The basic repeat unit of a *Chironomus Balbiani* ring gene

Helmut Bäumlein^{*}, Ulrich Wobus^{*}, Susan A. Gerbi[†] and Fotis C. Kafatos[‡]

^{*}Zentralinstitut für Genetik und Kulturpflanzenforschung, Akad. Wiss DDR, DDR-4325 Gatersleben, GDR, [†]Division of Biology and Medicine, Brown Univ., Providence, RI 02912, and [‡]Cellular and Developmental Biology, The Biological Laboratories, Harvard Univ., Cambridge, MA 02138, USA

Received 29 March 1982; Revised and Accepted 1 June 1982

ABSTRACT

A clone derived from the Balbiani ring b (BRb) gene of *Chironomus thummi* has been used to study the internal organization of that gene. Much of the gene consists of approximately 80 copies of a ca. 300 bp repeat unit, which are tandemly organized. The BRb clone contains a major part of that unit (242 bp). Sequence analysis shows that approximately 60% of the unit corresponds to short, tandemly organized subsequences, which encode peptides 8 to 11 residues long. In turn, each subsequence consists of even shorter internal repeats, corresponding to a tripeptide (consensus Proline.Serine.Lysine.). The remainder of the ca. 300 bp unit probably does not have obvious repetitive substructure.

INTRODUCTION

The Balbiani rings (BR), giant puffs in the *Chironomus* polytene chromosomes, have been studied extensively by cytogenetic, biochemical and, most recently, recombinant DNA procedures. These two to three major puffs contain genes that produce huge RNA transcripts (35-40 kb; Ref. 1-3). The BR transcripts are translated into giant (4) salivary proteins of unusual composition (5; A. Rudolph, L. Meyer and E. Serfling, unpublished results), which are spun into the elastic threads that make up the protective tube of the aquatic midge larva.

Whereas each BR gene is present in probably only one copy per genome, BR DNA appears to be highly repetitive (6-9). This paradox is explained by the fact that BR genes contain internal repeats. We have previously cloned a segment of the BRc gene of *C. thummi* (10) and have used it to demonstrate the existence of tandem arrays of a ca. 0.2 kb repeat unit. A similar tandemly repetitive element is present in the BRl gene of *C. tentans* (11).

Some DNA sequence data also exist for both BRc and BRl, but the regions sequenced represent either variant (10) or incomplete (11) repeat units.

We have obtained a 242 bp cloned DNA segment from the BRb gene of C. thummi. Although that clone is approximately 60 bp short of the full repeat length of BRb, it overlaps the homologous, BRl clone of C. tentans (11). Thus, these two clones combined permit us to define at the sequence level the full repeat unit of one type of BR gene. That unit, which is part of the translated transcript, is shown to consist partly of shorter, tandemly repetitive subunits, reminiscent of satellite DNA.

MATERIALS AND METHODS

The recombinant plasmid pCthBRb-16 was constructed from genomic DNA of Chironomus thummi thummi from a laboratory culture originating from larvae collected near Halberstadt, GDR. Cloning was as described (10, 12) except that A·T tails were used for insertion in the Bam HI site of pBR322 (insert tailed with T). Most methods used in the analysis of the clone have been described previously (10). For defining the transcribed strand of the cloned segment, strand-separated DNA fragments were transferred to DBM paper as described (13) and hybridized with alkali-fragmented, ³²P-γ-ATP labeled 75S BR-RNA for 7 hours at 65°C in a plastic bag containing 2xSSC, 5x Denhardt's solution, 0.2% SDS, 50 µg/ml E. coli tRNA, 10 µg/ml poly(A); prehybridization for two hours used the same solution at 65°C, but without labeled RNA. The paper was finally washed for 3x30 min., 3x15 min and 1x15 min in SSC solutions (2x, 1x and 0.5x, respectively, always containing 0.1% SDS and 0.1% Na-pyrophosphate) and exposed to ORWO HS11 x-ray film with an intensifying screen (Lightning Plus, du Pont) at -75°C. For sequencing by the Maxam-Gilbert methods, the four standard reactions plus a high-temperature, A>C reaction were used (14).

RESULTS

Construction and identification of a BRb clone.

