Volume 247, number 2, 298-302 FEB 07045 April 1989

Clones containing variant forms of complete human rRNA genes

Characterization and sequence of their transcription initiation region

Marie-Hélène Renalier, Nicole Joseph and Jean-Pierre Bachellerie

Centre de Recherche en Biochimie et Génétique Cellulaires du CNRS, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex, France

Received 14 February 1989; revised version received 1 March 1989

In order to assess the extent of structural heterogeneity among ribosomal genes in the human genome, several cosmid clones, each containing an entire transcription unit, have been isolated and analyzed. Five cloned genes were recognized as structurally different from each other to some extent, either within the transcription unit or in its immediate vicinity. The sequence of a 1.2 kb region encompassing the transcription start site has been determined for the five cloned genes. Point differences among the genes are observed at five nucleotide positions within the analyzed portion of the 5′ external transcribed spacer of the ribosomal gene. Moreover, upstream from the transcription start site, a unique difference is observed with the absence from the three genes of a 19-nucleotide long stretch which is present in the two other variant gene forms.

rRNA gene; Cosmid; Nucleotide sequence; Transcription initiation; Gene variant; Evolution

1. INTRODUCTION

A ribosomal RNA gene represents the interspersion of highly conserved and of more rapidly evolving domains: whereas some sequence motifs are universally preserved, others may diverge extensively, even between closely related species [1-3]. Like other multigene families, ribosomal genes have undergone concerted evolution in eukaryotes [4-7]: within a species individual genes are closely similar to each other and mature rRNA sequences appear to be essentially homogeneous [8]. However, the structural homogeneity is not absolute [9-11] and subclasses of ribosomal genes have been identified in some eukaryotes [12-14]. In humans, the approx. 200 copies of ribosomal genes present per haploid genome are arranged in clusters of tandem repeats distributed on the short arms of the five pairs of acrocentric (nos 13-15, 21, 22) chromosomes [15]. Evidence of some structural polymorphism has been reported for the ribosomal intergenic spacer [16,17] and also in 28 S rDNA sequences [4,18]. More recently, direct analyses on mature 28 S rRNA have revealed a limited but significant extent of sequence microheterogeneity, mostly located over the more rapidly evolving domains of the molecule [19] but also present over a phylogenetically conserved region [20]. However, thus far, clones containing an entire ribosomal gene have not been reported for mammals, hence precluding full assessment of the extent of polymorphism and recognition of potentially linked groups of variants. Here, we have characterized several clones containing variant forms of a complete human ribosomal gene and determined their sequences around the transcription start site.

Correspondence address: J.-P. Bachellerie, Centre de Recherche en Biochimie et Génétique Cellulaires du CNRS, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex, France

The nucleotide sequence presented here has been submitted to the EMBL/GenBank database under the accession no. Y07502

Published by Elsevier Science Publishers B.V. (Biomedical Division)

2. MATERIALS AND METHODS

A human cosmid library, constructed from a sized fraction (30–45 kb) of leukocyte DNA partially digested with *MboI* using pCV108 as cosmid vector [21], was screened by colony hybridization [22] with the 2 kb long *SalI-EcoRI* fragment of mouse rDNA which contains most of the 18 S rRNA sequence [23]. 28 positive clones were isolated from 3×10^4 screened clones. All 28 were subsequently probed, by colony hybridization, for the presence of an A or G at position 60 of the 28 S rRNA sequence, using a selective hybridization assay with a pair of cognate synthetic 22-mer oligonucleotides, as described [20]. Eleven of the 28 positive clones coded for the A variant form (all others coding for the G variant form). The 28 positive clones were next screened for the presence of the transcription initiation region, using successively two synthetic probes (see fig.1) selected according to the previously reported sequences [24,25]: one corresponds to the 5′-extremity of the transcription unit, i.e. to a sequence tract which is selectively conserved among mammals [24]. 15 clones devoid of the initiation region were not studied further. The 13 remaining positive clones were submitted to restriction analysis and Southern blot hybridization with various probes (the presence of the 3′-end of the transcription unit in the insert was tested with a 0.9 kb long *EcoRI-BamHI* fragment of mouse rDNA [1] containing the 3′-end of 28 S rRNA sequence). Five clones (fig.1) presented some evidence of corresponding to distinct rDNA loci and were submitted to sequence analysis of their transcription initiation region: the 1.2 kb *EcoRI-SalI* fragment from each recombinant cosmid was subcloned in M13 vectors [26] and sequenced by the dideoxynucleotide chain-termination method [27]. No ambiguity remained over the entire sequences which were determined on both strands over most of their length. Human genomic DNA was isolated from a placenta and HeLa cells, digested, and submitted to Southern blot hybridization with ^{32}P nick-translated probes [28].

