Evidence for a common ancestor sequence for the Balbiani ring 1 and Balbiani ring 2 genes in Chironomus tentans

(molecular evolution/gene family/repetitive sequences/structural proteins)

LARS WIESLANDER, JANOS SÜMEGI, AND BERTIL DANEHOLT

Department of Medical Cell Genetics, Medical Nobel Institute, Karolinska Institutet, S-10401 Stockholm 60, Sweden

Communicated by Alan Garen, August 6, 1982

ABSTRACT The Balbiani ring (BR) 1 and BR 2 genes in Chironomus tentans are functionally related and are only expressed in the salivary gland cells. Here we reveal the principal structure of the BR 1 gene and analyze the structural and evolutionary relationship between the BR 1 and BR 2 genes. The properties of the BR 1 gene, 37 kilobases in size, are derived from the analysis of a cloned cDNA sequence, pCt 21. A considerable part of the BR 1 gene consists of one or a few blocks of a tandemly repeated 246-base-pair (bp) major repeat unit. About half of this major repeat unit is in turn built from four tandem repeats of a 33-bp sequence. This hierarchical arrangement of repetitive sequences within the BR 1 gene suggests that the gene has evolved through two major amplification steps, starting from a short primordial sequence. A similar evolutionary model has been put forward for the BR 2 gene (Suáregi, J., Wieslander, L. & Danenholt, B. 1982) Cell 30, 579-587). The two putative primordial genes contain a similar, 102-bp-long sequence (86% nucleotide sequence homology), indicating that the BR 1 and BR 2 genes most likely arose from the same ancestor sequence. During the course of evolution, the two genes diverged, mainly due to differences in the length and sequence of the gene segments involved in the two amplification steps. Moreover, at least one of the BR genes was translocated to another chromosomal locus.

The Balbiani ring (BR) 1 and BR 2 genes in Chironomus tentans are both located on chromosome IV. They are expressed only in the salivary glands (1, 2) and code for two large-sized secretory proteins (3). These proteins are used by the larvae to construct a supermolecular structure, the larval tube (4). We have recently described the repetitive structure of the BR 2 gene in detail (5). The repetitive sequences within the gene exhibit a hierarchical arrangement that suggests that the large BR 2 gene has evolved from a short primordial sequence. This sequence forms a nonrepetitive part of the major repeat unit in the present-day BR 2 gene. Recently, a short BR 1 specific sequence also has been analyzed by Degelmann and Holllenberg (6). We noted that this BR 1 sequence shows a high degree of sequence homology with the nonrepetitive region of the BR 2 major repeat unit.

The aim of this study was to investigate further the structural, and possibly also the evolutionary, relationship between the BR 1 and BR 2 genes. We have isolated a BR 1 cDNA clone and demonstrated that a large part of the BR 1 gene consists of tandemly repeated major repeat units with an internal organization analogous to that observed in the BR 2 gene. Furthermore, the BR 1 gene conforms to the evolutionary model suggested for the BR 2 gene and it is suggested from sequence comparisons that both genes originate from the same ancestral sequence and thus constitute two members of a gene family.

MATERIALS AND METHODS

Isolation of RNA and DNA. C. tentans DNA, salivary gland 75S RNA, and BR RNA were extracted according to Suáregi et al. (5).

Cloning Procedures. cDNA was synthesized from 75S RNA by reverse transcriptase by using random calf thymus oligodeoxynucleotides as primers. Double-stranded cDNA was inserted into the Pl site of pBR322 with the oligo(G)-oligo(C) tailing method, propagated in Escherichia coli HB 101, and screened for 75S RNA type sequences (for further experimental details, see ref. 5). The cloning procedures followed the National Institutes of Health guidelines (P2/3K 1 containment conditions).

Hybridization Techniques. In situ hybridization, filter hybridization, and hybridization in solution were carried out as described (5).

DNA Sequence Determination. The Maxam and Gilbert procedure was applied (7).