Larval DNA fragments larger than 5 kb were selected by

Figure 1 In situ hybridization of ^3H -cRNA from clone pCthBRb-16 to Balbiani ring b of Chironomus thummi. Methods as in (10), exposure 10 days. Label is localized on BRb but not BRc, nor any other chromosomal region; the low amount of label at the nucleolus organizer (No) is unexplained.

sucrose gradient centrifugation, enriched for BR sequences by Actinomycin D-CsCl density gradient centrifugation (15), and cloned in the Pam HI site of pBR322 by the A·T tailing method. Transformed colonies were screened with iodinated 75S BR-RNA. The positive clone, pCthBRb-16, was shown by in situ hybridization to represent a BRb sequence (Fig. 1): in the in situ preparations, BRb was heavily labeled. Surprisingly, minor but consistent labeling of unknown significance was also observed in the nucleolar organizer; no labeling of BRc was evident under our conditions.

Clone pCthBRb-16 contains an unexpectedly small DNA insert, 242 bp. Although we cannot exclude other explanations (preferential cloning of contaminating small DNA fragments, tailing and therefore cloning at internal nicks), we suspect that the small size of the insert is due to post-cloning deletion(s), promoted by the internally repetitive nature of BRb DNA. Similarly, Degelmann and Hollenberg (11) consistently recovered small inserts (ca. 150 bp) when they attempted to clone 40 kb fragments of BR DNA from C. tentans. Furthermore, we have observed secondary deletions in a plasmid containing internally repetitive BRc sequences (10).

The pCthBRb-16 sequence represents part of a tandemly repetitive unit dominant in BRb DNA.

Southern hybridization of C. thummi genomic DNA, using nick-

Figure 2 Identification of the major BRb repeat unit by Southern hybridization. Lanes 1, 6, 10, 12 and 14 are markers of pCthBRC-19 DNA(10); remaining lanes are larval DNA of *Chironomus thummi*. For all samples the probe was nick-translated ^{32}P -pCthBRb-16 DNA, which hybridizes to the markers because of the shared pBR322 sequence and to chromosomal DNA because of the BRb DNA insert.

Lanes 1, 10, 12 and 14: Hinf I digested markers. Dots indicate bands corresponding, from top to bottom, to the following lengths (in bp; determined from the pBR322 sequence): 1435, 517 + 506; 396; 344; 298; 221 + 220. Lane 6: Alu I digested markers. Dots indicate bands corresponding, from top to bottom, to the following lengths (in bp): 910; 659 + 655; 520; 403; 281; 257; 226.

Lanes 2-5 and 7-9: 10 μg larval DNA digested fully with Hinf I (lanes 2, 4), fully with Alu I (lane 3), nearly fully with Alu I (lane 5), partially with Sau 3A (lane 7) and mildly with Ava II (lanes 8, 9: two different degrees of digestion). The repeat unit of BRb DNA can be estimated as 300 ± 20 bp on the basis of the samples in lanes 2, 4, 5, 8 and 9. Note in particular the 300 bp ladder in lanes 8 and 9 (dots); the presence of some off-ladder bands between 1200 and 1500 bp indicates limited heterogeneity in the repeat array. The two bands evident in Lane 7 (dots) are estimated as 280 and 170 bp, in good agreement with the sizes expected from Fig. 4a which maps Sau 3A sites at positions 5, 23, 41 and 173 within the 297 bp repeat. If a cut is made in, on the average, every third Sau 3A site, it would generate main fragments 279, 168 and 165 bp in length as observed in Lane 7.

Lanes 10-15: A reconstruction experiment that reveals the presence of ~ 80 copies of the ~ 300 bp BRb repeat unit per genome. The marker DNA samples were 34 ng (lane 10) or 17 ng (lanes 12, 14), and the *Chironomus* DNA 1.8, 3.6 and 5.4 μ g (lanes 11, 13 and 15, respectively; Hinf I-digested). From the nearly equal amount of radioactivity in the bands of comparable length in lanes 12, 14 and 15 one can calculate the copy number $X = (\text{grams of marker DNA}) \times (\text{genome size in bp [23]}) \div (\text{grams chromosomal DNA}) \times (\text{total bp of all marker fragments [10]}) = (17 \times 10^{-6} \times 1.4 \times 10^8) \div (5.4 \times 10^{-6} \times 5.7 \times 10^3) \approx 77$.

