Volume 12 Number 19 1984 Nucleic Acids Research

# Structure and organization of two linked ribosomal protein genes in yeast

C.M.T.Molenaar, L.P.Woudt, A.E.M.Jansen, W.H.Mager and R.J.Planta
Biochemisch Laboratorium, Vrije Universiteit, de Boelelaan 1083, 1081 HV Amsterdam,
The Netherlands

D.M.Donovan and N.J.Pearson
Department of Biological Science, University of Maryland Baltimore County, Catonsville,
MD 21228, USA

Received 18 July 1984; Accepted 13 September 1984

## ABSTRACT

The genes encoding yeast ribosomal proteins rp28 and S16A are linked and occur duplicated in the yeast genome. In both gene pairs the genes are approximately 600 bp apart and are both transcribed in the same direction. Both ribosomal protein genes resemble other ribosomal protein genes studied so far in many structural aspects. The genes are interrupted by an intron near the 5'-end of their coding sequence. In addition the flanking regions contain several conserved sequence elements, which may function in transcription initiation and termination.

In agreement with findings concerning other cloned yeast ribosomal protein genes, upstream homology blocks occur that may be involved in coordinate control of ribosomal protein gene transcription. The complete pattern of conserved and diverged sequences between the two duplicate gene pairs is presented.

## INTRODUCTION

Yeast ribosomes are composed of 4 ribosomal RNAs and about 75 ribosomal proteins. For an efficient assembly of all these ribosomal constituents into ribosomal particles, under various conditions of cell growth a simultaneous and equimolar production of these components is necessary (1,2). Thus ribosome biosynthesis requires a differential and balanced expression of a great number of different genes. To elucidate the molecular mechanisms involved in the coordinate synthesis of ribosomal proteins in yeast several ribosomal protein genes have been isolated by molecular cloning (see Ref. 1 for a review).

From the characterization of DNA fragments carrying ribosomal protein genes it had become apparent that most ribosomal protein genes in yeast are not clustered (3,4,5). In addition, many ribosomal protein genes turned out to be duplicated (3,5).

This paper deals with two ribosomal protein genes, encoding the large subunit protein rp28 and the small subunit protein S16A (rp55), which are exceptional in that they are physically linked. Strikingly, the duplicate

© IRL Press Limited, Oxford, England.

genes are also adjacent to each other suggesting a functional linkage of these ribosomal protein genes. Therefore we decided to investigate the structure and organization of these linked ribosomal protein genes.

Studies of cloned ribosomal protein genes carried out thus far have revealed some common structural features. Most, but not all ribosomal protein genes contain an intron of 300-500 bp near the 5'-end of the coding sequence (3, 5-10). The intron-exon boundaries show conserved sequences similar to the splice sites found in higher eukaryotes. In addition the introns contain near the 3'-end a conserved sequence, TACTAACA, which probably is involved in splicing of the mRNA (11-13).

Computer analysis of the DNA sequences upstream from a number of yeast ribosomal protein genes revealed several homologous sequence elements which may be involved in the regulation and/or expression of these genes (10). As will be demonstrated in this paper the genes coding for rp28 and S16A contain many of the common characteristics of yeast ribosomal protein genes. Apart from that, the duplicated gene pairs show some features which may be unique to these ribosomal protein genes.

## MATERIALS AND METHODS

### Nomenclature

Yeast ribosomal protein rp55 (14) is identical with S16A in the standard numbering system based on Kaltschmidt and Wittmann 2D-gels (15); rp28 cannot be identified in this gel system. However, rp28 is resolved on the gel system of Gorenstein and Warner (3,16). We have chosen to use the rp28-S16A nomenclature throughout this paper.

Genes isolated from the yeast strains *Saccharomyces carlsbergensis* and *Saccharomyces cerevisiae* are referred to as *S.ca* and *S.ce* genes, respectively.

### Yeast strains

Yeast strains used for the isolation of DNA were *Saccharomyces carlsbergensis* (NCYC74) and *Saccharomyces cerevisiae* A364A (ATCC22244).

### Isolation of genes

The initial isolation of the first-copy linked genes from *S.ce* and *S.ca* have previously been described (3,17). To isolate the second-copy genes for rp28 and S16A of *S. carlsbergensis* a library was constructed using cosmid pHC79 (Boehringer Mannheim) as a vector. Yeast DNA was isolated as described by Verbeet *et al.* (18), partially digested with restriction endonuclease Sau3A and size-fractionated by sucrose-gradient centrifugation (19). Appropriate fractions of this yeast DNA were treated with bacterial alkaline phosphatase and then ligated with a mixture of left-handed and right-handed vector ends (incapable of self-ligation; 20). After packaging in λ phage particles (19) the recombinant DNA was transduced to Escherichia coli 1046 (kindly provided by Dr. G.J.B. van Ommen). Colony screening was performed as described previously (20), using a 2.3 kb HindIII-generated fragment isolated from pBMCY44 as a probe. This fragment carries the genes encoding rp28 and S16A and was labelled in vitro by nick translation (21).

Appropriate fragments of the selected recombinant cosmid and of the recombinant bacteriophage A12 which carries the S.ce gene pair (3) were subcloned into pBR322.

Recombinant plasmids were purified from Triton-treated bacterial spheroplasts by CsCl ethidium bromide density gradient centrifugation (22).

### Heteroduplex analysis and electron microscopy

Heteroduplex analysis and electron microscopy were performed essentially as described by Verbeet et al. (18).

### DNA sequence analysis

DNA sequence analysis was performed using the chain termination method (23). Single-stranded templates were obtained by transforming JM101 or JM103 cells with recombinant bacteriophage M13 RF DNA (24). The M13 vectors used were M13 mp8, mp9, mp10, and mp11 (25).

### Restriction site mapping

Restriction endonucleases were obtained from Bethesda Research Laboratories, New England Laboratories or Boehringer (Mannheim) and used as recommended by the supplier. Cleavage sites for AluI, HpaI, MspI (HpaII) and TaqI were determined by the partial digestion procedure of Smith and Birnstiel (26).

### Southern blot analysis

Total yeast DNA was isolated from both S. carlsbergensis and S. cerevisiae and digested with EcoRI or HindIII. The resulting fragments were electrophoresed on 1% agarose gels, transferred onto nitrocellulose (27) and hybridized with $^{32}$P-labelled DNA fragments as described previously (6).