RESULTS

Construction of 75S RNA cDNA Clones and Selection of a BR 1 Clone. Salivary gland 75S RNA, containing mainly BR 1 and BR 2 RNA, was separated from other RNA species and transcribed into cDNA with random calf thymus oligodeoxynucleotides as primers. The cDNA molecules were subsequently cloned in the Pl site of pBR322. Individual recombinant plasmid DNAs were hybridized to cytological preparations of salivary gland chromosomes to identify BR 1-specific cDNA sequences. A majority of the cloned sequences was found to be BR 2 specific (57/64), probably reflecting the high abundance of BR 2 75S RNA during standard culturing conditions. Two BR 1-specific cDNAs were found; one of them, pCt 21, was chosen for further analysis. In cytological hybridization experiments the pCt 21 cDNA hybridizes predominantly to BR 1 and to some extent also to BR 2 (Fig. 1). Under more stringent hybridization conditions (see below), the pCt 21 cDNA hybridizes almost exclusively to BR 1. Therefore, the pCt 21 is likely to correspond to a BR 1 sequence, but evidently it also harbors sequences homologous, although not identical, to sequences in the BR 2 gene. We conclude that pCt 21 should be suitable for an analysis of a possible evolutionary relationship between the BR 1 and BR 2 genes.

Sequence Analysis of pCt 21 cDNA. To study the principal structure of the BR 1 gene, the restriction map and nucleotide sequence of the insert in pCt 21 were determined. Only a few different restriction enzyme sites are present in the pCt 21 insert and they are grouped characteristically at regular intervals (Fig. 2). The restriction map strongly suggests a periodic sequence arrangement with a periodicity of 240-250 base pairs.

Abbreviations: BR, Balbiani ring; kb, kilobase (pairs); bp, base pair(s).
The nucleotide sequence was determined according to the strategy outlined in Fig. 2. More than 95% of the pCt 21 cDNA sequence was established by sequence analysis of both strands, and each region was subjected to sequence analysis at least twice. Only one reading frame is open in the sequence, resulting in a protein with an amino acid composition characteristic of the putative BR 1 product (L. Rydlander, personal communication; lysine, serine, proline, and arginine are the most abundant amino acids). The 505-bp-long sequence flanked by the oligo(G)-oligo(C) tails and the amino acid sequence derived from the nucleotide sequence are presented in Fig. 3.

The pCt cDNA sequence exhibits a characteristic repetitive structure, with repetitive sequences on two different levels of organization (Fig. 3). First, the insert is built from a few major repeat units with a repeat length of 246 bp (see restriction map). Second, about half of each major repeat unit consists of four tandemly arranged 33-bp repeats; the remainder of the major repeat unit displays no obvious internal repeats. For simplicity, we have designated the two regions of the major repeat units as the repetitive and nonrepetitive regions, respectively.

Two repetitive regions are represented in the pCt 21 insert and have been indicated in Fig. 3 with solid lines; the vertical bars that cross the lines demarcate the 35-bp repeats. The two 132-bp-long repetitive regions differ at only one base pair position (number 183 in the first region and number 429 in the second region). When the four 33-bp repeats in each repetitive region are compared with a consensus sequence, there are five base pair differences in the first and four in the second repetitive region (Fig. 4). The distributions of the changes within the repetitive regions are remarkably similar; in fact, four of the five changes in the first repetitive region attain the same positions within the region as the four in the second repetitive region. On the amino acid level, the two regions are identical. The 33-bp repeat corresponds to the amino acid consensus sequence Lys-Pro-Ser-Lys-Gly-Ser-Lys-Pro-Arg-Pro-Glu. The four repeats exhibit the same amino acid sequence except for a substitution of Glu for Lys in the first repeat and of Glu for Gly in the second repeat. It can be concluded that there are several differences between the short repeats within a repetitive region, whereas the two repetitive regions are strikingly similar on the nucleotide level and identical on the amino acid level.

The single complete nonrepetitive region of the pCt 21 insert comprises 114 bp. Moreover, the insert starts and ends within incomplete nonrepetitive regions. Because these sequences seem to overlap, we have combined them to describe the properties of an additional 114-bp nonrepetitive region. The two regions are almost identical, showing the same amino acid sequence and differing in only 1 bp (numbers 36 and 282, respectively, in Fig. 3). Twelve different amino acids are represented in each nonrepetitive region.

It should be added that the nonrepetitive region of the major repeat unit in pCt 21 is almost identical to the short BR 1 sequence (147 bp) analyzed by Degelmann and Hollenberg (6). Our sequence differs at three base pair positions from that of Degelmann and Hollenberg. It is quite possible that the observed differences may be due to a minor divergence between the various nonrepetitive regions within the BR 1 gene or, alternatively, they may be due to polymorphism.