translated pCthBRb-16 DNA as probe (Fig. 2), revealed that BRb DNA has a tandemly repetitive substructure. When the chromosomal DNA was digested to completion with Hinf I, a single band hybridized with pCthBRb-16 DNA. The length of this band was estimated as 300 ± 20 , in side-by-side comparisons with sequenced DNA markers. According to reconstruction Southern, using known amounts of cloned and genomic DNAs, the hybridizing band was present in approximately 80 copies per genome (Figure 2). Furthermore, with chromosomal DNA mildly digested with Hinf I, Ava II or Alu I, 300 bp "ladders" were observed, extending to more than 15-mers. We conclude that a major part of the BRb gene consists of tandemly repetitive, 300 bp units, which hybridize with the pCthBRb-16 sequence. Each unit apparently contains a single Hinf I and a single Ava II site. A few off-ladder bands indicated that some of the repeat units are variant. Complete digestion of chromosomal DNA with Alu I yielded a 280 bp band that hybridized with pCthBRb-16; accordingly, the BRb repeat unit must contain at least two Alu I sites spanning 20 bp. Partial digestion with Sau 3A yielded approximately 280 and 170 bp fragments, which can be rationalized on the basis of several sites, one of which is non-clustered with the rest (see legend to Fig. 2, and Fig. 4a).

The repetitive units are transcribed, since they are represented in the 75S RNA. They are presumably translated, since their copy number suggests that they account for at least two thirds of the RNA length, and since the nascent polypeptide chain length (4) indicates that much of the transcript is translated.

Identification of the coding strand of pCthBRb-16.

To determine the amino acid sequence encoded by pCthBRb-16, we first determined the coding strand of the cloned insert. An

Ava II - Hae III fragment, containing 58 bp of the BRb sequence, approximately 80 bp of AT linker and 21 bp of pBR 322 DNA was strand-separated on a 5% polyacrylamide gel, transferred to DBM paper, and hybridized to 75S BR-RNA. Since there is a strong bias in the distribution of A and T in the two strands, strand separation was very effective. As shown in Figure 3, hybrids were formed exclusively with the slow-migrating, T-rich strand. Consistent results were obtained with the larger fragment extending from the Ava II site to the pBR322 Hae III site on the opposite side (data not shown).

Restriction and sequence analysis

Figure 4a presents the restriction map of the pCthBRb-16 insert, as determined by standard mapping techniques and subsequent sequence analysis. The insert contains a single Ava II site, and a cluster of three Sau 3A sites near one end plus a fourth one, 132 bp away. These sites are consistent with the results of Southern analysis of chromosomal DNA (Fig. 2); by subtraction we infer that the approximately 60 bp portion of the BRb repeat unit not represented in this clone contains one Hinf I site, at least two Alu sites spanning 20 bp, and no Ava II sites.

The entire insert was sequenced by the Maxam-Gilbert (14) procedure, after 5' end labeling at the unique Ava II site and strand separation. The sequence of the mRNA-synonymous A-rich DNA strand is presented in Fig. 4b, together with the derived

Figure 3 Determination of the transcribed strand in cloned BRb DNA. A ca. 160 bp AvaII-Hae III fragment containing both BRb sequence and AT linker (see Text) was strand-separated on a 5% polyacrylamide gel (14), transferred to DBM paper and hybridized with ^{32}P -labeled 75S BR-RNA. (a) Ethidium bromide staining pattern, (b) autoradiogram showing hybridization only with the slower, T-rich strand.

protein sequence: only one reading frame is open, the other two frames being interrupted by 1 or 12 termination codons, respectively. The correctness of the derived protein sequence is verified by the unusual amino acid composition; 73% of the residues are lysine, proline, serine or arginine, amino acids which are very abundant in total C. thummi or C. tentans saliva (Ref. 5; A. Rudolph, L. Meyer and E. Serfling, unpublished results.)

The BRb gene of C. thummi is thought to be homologous to the BRl gene of C. tentans (16). A portion of the BRl repeat unit has been sequenced by Degelmann and Hollenberg (11). Comparison of the two sequences revealed an unequivocal overlap (Fig. 4b): an uninterrupted segment of 70 bp is represented in both of the available partial sequences, with substitutions in only 16 positions (77% DNA homology; 74% protein homology; maximum complete match 22 bp). The overlap permits the construction of a composite BRb/BRl sequence that spans 323 bp, and therefore should encompass the entire repeat unit (ca. 300 bp for BRb; Fig. 2). Indeed, we note that a stretch of 20 bp at the end of the BRl sequence (nucleotides 304 to 323 in the composite numbering of Figure 4b) corresponds to a similar (80% homologous) stretch near the beginning of the BRb sequence (nucleotides 7 to 26 in Fig. 4b). Assuming these stretches are equivalent, the BRb/BRl unit is 297 bp long. The correct strand and frame in which the BRb/BRl sequence is translated agrees with those inferred by Degelmann and Hollenberg (11). The encoded protein sequence will be further discussed in the Discussion.