## RESULTS AND DISCUSSION

### Southern analysis

The linked genes coding for ribosomal proteins rp28 and S16A have been isolated from a colony bank of HindIII-generated S.ca DNA fragments in pBR322 as described previously (17). The genes have been mapped on plasmid pBMCY44 by electron microscopic R-loop analysis (17) as shown in Fig. 1A. To determine the copy number of the genes coding for the two ribosomal proteins, genomic

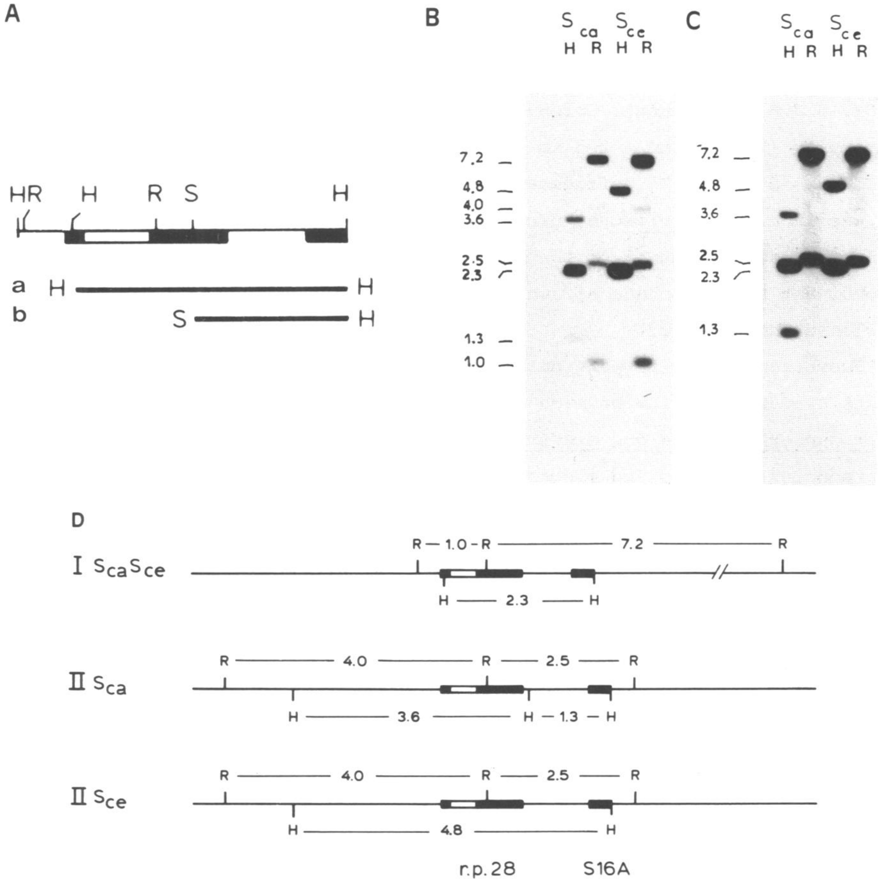

Fig. 1. Copy number analysis of the genes encoding rp28 and S16A.
In A the location of the S.ca rp28 and S16A genes on the physical map of the insert of recombinant pBMCY44 is depicted (17). Symbols for restriction sites are H (HindIII), R (EcoRI) and S (SacI). The intron in the rp28 gene is indicated by an open bar. The results of the hybridization of $^{32}$P-labelled fragments a and b (panel A) to S.ca and S.ce DNA digested with HindIII (H) or EcoRI (R) are shown in panel B and C, respectively. I ScaSce is the region in the genome corresponding with the insert of pBMCY44 as well as A12 in panel D. IISca and IISce are the regions of the genome corresponding to duplicate copies of rp28 and S16A in the two strains based on Southern analysis and EM R-loop analysis. Note that the 5'small exon of S16A cannot be detected by R-loop analysis but can be demonstrated by heteroduplex and sequence analysis (Fig. 3 and 4).

Southern analysis was performed using a 2.3 kb HindIII-generated fragment as well as a 1.3 kb (SacI plus HindIII)-generated fragment from pBMCY44 as a probe. The autoradiograms showing the result of hybridization to genomic digests of both S.ca and S.ce DNA are presented in Figs. 1B and 1C. It can be

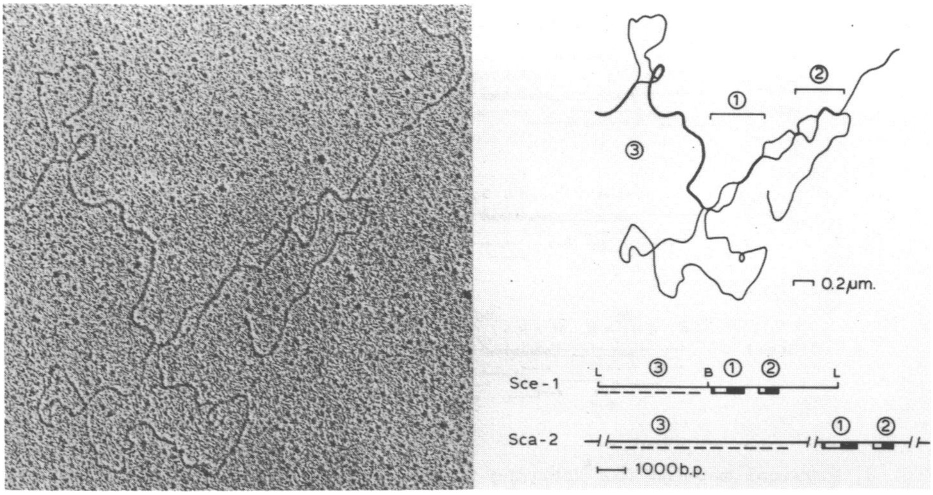

Fig. 2. Heteroduplex analysis of duplicate genes encoding rp28 and S16A. Recombinants *S.ce*-1 and *S.ca*-2 carry duplicate gene pairs coding for rp28 and S16A. *S.ce*-1 was linearized by cleavage with SalI; the 40 kb recombinant cosmid *S.ca*-2 was linearized by random shear. Only amino acid-coding regions of the rp28 (1) or S16A (2) gene and regions of homologous vector DNA (3; pBR322 and pHC79) form stable duplex structures. Similar results were obtained in the comparison of the two copies of the gene pair isolated from *S. carlsbergensis*.

concluded that portion of both genes is duplicated in the genome of both yeast strains and, in addition, that most probably the genes are also adjacent at the second locus. This agrees with the published results for *S.ce* (3). The latter finding suggests a functional linkage of the genes coding for S16A and rp28. The interpretation of the hybridization results described above is illustrated in Fig. 1D.

The duplicate gene pair was isolated from a cosmid bank as described in the Materials and Methods section. Isolation of the gene pair from *S. cerevisiae* that corresponds with the insert of pBMCY44 has been described previously (3). The *S. cerevisiae* clone carrying copy 1 of rp28 and S16A is designated A12.

We performed a comparative analysis of the structure and organization of the genes coding for rp28 and S16A to get insight into the pattern of conserved and diverged sequences in both the duplicate gene pairs and the two different yeast strains.