3. RESULTS AND DISCUSSION

On the basis of their restriction maps in the regions upstream and downstream from the transcription unit and taking into account the A/G variance at position 60 of the 28 S rRNA sequence [20], five positive clones containing the transcription initiation region were shown to correspond to distinct genomic loci. The structure of these inserts is schematized in fig.1. One clone, termed 12.8, corresponds to a 3′-truncated gene, possessing only a few hundred bp of the 28 S rRNA 5′-terminal sequence. The other four contain the 3′-end of the gene, as indicated by cross-hybridization with a mouse rDNA probe and by the presence of several characteristic restriction sites in the downstream portion of the intergenic spacer. Over the transcription unit, no discrepancy was detected between the five clones for the key restriction sites (fig.1) previously identified in genomic DNA [15,29]. However, some differences were apparent over the portions of intergenic spacers present in the cosmid inserts. Immediately downstream from the gene, a region of discrete size heterogeneity among human rDNA units had been detected [4], corresponding to the presence of variable numbers (from 1 to 4) of a 700 bp tandemly repeated sequence [17]: on the basis of Southern blot hybridization with a probe mapping at the 3′-end of the transcription unit (see fig.1), clone 11.9 appears to contain only 2 such sequences, while 3 are present in clones 11.36, 12.1 and 11.11 (this particular size class also appears to be the most abundant in human individuals [17]). In clone 11.36, the extent of upstream intergenic spacer present in the insert amounts to about 7 kb, with a restriction map in full agreement with the major pattern of rDNA genomic organization [15,30]. However, beyond the *SalI* site located about 0.6 kb upstream from the gene, the restriction maps of the other inserts differ from the canonical pattern represented in clone 11.36. Further analyses of these variant upstream regions (which map within the dashed portions of the inserts in fig.1) in clones 12.1, 11.9, 12.8 and 11.11 are definitely needed in order to assess the significance of these differences. They could correspond to the isolation of minor forms of rDNA units, such as those which are located at the 5′-border of a cluster of tandem repeats. In humans, the junction with non-rDNA has been mapped 3.7 kb upstream from the transcription start site of the first repeat unit [31]. From the restriction analysis data available so far and by reference to the restriction map reported in [31], the possibility remains that clone 12.1 represents such a 5′-border unit. The sequence of the *EcoRI-SalI* fragment encompassing the transcription initiation site was then determined for each insert (fig.2). It is noteworthy that this *EcoRI* site is polymorphic in human genomic DNA, as shown in fig.1 and in line with a previous report [15]. In fact, we have observed that the proportion of rDNA units resistant to *EcoRI* cleavage at this site may vary according to the origin of the DNA sample (not shown). However, all the clones that we have analyzed contain this *EcoRI* site, with the GAATTC recognition motif followed by a G, which provides a potential CpG substrate for m^5C methylation in genomic DNA. Accordingly, this