The Abundance and Arrangement of the pCt 21 cDNA Sequence in the BR 1 Gene. To obtain further information on the structure of the BR 1 gene, we have studied the occurrence and organization of the pCt 21 sequence within the BR 1 gene. Because about 75% of the BR 1 gene contains repetitive sequences (8), we considered it important to decide first to what extent the pCt 21 clone reflected the structure of the BR 1 gene, 37 kilobase pairs (kb) in size (9). Therefore, BR 1-specific 3H-labeled 7SS RNA was hybridized to a large excess of pCt 21 cDNA under conditions favoring DNA-RNA hybrid formation. Twenty-four percent of the RNA was found to be in RNase-resistant hybrids. Under the same stringent conditions, very little cross-hybridization with BR 27SS RNA (1.7%) could be detected. Therefore, the pCt 21 insert must be complementary to a large part of the BR 1 gene. It is difficult to calculate the exact proportion of the BR 1 gene complementary to the pCt 21 cDNA based on a 24%
RNAse resistance, because a 100% resistance cannot be experimentally reached even with a perfectly complementary DNA sequence; hence, the recorded resistance represents a minimal value. Furthermore, the BR 1-specific RNA was obtained from microdissected BRs that contain nascent 7SS RNA populations, in which the sequences at the 5' end of the gene are overrepresented and sequences at the 3' ends are underrepresented. Taking these factors into account, we estimate that no less than 15%, and perhaps as much as 55%, of the BR 1 gene is complementary to the pCt 21 cDNA insert.

The arrangement of the pCt 21 cDNA sequence in the BR 1 gene was studied in a Southern-type experiment. Total C. tentans DNA was partially cleaved with Mbo II. The DNA fragments were separated by electrophoresis and probed with the 32P-labeled pCt 21 cDNA insert (Fig. 5). A regular ladder pattern with a periodicity of about 240 bp was observed. This result implies that the pCt 21 cDNA insert is complementary to sequences building a periodic sequence block, which, according to Fig. 5, should be at least 6 kb in length.

In conclusion, a large part of the BR 1 gene consists of tandemly arranged major repeat units forming an uninterrupted sequence block that is at least 6 kb long and perhaps as long as 20 kb. It cannot be ruled out that several blocks may be present, separated by other sequences. Within the sequence block(s), the major repeat units are not likely to be separated by intron sequences, because the size of the major repeat units in the pCt 21 cDNA corresponds to the periodicity in the genomic ladder pattern.

**Sequence Comparison Between the BR 1 and BR 2 Genes.**

The BR 1 major repeat unit, 246 bp in size, was compared with the recently described, somewhat smaller major repeat unit in the BR 2 gene (about 215 bp) (5). The two repeat units exhibit the same principal structure, each consisting of a nonrepetitive part and a repetitive part. In the repetitive region, the BR 1 gene has four 33-bp repeats, whereas the same region in the BR 2 gene comprises about six 18-bp repeats.

To make a detailed sequence comparison feasible, we have presented the nucleotide sequence of the BR 1 and BR 2 major repeat units in Fig. 6; the nonrepetitive regions and the first repeat of the repetitive regions are displayed. It is evident immediately that the two BR genes show striking similarities. Based on the nucleotide and amino acid sequences, we have demarcated a high homology region (102 bp, corresponding to 34 codons) (thick line in Fig. 6). In both genes this region corresponds to almost the entire nonrepetitive region and, in the case of the BR 2 gene, it extends 3 bp into the repetitive region.
The degree of homology between the two 102-bp segments amounts to 86% on the nucleotide level and to 79% on the amino acid level (insertions and deletions allowed). The repeats of the two repetitive regions are less similar and differ in size (33 bp vs. 18 bp) as well as in position relative to the homology region (Fig. 6). However, it is striking that both the 33-bp and the 18-bp repeats contain a 9-bp sequence, AAA-CCA-AGC (Lys-Pro-Ser), or a minor variant of this theme (the sequence has been underlined in Fig. 6). Moreover, as indicated in Fig. 6, we noted that this very sequence is also present at the adjoining end of the 102-bp homology region. The possible significance of this finding will be commented upon below. We conclude that the remarkable sequence similarity in the major repeat units of the BR 1 and BR 2 genes, in particular in the nonrepetitive regions, strongly suggests that the BR 1 and BR 2 genes belong to the same gene family and might have a common origin. This conclusion is further strengthened when the evolution of the BR genes is considered on the basis of the hierarchic arrangement of the repetitive sequences in the BR genes (see Discussion).

**DISCUSSION**

The main feature of the analyzed large region of the BR 1 gene is the hierarchic arrangement of the repetitive sequences; the 246-bp repeat unit is tandemly arranged and about half of each major repeat unit consists of tandem repeats of a 33-bp sequence. A similar arrangement, in principle, is present also in the functionally related BR 2 gene. A detailed analysis of the BR 2 gene has led to the proposal that this gene has evolved from a short primordial sequence that is 110–120 bp long (5). During evolution this unit expanded in a first step by an amplification of a minor segment of the unit, resulting in an entity consisting of the remainder of the primordial sequence and an (approximately) equally large block of the amplified segment. Subsequently, the entire unit was amplified a large number of times forming the long array of major repeat units characteristic of the present-day structure of the BR 2 gene.