### Heteroduplex analysis

First heteroduplex analysis under the electron microscope was performed. On

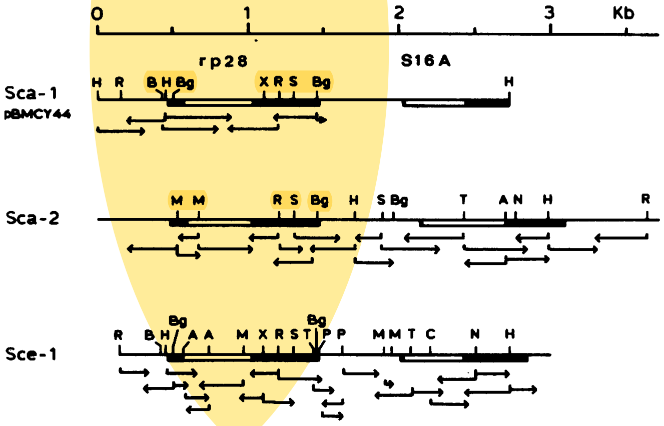
0
1
2
3
Kb
rp28
S16A
Sca-1
pBMCY44
H R B H Bg X R S Bg H
Sca-2
M M R S Bg H S Bg T A N H R
Sce-1
R B H Bg A A M X R S T Bg P P M M T C N H

Fig. 3. Physical maps of DNA carrying the rp28 and S16A genes and the DNA sequencing strategy.
Symbols for restriction endonucleases are A (AluI), B (BamHI), Bg (BglII), C (BclI), H (HindIII), M (MspI, HpaII), N (HincII), P (HpaI), R (EcoRI), S (SacI), T (TaqI) and X (XbaI). Introns and exons are indicated by open and solid bars, respectively.
S.ca-1 Map of the yeast DNA insert of pBMCY44 (17).
S.ce-1 Map of part of the insert of the recombinant phage A12 (3).
S.ca-2 Map of part of the recombinant cosmid carrying the second-copy genes for rp28 and S16A (this paper).
The arrows indicate the direction and extent of nucleotide sequence analysis.

this level of examination the corresponding DNA fragments from S.ce and S.ca turned out to be completely homologous (results not shown). Heteroduplex analysis of the duplicate gene pairs, however, indicated that, apart from the presumed coding regions, they have very little sequence homology (see Fig. 2). The electron micrograph suggests that both genes are interrupted by an intron (as has been established previously for the rp28 gene [17]). Unlike the exons, which are highly conserved and therefore form stable heteroduplexes, the intervening and flanking sequences including the intergenic region show a high degree of divergence since no stable duplex structures of these regions can be observed. Since it is a general characteristic of yeast ribosomal protein genes studied so far that these genes are split by an intron near the 5'-side of the coding sequence, the heteroduplex analysis predicts a head to tail arrangement of these genes.

