suggests that a first duplication event has been followed by the dispersal of the duplicated elements by the insertions of DNA fragments of various sizes and gene compositions. Most of these clustered duplications in chromosome IV are localized in the pericentromeric region. The centromere itself is included in the longest duplicated region, which occurs between chromosomes IV (coordinates 450,000–570,000) and II (238,000–407,000). Similar localizations of clusters have already been noticed on other chromosomes\(^\text{3}\)\text{13,23}. To explain the proximity of the centromere, it was suggested that the gene dispersion of the initial cluster of duplicated genes might be slower in the centromeric regions than in the telomeres owing to the adverse effects of rearrangements on chromosome segregation\(^\text{32}\). Alternatively, the centromeric duplications might have been essential steps in the construction of the yeast genome\(^\text{2}\). In agreement with these ideas, the central domain of chromosome IV contains few traces of clustered redundancies, perhaps because of its genetic plasticity. These preliminary observations indicate that the availability of the complete sequence of the yeast genome will allow a greater understanding of the processes involved in creating the genome architecture.

Methods

The sequence was assembled from a set of 44 partly overlapping cosmids and lambda phages from two independent contigs of chromosome IV. The 650-kb cosmids corresponding to the left part of chromosome IV was constructed mainly from a specific cosmid library obtained from a gfp-purified chromosome portion (J. D. H., et al., unpublished), and a few other cosmids cloning from this contig came from a library\(^\text{3}\). The rest of the chromosome sequence was established from a cosmids lambda phage library (L. Riles & M. Olson, unpublished, and ref. \text{30}). The two cosmid contigs were made from two closely related yeast strains (AB972, derived from S288C\(_A\), and FY1679, a diploid strain isolated from the cross between FY23 and FY73, both of which are isogenic with S288C except for the markers indicated. Sequence analysis of a large overlapping fragment (of about 170 kb) confirmed that the strain AB972 and FY1679 are very similar, as the number of base differences was below the estimated error rate. Only the extreme left telomeric regions of the two strains clearly differ, probably in the number of their TG1\(_A\) repeats (C. B. and J. C., unpublished). The telomeres were isolated independently and sequenced from a plasmid clone generated by integration at the TG1\(_A\) repeats of the telomere, followed by excision of the plasmid and capture of the flanking sequences\(^\text{3}\). Two gaps in the 650-kb left cosmids contig (constructed from the strain FY1679) were filled with lambda clones from the library constructed from AB972. They correspond to the regions 9,756–11,360 and 363,100–368,150. The left 600-kb region was sequenced according to the rules followed by the European consortium and the 20 cosmids and phages were distributed to 18 contractors, whereas the central part was sequenced by the Sanger Centre (EMBL database SCCHRIV, accession no. Z71256), the rest of the chromosome was sequenced by groups from Washington University in St. Louis and Stanford University in the United States.

There were very few base differences in the overlapping fragment sequenced in parallel by the Sanger Centre and by the European consortium, demonstrating that both approaches are reliable. However, a verification procedure was necessary because of the greater heterogeneity of the European approach. This was done on 25 regions of the left part of the chromosome, according to the protocol of G. V. (manuscript in preparation). This allows direct polymerase chain reaction (PCR) sequencing of a 300-bp region of the yeast genome limited by two previously designed oligonucleotides. We could thus correct a sequence in which a bacterial transposon had been inserted during the cloning process but no real sequence error could be detected at this final step of the sequencing project. Sequence errors could only be corrected after examination of the raw sequence data. From these data, the error rate of this part of the yeast chromosome IV sequence presented is less than four errors per 10 kb. In the central part of the chromosome the error rate is estimated as less than one error per 10 kb. Specific strategies were developed to sequence difficult parts of the chromosome IV. Thus, for example, to finish the regions between the two transposons located in cosmid BL42 (http://www.sanger.ac.uk/~yeastpub/sw/sequencing.html), a PCR product covering this region from strain FY1679 was filled with lambda clones from the library constructed from AB972. They correspond to the regions 9,756–11,360 and 363,100–368,150. The left 600-kb region was sequenced according to the rules followed by the European consortium and the 20 cosmids and phages were distributed to 18 contractors, whereas the central part was sequenced by the Sanger Centre (EMBL database SCCHRIV, accession no. Z71256), the rest of the chromosome was sequenced by groups from Washington University in St. Louis and Stanford University in the United States. There were very few base differences in the overlapping fragment sequenced in parallel by the Sanger Centre and by the European consortium, demonstrating that both approaches are reliable. However, a verification procedure was necessary because of the greater heterogeneity of the European approach. This was done on 25 regions of the left part of the chromosome, according to the protocol of G. V. (manuscript in preparation). This allows direct polymerase chain reaction (PCR) sequencing of a 300-bp region of the yeast genome limited by two previously designed oligonucleotides. We could thus correct a sequence in which a bacterial transposon had been inserted during the cloning process but no real sequence error could be detected at this final step of the sequencing project. Sequence errors could only be corrected after examination of the raw sequence data. From these data, the error rate of this part of the yeast chromosome IV sequence presented is less than four errors per 10 kb. In the central part of the chromosome the error rate is estimated as less than one error per 10 kb. Specific strategies were developed to sequence difficult parts of the chromosome IV. Thus, for example, to finish the regions between the two transposons located in cosmid BL42 (http://www.sanger.ac.uk/~yeastpub/sw/sequencing.html), a PCR product covering this region from strain FY1679 was sequenced. This strain, a gift from E. Louis, did not contain the transposons.
Here we report the sequence of 569,202 base pairs of *Saccharomyces cerevisiae* chromosome V. Analysis of the sequence revealed a centromere, two telomeres and 271 open reading frames (ORFs) plus 13 tRNAs and four small nuclear RNAs. There are two Ty1 transposable elements, each of which contains an ORF (included in the count of 271). Of the ORFs, 78 (29%) are new, 81 (30%) have potential homologues in the public databases, and 112 (41%) are previously characterized yeast genes.