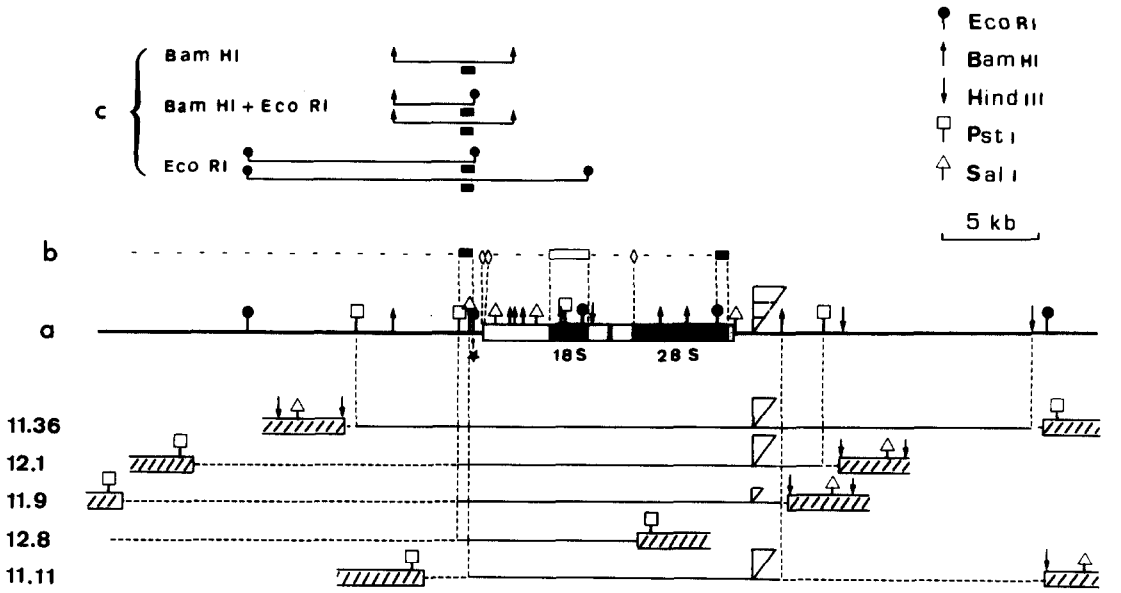
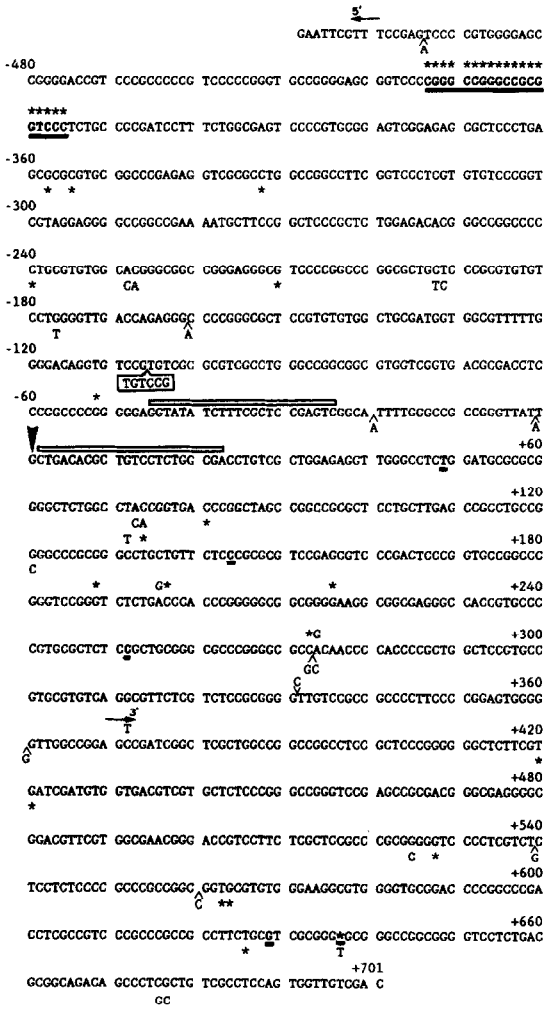


Fig.1. Structure of the humn rDNA cosmid clones. (a) Map of human genomic rDNA for key restriction sites. In the transcription unit, spacer regions are denoted by open boxes. A region of length heterogeneity downstream from the gene is denoted by triangles. (b) Location of DNA probes used for screening of the cosmid library (open symbols, with lozenges referring to synthetic oligonucleotides) and analysis of cloned and genomic DNAs (filled symbols). (c) Restriction fragments of genomic DNA identified by Southern blot hybridization with the *Pst*I-*Eco*RI probe (from the 11.36 clone) after *Bam*HI, *Bam*HI + *Eco*RI and *Eco*RI digestions, showing the existence of a polymorphic *Eco*RI site (denoted by a star in a). Extents of each cosmid insert are shown at the bottom: portions represented as a continuous line possess a restriction map in complete agreement with that in (a) for genomic rDNA, while dashed segments either differ or have not been characterized further (hatched boxes denote the cosmid vector).