Sequence comparison: S.ca vs S.ce

The respective restriction maps of the cloned recombinant DNAs are presented in Fig. 3. Correlation of the physical maps with the results of the genomic Southern analysis (cf. Fig. 1D) confirmed our preliminary conclusion with

```
                    -300
1                           T  T      ACT                     A                      T                   90
   TTCATAATTT CGAAGACTGT TTT TA ATA TC   TGTGG TGTACGGATA TGAA TTTTT TTACCGAAGA CAATA TGCA TGGGTGTTAA
2                           C  C      TTA                     -                      A                   89
   ****** *   **   *     *          *      **    *    *
3  TTCATATTCA CGCCAAGAAA TCAGGCTGCT TTCCAAATGC AATTGACACT ---------- ---------- ---------- ----------    50
         -290

                                -200
1  A                                                                                 T                   173
    TCTATTTTT TTTTTTTTCT AATTCAGAAT GAAATTTTTT TCTTGCTGGA GATTACAGTG GAAATT---- ---TC CTCC AACACACGCA
2  T                                                                                 G                   172
                  *  *           *   ***  * ** ********** * **    *   ***          *    *  *****
3  ---------- ----TCATTA GCCATCACAC AAAAGCTCTT TCTTGCTGGA GCTTCTTTTA AAAAAGACCT CAGTACACCA AACACGTTAC    126
                                                                     -200

                                                                     -100
1                                                                                                        247
   CTCTCTCTGA CTTGGTGCTG ------CCGT GAGATTCTGG TTCCTCAAGG AA-------- --TATTGTAT GAGATAATTT AAAACAAATT
2                                                                                                        246
   *   *   *    *  * *            *  * ******        **   **          **       *   * ****  *** **
3  CCGACCTCGT TATTTTACGA CAACTATGAT AAAATTCTGA AGAAAAAATA AAAAAATTTT CATACGTCTC GCTTTTATTT AAACCATTGA    216
                                                                                -100

1                                  A                                                                     325
   GAGTTGAAGA TTCATCAATT CTGCCATTT  GGATCC---- --ATCACAAG AAGAGATCAT CAAG----CT TACGGATTCA CA   ATG GGT
2                                  T                                                                     324
     * *      ** * ***  **  * * *    * *        **   ***  ** * ****    **            *  **  *   *** ***
3  ATGATTTCTT TTGAACAAAA CTNGCCTGTT TCACCAAAGG AAATAGAAAG AAAAAATCAA TTAGAAGAAA ACAAAAAACA AA   ATG GGT  304
                                                                                                met gly
                                                                                                 1

1                                                                                                        400
   ATC GAT CAC ACT TCC AAG CAA CAC AAG AGA TCT GGT CAC AGA ACT GCT CCA AAG TCT GAC AAT GTC TAC TTG AAA
2                                                                                                        399
   ** *** *** *** *** *** *** *** **  *** **  *** **  *** **  *** *** **  **  **  **  *** **  *** ***
3  ATT GAT CAC ACT TCC AAG CAA CAC AAA AGA TCC GGT CAT AGA ACC GCT CCA AAA TCC GAT AAC GTC TAT TTG AAA   379
   ile asp his thr ser lys gln his lys arg ser gly his arg thr ala pro lys ser asp asn val tyr leu lys
                                                                       20

1                                                          A                                             481
   TTG TTA GTC AAA TTA TAC ACT TTC TTA GCT C   GTATGTTCAA  AATAATATC AAGGG-TTTT AACCAACGCC AATTATGGTT
2                                                          G                                             480
   **  **  *** ***  ** *** *** ***  ** *** *   ******       **   *  *       ****              **  ** *
3  TTA TTG GTC AAA CTA TAC ACT TTC CTA GCT C   GTATGTATT- GAAATCTCCC TCCCAATTTT C--------- --TTCAGGAT    448
   leu leu val lys leu tyr thr phe leu ala
                                       37

1                                                                                                        571
   AGAAGGATTC TGTTACATTG AATATAATAG ATTATCCTGA ACTTTGGTAA GTTCATTGGT TTTCTATAAA TATTCGTTAT GGTTTGCTGA
2                                                                                                        570
   * ******** *** * ***      *        *      **  ***       * *      * *
3  AAAAGGATTC TGTGATATTA TTAACTTCTA GTCCCGGCGA TCTTCTAATG GCTTCAGTCT T--------- ---------- ----------    509

1                                                           -             G                              660
   AAAATATTGG TTAATGTTTT TCGCGATCAT CTAGCTTTGC TTCTGATGTG AC AATTATT GGTCA TTTA TTGAACTGAA TCTGGCTAAA
2                                                           A             A                              660
                                  *  ** *       **   * *** *      *  *   *   * ***  **    *    * *     *
3  ---------- ---------- -----AGTAT CGTTGCGTGC GACGGATCTC TTAGACAAAA GTGGAATTTG TTCTCTTTTG TGTTAAAGAT    574

1                                                                                                        749
   ATAGTCAT-A CCAGAAGAAG CTAGACGTAT TAAAATTCTG CGAGTGCAAA CGGGTATATT TGAGGAACGA TATGTTTCTT TTACATCACG
2                                                                                                        749
    *   *   *  *****   *  ***   * *   ******** ***** * *                         * *               *
3  TTCCACTGGA ACAGAAAGCG GTAGTTTTCT GGAAATTCTG CGAGTACTA- ---------- ---------- AAGGAAGTAC AATACGACCT    643

1                                                                                                        823
   TTTTGTGCGG TATTGCAAAC CAGTGAGCAG AATCTTTTTC TAATTTAATG AAATAT---- TTGTACGGAA AACTATTG-- ----------
2                                                                                                        823
   *   ******      * *    * ******* *********        *  * *** *      * ****        * ***
3  TCAAGTGCGG ACCAACTATT AACTGAGCAG AATCTTTTTT CTTCAAAGAG AAAAACAAAT TGGTACTTTT TCGTTTTGCA CTCTTTTTGG    733

1                                                                                                        859
   ---------- ---------- ---------- --------TT TTCCGGTCCT TCGATCCATT AGTTACTAAC AT-------- --------TT
2                                                                                                        859
                                               * **  *    **  *        ******** **                 **
3  CTTCAATTAA TGGAAGAAAT TAAAAGCTGC CTATCGAGAA TACCAATAAA TCCTTTATCC CTTTACTAAC ATAACGTAAA AACTTTTGTT    823

1                                                                                                        937
   TTTCATTTTT TTTTTTTAG GT CGT ACT GAT GCT CCA TTC AAC AAG GTT GTC TTG AAG GCT TTG TTC TTG TCT AAG ATC
2                                                                                                        937
   **** * * * *  *   ** ** *** *** *** *** *** *** *** *** *** **  *** *** *** *** *** *** *** *** ***
3  TTTCTTGTCT TAATCACAG GT CGT ACT GAT GCT CCA TTC AAC AAG GTT GTT TTG AAG GCT TTG TTC TTG TCT AAG ATC   901
                       arg arg thr asp ala pro phe asn lys val val leu lys ala leu phe leu ser lys ile
                        38     40

1                                                                                                        1012
   AAC AGA CCA CCT GTT TCT GTC TCT AGA ATT GCT AGA GCT TTG AAG CAA GAA GGT GCT GCT AAC AAG ACT GTT GTC
2                                                                                                        1012
   *** *** *** *** *** **  *** **  *** *** *** *** *** *** *** *** *** *** *** **  *** *** *** *** ***
3  AAC AGA CCA CCT GTT TCA GTC TCC AGA ATT GCT AGA GCT TTG AAG CAA GAA GGT GCT GCC AAC AAG ACT GTT GTC   976
   asn arg pro pro val ser val ser arg ile ala arg ala leu lys gln glu gly ala ala asn lys thr val val
           60                                                                                80

1                                                                                                        1087
   GTT GTT GGT ACT GTT ACT GAC GAT GCC AGA ATC TTT GAA TTC CCA AAG ACC ACT GTT GCT GCT TTG AGA TTC ACT
2                                                                                                        1087
   *** *** *** **  *** *** *** *** *** **  *** **  *** *** *** *** *** *** *** *** *** *** *** *** ***
3  GTT GTT GGT ACC GTT ACT GAC GAT GCC AGG ATC TTC GAA TTC CCA AAG ACC ACT GTT GCT GCT TTG AGA TTC ACT   1051
   val val gly thr val thr asp asp ala arg ile phe glu phe pro lys thr thr val ala ala leu arg phe thr
                                                                       100
```

```
1                                                                                                        1162
2 GCT GGT GCC AGA GCC AAG ATT GTT AAG GCT GGT GGT GAA TGT ATC ACT TTG GAT CAA TTA GCT GTC AGA GCT CCA    1162
  *** *** *** *** *** *** **  *** *** *** **  *** *** *** *** *** **  *** *** *** *** *** *** *** **
3 GCT GGT GCC AGA GCC AAG ATC GTT AAG GCT GGC GGT GAA TGT ATC ACT TTA GAT CAA TTA GCT GTC AGA GCT CCT    1126
  ala gly ala arg ala lys ile val lys ala gly gly glu cys ile thr leu asp gln leu ala val arg ala pro
                                                  120

1                                                                                                        1237
2 AAG GGT CAA AAC ACT TTG ATC TTG AGA GGT CCA AGA AAC TCC AGA GAA GCT GTC AGA CAC TTC GGT ATG GGT CCA    1237
  *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** **  *** *** *** *** **  *** *** ***
3 AAG GGT CAA AAC ACT TTG ATC TTG AGA GGT CCA AGA AAC TCC AGA GAA GCG GTC AGA CAC TTC GGC ATG GGT CCA    1201
  lys gly gln asn thr leu ile leu arg gly pro arg asn ser arg glu ala val arg his phe gly met gly pro
                              140

1                                                                                                        1312
2 CAC AAG GGT AAG GCT CCA AGA ATC TTG TCC ACC GGT AGA AAG TTC GAA AGA GCT AGA GGT AGA AGA AGA TCT AAG    1312