As part of an international collaborative effort to sequence the total genome of the yeast *Saccharomyces cerevisiae*, we have deduced the DNA sequence of 569,202 base pairs of yeast chromosome V. We used an overlapping set of recombinant yeast cosmid and lambda clones that together cover the entire chromosome (except for the extreme ends of the telomeres). A line drawing of chromosome V and the identification of the recombinant DNAs sequenced are shown in Fig. 1. The sequence was broken arbitrarily into 11 slightly overlapping pieces for ease of handling and deposited in Genbank (see Fig. 1 for accession numbers).

Sequencing was accomplished in two phases: the ‘shotgun’ phase, using dye-primer chemistry, and the ‘finishing’ phase, using the polymerase chain reaction (PCR) and dye-terminator chemistry. There were no gaps in the sequence at the end of shotgun sequencing and assembly. The assembled continuous sequence of chromosome V has 569,202 bp, starting from the guanine residue of the Sau3A site on the left vector boundary of the leftmost clone (1160 in Fig. 1). The 569-kilobase sequence results from the reads 32,631 individual lanes of sequencing gels, or reads. The average depth of coverage was 12.5-fold. The minimum acceptable coverage was three, with at least one read from each strand.

After shotgun sequencing and assembly, problems remained in the sequence at a frequency of (roughly) two per kilobase and were of several types. They included the inability to count unambiguously the number of repeating units, such as poly (dA), and guanine compressions. There were also small regions in which only one of the two strands had been sequenced. These difficulties were resolved during the finishing phase.

After finishing, the 569-kb contig was checked against three external sets of data. First was the use of tetra segregation data to derive a genetic map for yeast1. The chromosome V gene order based on DNA sequence was in complete agreement with the tetrad segregation data. There were two locations on the genetic map (CEN V at 151 kb and PRO3 at 200 kb) where closely spaced loci had been mapped against distant markers and not against each other, resulting in ambiguities of relative locus order2,3, which were resolved using the DNA sequence. The gene order across the centromere is GLC3 rRNA-Arg GCN4 CEN V MNN1. In the region of PRO3, at 200 kb, the gene order is PRO3 GPA2 GCP11 CH01 GAL83. Second, our sequence was compared to the S. cerevisiae sequences already deposited in Genbank, using both the FASTA and BLAST programs2,3. In the rare cases of sequence difference, we re-examined our trace files. Remaining ambiguities were resolved using the same methodology as finishing. Third, we checked our data against the primary EcoRI/HindIII double-digestion fragment maps of the recombinant yeast DNA2,3. Our sequence was examined for EcoRI and HindIII cleavage sites. Of 534 mapped fragments, there were only five discrepancies, which is a tribute to the care taken in preparing the cleavage site maps2,3. The five apparent discrepancies between the double-digest map and our sequence are the map had doubles where the sequence predicts singletons after bases 272, 193; 280,936; and 441,102; the map has a fragment that was not found in the sequence after base 414,946; and the sequence is missing a cleavage site after base 506,807.

We examined all six possible reading frames of the 569-kb sequence for ORFs of at least 300 bp that began with a start codon and ended with a stop codon. As a special case, an ORF could be interrupted if there were splice donor/acceptor branchpoint sequences present at the appropriate intervals. The remaining sequence was examined using FASTA and BLAST for homology to sequences in the public databases. This enabled us to find small ORFs, as well as the centromere, 13 tRNAs, two Ty1 elements (which each contain an ORF), four small nuclear RNAs, many delta and delta-like elements, and the highly conserved X and Y sequences characteristic of yeast telomeres (see refs 5, 6) at the far left and right ends.

Initially, 271 ORFs were identified in the 569-kb sequence, although this number has changed as evaluation continued. The 271 ORFs make up roughly 70% of the sequence, with an average of 2.1 kb per ORF. The ‘average’ ORF (1.4 kb) encodes 475 amino acids. Of the ORFs, 112 (41%) have been characterized previously, 81 (30%) have apparent homologues in the public databases, and 78 (29%) are new; six (2%) are spliced. Of the 81 apparent homologues, 55 of these are to other S. cerevisiae sequences.

The fractional G+C content of the 569,202 bp of chromosome V is 0.384. The combined ORF DNAs have a fractional G+C content of 0.401, and the combined ‘non-ORF’ DNA has a G+C content of 0.351.