*Eco*Ri polymorphism could correspond to variations in the methylation patterns of rDNA units. As shown in fig. 2, the sequences for the five clones are almost perfectly identical except for 5 nucleotide positions and also for the variable presence of a 19-nucleotide tract (present in clones 11.9 and 12.1 but missing in the 3 others). The latter variation seems likely to have been generated through a slipped-strand mispairing during replication, since the 19-nucleotide motif is repeated almost exactly at the 5′-border of this insertion/deletion. None of the variations among the five clones maps within the two major control elements (termed core and upstream control elements) identified in the human rRNA gene promoter [32]. The sequence of all, or a part, of this 1.2 kb region of a human rRNA gene has been reported previously [24,25]. A substantial level of divergence is observed (fig. 2), with 33 point changes (and a hexanucleotide insertion) over the 1.2 kb sequence by reference to [24]. The divergence is much less extensive when referring to [25], with only 11 point differences (fig. 1) over the 875 nucleotides sequenced in the latter analysis.

As for the variant 19-nucleotide tract, it is present in the fragment of human rDNA analyzed in [24] but deleted in the other case [25]. Thus, while most of the above-mentioned point nucleotide differences could be attributed to the different origin of the human DNA libraries in our work and in previous studies [24,25], our results suggest that the 19-nucleotide tract variance represents a general feature of human rDNA polymorphism, which is widespread in the population. Such a sizable segmental mutation can introduce a significant bias against non-reciprocal exchanges between the two types of units, thereby favouring the propagation of linked groups of variations within each size class rather than within the entire gene family. In support of this notion, it is remarkable that the three cloned genes which are devoid of the 19-nucleotide tract, i.e. clones 12.8, 11.11 and 11.36, have a perfectly identical sequence over the 1.2 kb region analyzed here (fig. 1). However, for this group the linked variance does not extend to the sequence heterogeneity detected at position 60

Fig.2. Sequence of the cloned human rDNA genes around the transcription start site. The sequence shown corresponds to the 11.9 clone. Nucleotide positions where differences among the five sequenced clones have been detected are underlined by a bar. Positions are numbered by reference to the transcription start site (large arrowhead). Differences with two other previously reported [24,25] sequences for this region of human rDNA are indicated above [25] or below [24] the sequence, with small arrowheads denoting additional nucleotides found in these two sequences [24,25]. The boundaries of the portion analyzed in [25] are denoted by horizontal arrows above the sequence. The two open bars above the sequence delineate two synthetic oligodeoxynucleotides used for probing the presence of the 5′ boundary of the transcription unit in the cloned cosmids. (Inset) Comparison of the sequences for the five clones, with indication of all variant nucleotides (the open triangle at −415 denotes the 19-nucleotide long tract which is missing in three clones but present, and perfectly identical, in the other two).



of 28 S rRNA (with an A present in clone 11.36, but a G in clones 12.8 and 11.11), possibly in relation with the much greater length of intervening DNA involved in that case (i.e. almost 9 kb). Thus the variant 19-nucleotide tract definitely appears to be a valuable marker for studying the genomic distribution of variant ribosomal genes (within a cluster of tandem repeats and among the five human acrocentric chromosomes). In turn, this should allow better evaluation of the extent and dynamics of the intra- and interchromosomal recombinational events which result in the concerted evolution of this multigene family [6,7].

Although the cloning of fragments containing a large portion of the human rRNA transcription unit has been reported recently [33], the clones described here represent the first cases of complete ribosomal genes reported for a mammal. These different units contain variant structural features, both within and around the transcription unit, particularly in an area of the intergenic spacer (extending over 2–3 kb upstream from the major promoter) which might play an important role in the transcriptional control of mammalian ribosomal genes [34,35]. They should therefore prove to be useful tools for further analyses of rRNA gene expression, at the level of transcription or rRNA processing.

Acknowledgments: The support of Professor J.P. Zalta was greatly appreciated. This work was aided by grants from INSERM (CRE no. 851001) and ARC (no. 6158).

REFERENCES

[1] Hassouna, N., Michot, B. and Bachellerie, J.P. (1984) Nucleic Acids Res. 12, 3563–3583.
[2] Michot, B., Hassouna, N. and Bachellerie, J.P. (1984) Nucleic Acids Res. 12, 4259–4279.
[3] Gorski, J.L., Gonzalez, I.L. and Schmickel, R.D. (1987) J. Mol. Evol. 24, 236–251.