  *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** ***
3 CAC AAG GGT AAG GCT CCA AGA ATC TTG TCC ACC GGT AGA AAG TTC GAA AGA GCT AGA GGT AGA AGA AGA TCT AAG    1276
  his lys gly lys ala pro arg ile leu ser thr gly arg lys phe glu arg ala arg gly arg arg arg ser lys
          160                                                                             180

1                                                                                                        1384
2 GGT TTC AAG GTG TAA   GTTAACTGA- AATGAAAA-- ---------- TTTCATATTT ACTTTTTTAT TGTTACTCAT TTGTAATTCA     1384
  *** *** *** *** ***     **         *****                *** *  **  ******  * **** *     *
3 GGT TTC AAG GTG TAA   TCTAGTATGG TTTGAAACCT TACAATTTTT CTTCTTTGTT CCTTTTTCCT TGTTTCAGTG TATATTAGGT     1361
  gly phe lys val end
              186

                                              +100
                                                .
2 TAAACTACAT ACACTTTCAA TCGTTTCTTT CAAACTACAT AATTTTTCCT GGCGATCAAT AACGCATTTA GTTCATAAAG TGAGTCAAAG     1474
  *                   *   * ** **  ** ** * **  * *              *        * * *  *******   * *  ***
3 TGGGAA---- -------AGA GGGATTTTTC CATACCATAT GACTGACTAC AATATATACA TGTATAATAA CTTCATAATC TAAACCAACC     1440
                                                               ~~~~~~~~~~
                                              .
                                            +100

                                                          +200
                                        ~~~     ~~~~        .    ~~~~~~~          ~~~
2 TTAACACTAG AATATTTGCT ACACCATCAA TAGGCTAGAC CATAGTTGAA AACTTNCATA ATAAATTCTT CCGTTTTCAA TC-TTCATAT     1563
  *      *                                       * *       *        *****           *   * *  **** *
3 TATCAGTTCA GTATCAAG-- ---------- ---------- TCAACTATTC CGCCCTATGC ATAAACCTAC TAAACTATCA TTCTTCACAC     1508
                                                                                .
                                                                              +200

        -300
          .                              ______________
2 ACTGTGTCTC TA---ACCAT GATACCGTGA CACAACTACA TCCGTACACA TGTGACGTTC GTTCAACCCG TACATTTATA TAAAACCGTT     1650
           * *     *  *   *  *   * *     **** **** ***   *  **    *    * *                   *  * **
3 TTCACAAAGC TTTTCCCATT TTTTTCAATA CTACTTTACA TCCGAACATT TTAGAAACCC ACACCAT--- ---------- ---ATACCTT     1582
                             .           ‾‾‾‾‾‾‾‾‾‾‾‾‾‾
                           -400

                       -200
                         .
2 CTGGCGGCCT TTTATTTTTT TACATTTCTT ATGATCGGGA TTGCAGAACG CCGTG----- ---------- ---------- ----------     1705
    *    *   ** ***** * * *   * *  *   *   *       **
3 TGGTGCACTA TTGATTTTCT TCCTGATGTC AGCTTTTTGT GCTTTGACAA AAAAATCGCG TCTACGTCCG TCCGTTCTCC CTGAATCAAT     1672
                                                          .
                                                        -300

                                                ___________
2 ---------- ---------- ---------- ---------- --AAATTTTT CAATGTGAGG TTCGGCCT-- ---TGTTTGC AAAAGCCCTA     1748
                                                   *  *      ****** *   **        ****         *  **
3 TAGGCGCGTT TGAGCCCAGC AGGACGGAGC TCTAGTGACA AGCCCTGGTG TTTGGTGAGG TAATGCACAT TGCTGTTCCT TTCTACTGTA     1762
                                                                     .
                                                                   -200

                         -100
                           .                                                     ~~~~~ ~~~~~~~~
2 TTGAGATACC GGAAAGATAT AGGTGAAATG AAGAAAA--C TATGGGTTGT ATATCTAATA CCCCGGTGCG TTTATTAATA TTTTAGCTTG     1836
  *******  *       *                  *           *****                                *
3 TTGAGATCTC CAGTTTACGG CTCCCTGGGT GCCACCCGTA ACGCGGTTGG ---------- ---------- ------TGTG CCCATTTCAA     1826

2 AAAGCGAAG- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -TGATACGAT     1854
   *******                                                                                        *
3 TAAGCGAACA TTAGTGAAGA CACAATCGTT AAAATGGACT AATGAAATTT TAAAGTGGGA TTTTTGTGAA TATTGACAAC AAAGGTATAG     1916
         .                                       ~~~~~~~~~~~~                           ‾‾‾‾‾
       -100

     ____                                         ______
2 CGACAAATAG AGTAAAA   ATG CCA GGT GTT TCC GTT AG  GTACGT AAAAATGACA TATCAT--AA TGATTATTAC TAT-ATGTTT     1934
     **** *  * ****   ***  ** *** *** *** *** **  ****** * ** *  **    ***  ** *** ****    *  ***  *
3 AACCAAAGAT AATAAAG   ATG GCA GGT GTT TCC GTT AG  GTACGT ATAACTTTCA CT-CATTGAA TGAGTATTTA GAGAATGAAT     1998
     ‾‾‾‾                                         ‾‾‾‾‾‾
                      met pro gly val ser val arg
                       1  ala                  7

2 CGAGTGGTTA GATGGAAATA AAACGCAGTC ATTCCAGCAG CATAGACACG CAAGCAAAGT ACTTCTTC-- ---------- ----------     2002
   **    *** * ******   **    *    *  *    **  ** ** *    * * *
3 GGA--TATTA GGTGGAAAAC AACGAGATGA AAACATATAG GATTGAGAAA AAGGAATTAG TGCAGAATTT TCCCATTACC TGATAAATTG     2086

2 ---------- ---------- ---------- --GTCAAGTT AAATAGCACC CATGGTTAAA TCAATGATAT ATCAACTGTC CCATA-ATAA     2059
                                         **** ***         **               * *     *   *     *     *
3 AAGTTCATCT TTACTGTGCA TTTACCGAAA CGAATGAGTT AAACTATTAT GATATAATCG CTGTAATTGT GGAGAGAATA TTTTGAGCCA     2176

2 TTGAAACAAA ATATTGATCA AATTTCGCGA ATGAGTGCAA TGTGTAAATA TATGGAAGAA GAGGGAGATT AAGGGTAGTG GTGCTATTTG     2149
     * ***** **   *                                   *      ** *** **** *          *   * *   *   *
3 ACTATACAAA ATGAAG---- ---------- ---------- --------TT CTCAGACGAA GAGGTAATCA TTTATTTATT GGAATCGATT     2234
```

```
2 CTCTT----- ---------- ---------- ----TGTTAT TACTAGGATT TCAATTCCTT CCGATAAGTT CTCATGAAAC ----------  2200
                                           *  **  ** *   * **       *   ******    * ****
3 GATAGGAGGT TTTAACTAGG TGGGATGTTT TCTAATGAAA TAGGAGAACC TGAACCAAGA CGTTTAAGTT GGAAGGAAAT TTGTAACCGT  2324


2 -GTAATG--- ---------- ---------- ----GTTAAT GAAAG----- ---------- ---------- ---------- ----------  2217
   ******                              ******   ***
3 GGTAATGCAA AGCAATGATA AGCGATTTTA TGCCGTTAAT TTAAGCAACA ACCCGAGACC ACGATACTGC CGCACCAAGA AGTTGTCTCA  2414


2 ---------- ---------- -------TAT TTTTTTTAAA CTGTATCATT TACTAACAAT AACTTTTTTG TTTTGATTTT GTTTTCT-AC  2279
                               * *  *****   *     *   ** *********  *   ***    ** **  ***  **  *
3 ATAAATATTT CACTGAGCAT TCCATATTCT ATTTTTACGA GAAAACTTTT TACTAACAAA AGCATTTCAA TTGTGCATTT TTT--CAATT  2502


2 AG     A GAC GTT GCT GCT CAA GAT TTC ATT AAT GCT TAC GCT TCT TTC TTG CAA AGA CAA GGT AAG CTA GAA GTT  2351
  **     * *** *** **  *** *** *** *** **  *** *** *** *** *** *** *** *** *** *** *** **    ** *** **
3 AG     A GAC GTT GCA GCT CAA GAT TTC ATC AAT GCT TAC GCT TCT TTC TTG CAA AGA CAA GGT AAA TTA GAA GTC  2574
           asp val ala ala gln asp phe ile asn ala tyr ala ser phe leu gln arg gln gly lys leu glu val
            8                                              20

2 CCA GGT TAC GTT GAC ATT GTC AAG ACC TCT TCT GGT AAC GAA ATG CCA CCA CAA GAT GCC GAA GGT TGG TTC TAC  2426
  *** *** *** *** *** *** *** *** *** *** *** *** **  *** *** *** *** *** **  **  *** *** *** *** ***
3 CCA GGT TAC GTT GAC ATT GTC AAG ACC TCT TCT GGT AAT GAA ATG CCA CCA CAA GAC GCT GAA GGT TGG TTC TAC  2649
  pro gly tyr val asp ile val lys thr ser ser gly asn glu met pro pro gln asp ala glu gly trp phe tyr
                                       40

2 AAG CGT GCT GCC TCT GTT GCC AGA CAC ATT TAC ATG AGA AAA CAA GTT GGT GTT GGT AAA TTG AAC AAA TTA TAC  2501
  *** *** *** *** **  *** **  *** *** *** *** *** *** **  *** **  *** *** *** **  *** *** *** **  ***
3 AAG CGT GCT GCC TCC GTT GCT AGA CAC ATT TAC ATG AGA AAG CAA GTC GGT GTT GGT AAG TTG AAC AAA TTG TAC  2724
  lys arg ala ala ser val ala arg his ile tyr met arg lys gln val gly val gly lys leu asn lys leu tyr
                   60                                                                              80

2 GGT GGT GCC AAG AGC AGA GGT GTT AGA CCA TAC AAG CAC ATT GAC GCT TCC GGT TCT ATC AAC AGA AAG GTT TTG  2576
  *** *** *** *** *** *** *** **  *** *** *** *** *** *** **  *** *** *** *** *** *** *** *** **  ***
3 GGT GGT GCC AAG AGC AGA GGT GTC AGA CCA TAC AAG CAC ATT GAT GCT TCC GGT TCT ATC AAC AGA AAG GTC TTG  2799
  gly gly ala lys ser arg gly val arg pro tyr lys his ile asp ala ser gly ser ile asn arg lys val leu
                                                                              100

2 CAA GCT TTG GAA AAG ATT GGT ATC GTC GAA ATC TCT CCA AAG GGT GGT AGA AGA ATC TCT GAA AAC GGT CAA AGA  2651
  *** *** *** *** **  **  *** **  *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** ***
3 CAA GCT TTG GAA AAA ATC GGT ATT GTC GAA ATC TCT CCA AAG GGT GGT AGA AGA ATC TCT GAA AAC GGT CAA AGA  2874
  gln ala leu glu lys ile gly ile val glu ile ser pro lys gly gly arg arg ile ser glu asn gly gln arg
                                                          120

2 GAT TTG GAT CGT ATT GCC GCT CAA ACT TTG GAA GAA GAC GAA TAA  ATAT AAAATCTATA ATTTATATAT ATATACTACT  2730
  *** *** *** *** *** *** *** *** *** *** *** *** *** *** ***           *  * * **  * *  * * * * *  *
3 GAT TTG GAT CGT ATT GCC GCT CAA ACT TTG GAA GAA GAC GAA TAA  GCGA GTTGTTAAAA ATAAAAAGCT A-AGATTTTT  2952
  asp leu asp arg ile ala ala gln thr leu glu glu asp glu end
                                      140             144

2 ACTACAATTA TCATCATACA AGTAATAATA A--------- ---------- ---------- ---------- ---------- ----------  2761
  * **  **** **** ****  *
3 ATTAATATTA TCATTATACT ATAGTATTAT CGTTTAACTC AAACAAAATT GATTTTAAAT AATTTATTTA TTTAAATTTT GCCATTTTAT  3042
                                                                       +100

                                                                                +200
3 ATGTACATTC ATCATCATCA ATATAGCCGA AACCAAAATA AATAGATTTT GTTTTTTTAA GATAGAATGT GTCAAACCAT GGATTGCGAT  3132
                                                                                           +300
3 CTGGATAAAA ATGACAACTT TTAAGTTTTC GTCTACAAAA CATTAAAAAC ACCATTGAAG AATAAATGGT AGTAAATAAT TTTACAATGA  3222
```