Most likely, the BR 1 gene has evolved in a similar manner. First, when the four 33-bp repeats in each repetitive region of the major repeat units were compared with the consensus sequence, it was noted that five base pair changes had occurred in the first and four base pair changes in the second repetitive region. No less than four of the five changes in the first repetitive region appeared at identical positions in the second region. This crucial observation can be understood best, as suggested in the evolutionary model, if the amplification giving rise to the block of 33-bp repeat units preceded the amplification, result-
ing in the long array of major repeat units. After the first amplification some sequence divergence took place, the result of which is apparent in all the major repeat units after the second round of amplification. The major repeat units in the BR 1 gene seem to have the same length and are almost identical at the sequence level; we assume that the few base differences present appeared after the establishment of the array of the major repeat units.

The evident similarities in the principal structure of the BR 1 and BR 2 genes and the proposed consequential similarity in evolutionary history focus the interest on the structure of the two primordial genes and their sequence relationship. If we assume that the nonrepetitive region and the first repeat of the repetitive region of the BR 1 gene (114 bp and 33 bp) and the corresponding parts of the BR 2 gene (105 bp and 18 bp) represent the structure of the BR 1 and BR 2 primordial genes, we can make direct comparisons. It is of interest then to note that a major part of the two putative primordial genes exhibits a conspicuous similarity: a 102-bp region can be recognized, showing 96% homology on the nucleotide level (Fig. 6). We conclude that the two BR genes are closely related and probably emanate from the same ancestral sequence, the structure of which should be reflected, at least partially in the 102-bp homology region.

After duplication, the ancestor sequence participated in the formation of the two primordial genes, which differ in length and, perhaps more important, in the size and sequence of the segment being amplified. It could be proposed that this latter difference is due to the incorporation into the primordial genes of sequences quite unrelated to the 102-bp ancestral sequence and specific for each BR gene. The major alternative would be that the two primordial sequences originate in their entirety from one and the same ancestral sequence but that the two genes evolved differently. We favor the latter alternative mainly on the basis of two observations. First, the segment of the BR 2 primordial gene to be amplified extends into the homology region (Fig. 6). Therefore, at least part of this segment of the BR 2 gene is likely to have originated from the common ancestral sequence. Second, in both genes a 9-bp sequence, A-A-A-C-C-A-A-G-C, or a slight variant of this sequence, constitutes the major part of each amplified segment (the 9-bp sequence has been underlined in Fig. 6). Furthermore, this very sequence also appears at the end of the homology region, indicating for each region a relatedness between the homology region and the remainder of the primordial gene (Fig. 6).

For more conclusive statements on the processes involved in the formation of the primordial BR 1 and BR 2 genes and the selection of the segments to be amplified, it is necessary to compile more information on the transition regions between the repetitive and nonrepetitive parts of the major repeat units within the BR 1 and BR 2 genes and perhaps also in other BR genes in Chironomus as well as in other Chironomus species. For example, comparisons between C. tentans and C. thummi might turn out to be quite valuable, because Wobus et al. (10) have described a BR sequence, a part of which (79 bp) is highly homologous (92%) to the nonrepetitive region of the major repeat units in the BR 1 and BR 2 genes. Finally, it remains to be established at what stage in the evolutionary process the BR 1 and BR 2 genes attained their present-day positions on chromosome IV (segment 2A and 3B, respectively); because they are now widely separated on chromosome IV, at least one of them must have been translocated during evolution from the locus of the ancestral sequence to its present chromosomal site.

We are grateful to Dr. L. Rylander for communicating unpublished results. Dr. J. W. Beard, Life Sciences, generously supplied reverse transcriptase. We are indebted to Barbro Dahlberg, Catherine Engberg, and Elisabeth Long for technical assistance and to Kerstin Yettman for typing the manuscript. The research was supported by the Swedish Cancer Society, the Swedish Natural Science Research Council, Magnus Bergvalls Stiftelse, Gunvor och Josef Anders Stiftelse, and Karolinska Institutet (Reservationsanslaget). J. S. obtained a Guest Research Fellowship from the Swedish Natural Science Research Council and a Postdoctoral Fellowship from the European Molecular Biology Organization.