[4] Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. and Zimmer, E. (1980) Proc. Natl. Acad. Sci. USA 77, 7323–7327.
[5] Long, E.O. and Dawid, I.B. (1980) Annu. Rev. Biochem. 49, 727–764.
[6] Dover, G. and Coen, E. (1981) Nature 290, 731–732.
[7] Dover, G.A. (1986) Trends Genet. 2, 159–165.
[8] Maden, B.E.H., Forbes, J.M., Stewart, M.A. and Eason, R. (1982) EMBO J. 1, 597–601.
[9] Stewart, M.A., Hall, L.M.C. and Maden, B.E.H. (1983) Nucleic Acids Res. 11, 629–646.
[10] Yagura, T., Yagura, M. and Muramatsu, M. (1979) J. Mol. Biol. 133, 537–547.
[11] Kolosha, V.O., Kryukov, V.M. and Fodor, I. (1986) FEBS Lett. 197, 89–92.
[12] Dame, J.B., Sullivan, M. and Mc Cutchan, T.F. (1984) Nucleic Acids Res. 12, 5943–5952.
[13] Back, E., Müller, F. and Tobler, H. (1984) Nucleic Acids Res. 12, 1313–1332.
[14] Briner, G., Müller, E., Neuhaus, H., Back, E., Müller, F. and Tobler, H. (1987) Nucleic Acids Res. 15, 6515–6538.
[15] Wilson, G.N. (1982) in: The Cell Nucleus (Busch, H. and Rothblum, L. eds) vol. 10, pp. 287–318, Academic Press, New York.
[16] Krystal, M. and Arnheim, N. (1978) J. Mol. Biol. 126, 91–104.
[17] La Volpe, A., Simeone, A., D'Esposito, M., Scotto, L., Fidanza, V., De Falco, A. and Boncinelli, E. (1985) J. Mol. Biol. 183, 213–223.
[18] Gonzalez, I.L., Gorski, J.L., Campen, T.J., Dorney, D.J., Erickson, J.M., Sylvester, J.E. and Schmickel, R.D. (1985) Proc. Natl. Acad. Sci USA 82, 7666–7670.
[19] Gonzalez, I.L., Sylvester, J.E. and Schmickel, R.D. (1988) Nucleic Acids Res. 16, 10213–10224.
[20] Nicoloso, M., Qu, L.H. and Bachellerie, J.P. (1989) Biochem. Biophys. Res. Commun., in press.
[21] Lau, Y.-F. and Kan, Y.W. (1983) Proc. Natl. Acad. Sci USA 80, 5225–5229.
[22] Grunstein, M. and Hogness, D. (1975) Proc. Natl. Acad. Sci USA 72, 3961–3965.
[23] Raynal, F., Michot, B. and Bachellerie, J.P. (1984) FEBS Lett. 167, 263–268.
[24] Financsek, I., Mizumoto, K., Mishima, Y. and Muramatsu, M. (1982) Proc. Natl. Acad. Sci USA 79, 3092–3096.
[25] Miesfeld, R. and Arnheim, N. (1982) Nucleic Acids Res. 10, 3933–3949.
[26] Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucleic Acids Res. 9, 309–321.
[27] Sanger, F., Nicklen, S. and Coulson, A. (1977) Proc. Natl. Acad. Sci USA 74, 5463–5467.
[28] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[29] Erickson, J.M., Rushford, C.L., Dorney, D.J., Wilson, G.N. and Schmickel, R.D. (1981) Gene 16, 1–9.
[30] Sylvester, J.E., Whiteman, D.A., Podolsky, R., Pozsgay, J.M., Respess, J. and Schmickel, R.D. (1986) Hum. Genet. 73, 193–198.
[31] Worton, R.G., Sutherland, J., Sylvester, J.E., Willard, H.F., Bodrug, S., Dubé, I., Duff, C., Kean, V., Ray, P.N. and Schmickel, R.D. (1988) Science 239, 64–68.
[32] Haltiner, M., Smale, S.T. and Tjian, R. (1986) Mol. Cell. Biol. 6, 227–235.
[33] Maden, B.E.H., Dent, C.L., Farrell, T.E., Garde, J., Mc Callum, F.S. and Wakeman, J.A. (1987) Biochem. J. 246, 519–527.
[34] Kuhn, A. and Grummt, I. (1987) EMBO J. 6, 3487–3492.
[35] Cassidy, B.G., Yang-Yen, H.F. and Rothblum, L.I. (1987) Mol. Cell. Biol. 7, 2388–2396.