Fig. 4. Primary structure of the rp28 and S16A gene pairs of two yeast strains.
1, 2 and 3 represent the primary structures of S.ca-1, S.ce-1 and S.ca-2 (cf. Fig. 3), respectively. The nucleotides for S.ca-1 and S.ce-1 are indicated separately only in the case of base substitutions. Dashed lines indicate deletions. The positions at which both copy genes (S.ce-1 and S.ca-2) have identical nucleotide sequences are marked by asterisks. Notable sequences are referred to in the text.

////// ////// TATA-like elements (36).

|||||||| PyAAPu (37,38)

〰 〰 〰 putative transcription-termination and/or polyadenylation signals (10,39,41).

respect to the duplicate and linked nature of these genes. Southern and heteroduplex analyses permitted us to map the genes as indicated in Fig. 3. To analyse the structure of the genes coding for rp28 and S16A the complete nucleotide sequences of the *S.ce* gene pair and the second-copy *S.ca* gene pair have been determined. To be able to evaluate a comparison of duplicate gene copies isolated from two yeast strains, we also have sequenced the "first-copy" *S.ca* rp28 gene carried by pBMCY44 that corresponds with the *S.ce* rp28 gene. In Fig. 3 the strategy to determine the complete sequence is outlined. The primary structure of the rp28-S16A gene pairs that resulted from these sequence analyses is presented in Fig. 4.

First we compared the sequence of the homologous *S.ca* and *S.ce* gene copies coding for rp28. Only 13 nucleotide differences were observed in the non-translated and flanking regions of the gene and none in the coding region. Three of these changes were in the intron and ten were in the 300 nucleotides 5' to the coding region (Fig. 4). A low degree of divergence on the DNA level between *S.ca* and *S.ce* has been reported previously for the actin gene (28) and the ribosomal RNA genes (29) and was also observed for the histon H4 gene (30,31). Our results support the present idea that *S.ca* and *S.ce* are similar or identical yeast strains (32). Therefore any alterations (except for minor base substitutions) found by comparing the DNA sequences of the first-copy gene pair of *S.ce* with the second-copy gene pair of *S.ca* can be considered as differences between the duplicated gene pairs rather than strain differences.

The sequence data presented in Fig. 4 indicate that the genes coding for rp28 and S16A are transcribed in the same direction in both gene pairs and that there are 544 and 642 base pairs between coding regions in *S.ce* gene pair 1 and *S.ca* gene pair 2, respectively.

### Sequence comparison: duplicate genes, coding region and intron

The genes contain some structural features which many ribosomal protein genes have in common. First of all, as predicted above, both genes are interrupted by an intron near the 5'-end of the coding sequence. These data are summarized in Table I. The intron-exon boundaries are conserved, also near the 3'-end of the introns another common sequence, TACTAACA, is present. This sequence has been shown to be essential for splicing of intron-containing pre-mRNAs in yeast (11-13). It is clear that the intervening sequences interrupting the duplicate genes differ not only in length (*cf.* Table I) but also in nucleotide sequence. By computer analysis some regions could be identified that are conserved only in the introns of duplicate genes encoding the same protein. Two such conserved regions within the rp28-introns show a dyad symmetry of

Table I. Structure of the ribosomal protein genes encoding rp28 and S16A.

| | | | 5' ↓ | | | | | | 3' ↓ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sce rp28: | ATG - | 36 codons | - | GTATGT | - 411 n. - | TACTAACA | - 19 n. - | TAG | - | 149 codons - TAA |
| Sca rp28: | ATG - | 36 codons | - | GTATGT | - 377 n. - | TACTAACA | - 35 n. - | CAG | - | 149 codons - TAA |
| Sce S16A: | ATG - | 5 codons | - | GTACGT | - 343 n. - | TACTAACA | - 30 n. - | CAG | - | 138 codons - TAA |
| Sca S16A: | ATG - | 5 codons | - | GTACGT | - 505 n. - | TACTAACA | - 29 n. - | TAG | - | 138 codons - TAA |
| consensus: | | | | GTAPyGT | | TACTAACA | | PyAG | | |

13 base pairs, which would permit a secondary structure element in the precursor rp28-mRNAs (indicated in Fig. 4 with arrows).

The introns interrupt the coding regions of the duplicate genes at the same position (cf. Table I).

With respect to the amino acid coding regions, we observed 95% homology between both duplicate gene pairs. Most base substitutions do not give rise to an amino acid substitution. For the S16A genes only 1 out of 21 nucleotide substitutions causes an amino acid substitution, viz. Pro/Ala at position 2 (cf. Fig. 4). Concerning the rp28 genes, the 27 base changes do not result in amino acid substitutions. It is interesting to know that the silent base changes in rp28 are clustered in the short 5'-exon. 14 out of 112 bases are substituted in the first exon and only 13 out of 446 are changed in the large exon. This clustering of silent base changes is not seen in S16A. By Otaka et al. (33) the amino acid sequence of the first 49 N-terminal residues of the S16A protein has been established. Their data are completely consistent with the amino acid sequence deduced from the DNA sequence. Moreover, Otaka et al. found both proline and alanine as N-terminal amino acids at a ratio of 1:2. This finding is in excellent agreement with our results and, moreover, provides evidence that both S16A genes are expressed.

The coding sequences of all four genes examined show a rather preferred codon usage consistent with previous findings for other efficiently transcribed yeast genes (34,7).

Sequence comparison: noncoding regions

Comparison of the regions upstream from the ATG initiation codons only reveal short regions that are conserved between duplicate genes encoding the same protein. The presumed leader sequences of the duplicate genes show a low degree of homology as can be deduced from Fig. 4. It is unlikely, therefore, that specific primary structure elements in this part of the mRNAs are involved in translational control in the autogenous way it has been observed in bacterial cells (35).

In the 5'-flanking region of the S16A genes the most remarkable conserved sequence is TACATCCG(T/A)ACA at positions -274 and -389 in gene copy 1 and gene copy 2, respectively (Fig. 4). This sequence is very similar to consensus sequence HOMOL1 observed by Teem et al., viz. AACATC(T/C)(G/A)T(A/G)CA (10). Since this sequence element occurs at a position of about -300 for 6 yeast ribosomal protein genes investigated so far and does not precede any of the 20 yeast non-ribosomal protein genes examined, it was proposed to play a role in the coordinate expression of these genes (10). A similar sequence, GACCTCAGTACA, matching in 9 out of 12 positions with consensus sequence HOMOL1 was observed at -197 from the rp28-2 coding region (Fig. 4), but not in the 5' flanking region of the rp28-1 gene. In the upstream region of both rp28 genes, however, we observed the conserved sequence T(T/C)TTTCTTGCTGGAG(A/C)TT at position -193 and -226 in the first- and second-copy genes, respectively (Fig. 4). We suggest that additional sequence elements of this kind may function in the coordinate expression in subsets of yeast ribosomal protein genes.

Sequences that fit other consensus sequences proposed by Teem et al. (10) are AATTTTTCA at position -165 in the S16A-1 gene (consensus sequence HOMOL4: (T/A)AT(T/A)TTnCA), TATTTA at position -97 in rp28-2 and TATT(T/A)(T/A) at positions -47 and -53 in the S16A-1 gene (consensus sequence HOMOL5: TATT(T/A)(T/A), resembling the "TATA" box [36]). More general TATA-like structures are found in the two duplicate pairs of genes that lack consensus sequence HOMOL5 (Fig. 4). Finally, in each gene one or more sequences can be identified that fit the consensus sequence PyAAPu that has been proposed to act as a transcriptional start site for yeast polymerase B (37,38) (see Fig. 4).

From the sequence data presented in Fig. 4 it is apparent that the sequences downstream of the TAA stopcodon have diverged to a similar extent as the other noncoding regions. Several sequences have been proposed to be involved in transcription termination and/or polyadenylation in yeast (10,39,40) or Eumetazoa (41). All four genes examined contain one or more of these consensus sequences as indicated in Fig. 4.

In summary, we have analysed the structure of two linked ribosomal protein genes in yeast by heteroduplex - and sequence analysis. We have demonstrated that these genes are duplicated in the genome in the same linkage arrangement. The almost complete conservation of amino acid sequence in duplicate copies of these genes, as deduced from the DNA sequence, and the work of Otaka et al. (33) described in the results section, strongly suggests that all four genes make functional products. The fact that a similar linkage arrangement is maintained in both gene pairs, even though the distance and sequence between them is different, suggests that the linkage arrangement may play some role in the regulation of expression of these genes. Finally, it will be of interest to determine whether these duplicated genes, which code for almost identical gene products, are regulated in the same or different manner, considering that the regions 5´ to the ATG start codons are almost totally diverged.

## ACKNOWLEDGEMENTS

C.M.T.M. et al. are indebted to Dr. R.J. Leer and Mrs. M.M.C. van Raamsdonk-Duin for stimulating discussions and technical advices, to Mrs. C. Sondermeijer and Mr. A.B. Smit for technical assistance during part of this work and to Mrs. P.G. Brink for preparing the typescript. The S.ca. work was supported by the Netherlands Foundation for Chemical Research (S.O.N.) with financial aid from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

N.J.P. and D.M.D. would like to thank M.P. Remington for her excellent technical assistance; Dr. J. Scocca, G. Pearson, and W. Schwindinger for assistance with computer analysis; and M. Neeper for helpful discussions. The S.ce. work was supported by NIH grant GM 31812 and an ACS Institutional Research Grant #IN-174B (to D.M.D.) University of Maryland/Maryland Cancer Program. Preliminary work was begun in Dr. J. Warner's laboratory supported by NIH Grant GM 25532.

## REFERENCES

1. Planta,R.J. and Mager,W.H. (1982) in The Cell Nucleus, Busch,H. and Rothblum,L. Eds., Vol 12, pp.213-225, Academic Press, New York.
2. Warner,J.R. (1982) in The Molecular Biology of the yeast Saccharomyces, Strathern,J.N., Jones,E.W. and Broach,J.R., pp.529-560, CSH, New York.
3. Fried,H.M., Pearson,N.J., Kim,C.H. and Warner,J.R. (1981) J. Biol. Chem. 256, 10176-10183.
4. Woolford,J.L. and Rosbash,M. (1981) Nucl. Acids Res. 9, 5021-5036.
5. Molenaar,C.M.T. (1984) Ph.D. Thesis, Free University, Amsterdam.
6. Leer,R.J., Raamsdonk-Duin,M.M.C. van, Molenaar,C.M.T., Cohen,L.H., Mager,W.H. and Planta,R.J. (1982) Nucl. Acids Res. 10, 5869-5878.
7. Leer,R.J. Raamsdonk-Duin,M.M.C. van, Schoppink,P.J., Cornelissen,M.T.E., Cohen,L.H., Mager,W.H. and Planta,R.J. (1983) Nucl. Acids Res. 11, 7759-7768.
8. Käufer,N.F., Fried,H.M., Schwindinger,W.F., Jasin,M. and Warner,J.R. (1983) Nucl. Acids Res. 11, 3123-3127.
9. Teem,J.L. and Rosbash,M. (1983) Proc. Natl. Acad. Sci. USA 80, 4403-4407.
10. Teem,J.L., Abovich,N., Käufer,N.F., Schwindinger,W.F., Warner,J.R., Levy,A., Woolford,J., Leer,R.J., Raamsdonk-Duin,M.M.C. van, Mager,W.H., Planta,R.J., Schultz,L., Friesen,J.D. and Rosbash,M. (1984) Nucl. Acids Res. submitted.
11. Langford,C.J. and Gallwitz,D. (1983) Cell 33, 519-527.
12. Pikielny,C.W., Teem,J.L. and Rosbash,M. (1983) Cell 34, 395-403.
13. Langford,C.J., Klinz,F., Donath,C. and Gallwitz,D. (1984) Cell 36, 645-653.

14. Warner,J.R. and Gorenstein,C. (1978) Methods in Cell Biol. 20, 45-60.
15. Kaltschmidt,E. and Wittmann,H.G. (1970) Anal. Biochem. 36, 401-412.
16. Warner,J.R. and Gorenstein,C. (1977) Cell 11, 201-212.
17. Bollen,G.H.P.M., Molenaar,C.M.T., Cohen,L.H., Raamsdonk-Duin,M.M.C. van, Mager,W.H. and Planta,R.J. (1982) Gene 18, 29-37.
18. Verbeet,M.Ph., Klootwijk,J., Heerikhuizen,H. van, Fontijn,R.D., Vreugdenhil,E. and Planta,R.J. (1983) Gene 23, 53-63.
19. Grosveld,F.G., Dahl,H-H.M., Boer,E. de and Flavell,A. (1981) Gene 13, 227-237.
20. Ish-Horowicz,D. and Burke,J.F. (1981) Nucl. Acids Res. 9, 2989-2998.
21. Rigby,P.W.J., Dieckmann,M., Rhodes,C. and Berg,P. (1977) J. Mol. Biol. 113, 237-251.
22. Clewell,D.B. and Helinsky,D.R. (1969) Proc. Natl. Acad. Sci. USA 62, 1159-1166.
23. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
24. Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) J. Mol. Biol. 143, 161-178.
25. Messing,J. and Vieira,A. (1982) Gene 19, 269-276.
26. Smith,H.O. and Birnstiel,M.L. (1976) Nucl. Acids Res. 3, 2387-2398.
27. Southern,E.M. (1975) J. Mol. Biol. 98, 503-517.
28. Nellen,W., Donath,C., Moos,M. and Gallwitz,D. (1981) J. Mol. Appl. Gen. 1, 239-244.
29. Skryabin,K.G., Eldarov,M.A., Larionov,V.L., Bayev,A.A., Klootwijk,J., Regt,V.C.H.F. de, Veldman,G.M., Planta,R.J., Georgiev,O.I. and Hadjiolov,A.A. (1984) Nucl. Acids Res. 12, 2955-2968.
30. Woudt,L.P., Pastink,A., Kempers-Veenstra,A.E., Jansen,A.E.M., Mager, W.H. and Planta,R.J. (1983) Nucl. Acids Res. 11, 5347-5360.
31. Smith,M.M. and Murray,K. (1983) J. Mol. Biol. 169, 663-690.
32. Barnett,J.A., Payne,R.W. and Yarrow,D. (1983) in Yeast: Characteristics and identification, pp.467-469, Cambridge University Press.
33. Otaka,E., Higo,K. and Osawa,S. (1982) Biochemistry 21, 4545-4550.
34. Ammerer,G., Hitzeman,R., Hagie,F., Barta,A. and Hall,B.D. (1981) in Recombinant DNA, Walton,A.G. Ed., pp.185-197, Elsevier Scientific Publishing Co, Amsterdam.
35. Dean,N. and Nomura,M. (1982) in The Cell Nucleus, Bush,H. and Rothblum, L., Vol. 12, pp.213-225, Academic Press, New York.
36. Benoist,C. and Chambon,P. (1981) Nature 290, 304-310.
37. Dobson,M.J., Tuite,M.F., Roberts,N.A., Kingsman,A.J., Kingsman,S.M., Perkins,S.E., Conroy,S.C., Dunbar,B. and Fothergrill,L.A. (1982) Nucl. Acids Res. 10, 2625-2637.
38. Burke,R.L., Tekamp-Olsen,P. and Najarian,R. (1983) J. Biol. Chem. 258, 2193-2201.
39. Bennetzen,J.L. and Hall,B.D. (1982) J. Biol. Chem. 257, 3026-3031.
40. Zaret,K.S. and Sherman,F. (1982) Cell 28, 563-573.
41. Fitzgerald,M. and Shenk,T. (1981) Cell 24, 251-260.