DISCUSSION

We have obtained a cloned fragment of the BRb gene of C. thummi. When this clone is used as probe in Southern hybridizations of partially and completely digested chromosomal DNA, it is seen to represent a unit which is tandemly repeated within the gene in array(s) of at least 15 copies, with a total repetition frequency of approximately 80, and thus accounts for a major part of the 35 to 40 kb long BRb gene. The tandemly repetitive internal structure of the BRb gene is shared with its homologue in C. tentans, BRl (11) and with the other major Balbiani ring of

Figure 4a (top) Restriction map of the reconstructed BRb/BRl repeat unit. The cloned portions of BRb (this report) and BRl (11) are shown as open and filled bars, respectively. Restriction sites are numbered in bp from the left end.

Figure 4b (bottom) Sequence and conceptual translation of the BRb/BRl repeat unit. The sequences derived from BRb and BRl

(this report and Ref. 11, respectively) are appropriately identified, aligned according to their overlap, and numbered from the first nucleotide of the available BRb sequence. Tandemly repetitive subunits are boxed (See Fig. 5). Residues of BRl are shaded if known to differ relative to BRb. Underlining indicates that nucleotides 302-323 are considered to be equivalent to nucleotides 7-16, defining a 297 bp repeat unit (in brackets).

C. thummi, BRc (10). Our cloned BRb fragment, together with the BRl clone of Degelmann and Hollenberg (11), define for the first time the full repeat of one type of Balbiani ring gene.

The size of the BRb repeat is estimated as 300 ± 20 , from Southern analyses of chromosomal DNA using sequenced DNA fragments as markers (e.g. Fig. 2). We are confident of this estimate (which corrects an earlier estimate of 200 bp; Ref. 10), since it is based on side-by-side comparisons of both complete digests and partially digested "ladders" with sequenced markers. Degelmann and Hollenberg (11) estimated from Southern blots that the homologous BRl unit of C. tentans is about 240 bp long. Therefore, since their last common ancestor either BRb has increased in length or BRl has contracted. According to recent sequence and Southern blot data, the size of the repeat unit in the other major Balbiani ring of C. thummi, BRc is 249 bp (Bäumlein et al., in preparation), instead of 175 bp as originally estimated (10).

The homology of BRb and BRl (16) is reflected in the 70 bp segment which is shared between the two respective partial sequences (Fig. 4b). The two sequences are read in the same translational frame; and in combination correspond to a polypeptide that shares the very unusual amino acid composition of Chironomus saliva. The combined sequence (323 bp) represents more than one complete copy of the tandemly repetitive unit, although it is a consensus of two sequences that apparently differ by 23% in mismatching and approximately 25% in length.

The repeat unit contains a tandem array of smaller subunits. As shown in Fig. 5, approximately 60% of the length of the combined BRb/BRl sequence consists of tandem, inexact repeats of an 11 residue polypeptide unit, with the consensus structure PSKPSKSGSKPR (Proline.Serine.Lysine.Proline.Serine.Lysine.Glycine.Serine.Lysine.Proline.Arginine). This subunit in turn contains even shorter tripeptide repeats (consensus Proline.Serine.Lysine),

Figure 5 Diagrammatic representation of the polypeptide repeat unit encoded by BRb/BRl DNA. The amino acids are shown in the one-letter code. Tandem 8 to 11 residue subunits are vertically stacked and aligned; they correspond to nucleotides 1 to 186 (Fig. 4b) and should be read left to right, top to bottom. The non-repetitive segment (corresponding to nucleotides 187 to 303) is offset to the right. Conserved residues in the repeat subunit are shown in white on black background. Italics indicate residues sequenced only in BRl or known to differ in BRl as compared to BRb.

which differ substantially at the nucleotide level. In sum, a major part of this BR DNA is reminiscent of certain satellite sequences, in consisting of two or three tiers of repeats: possibly 9 bp within 33 bp within the overall unit of 297 bp.

This hierarchically repetitive structure is unequivocally established for the BRb sequence. However, we presume that it is also applicable to BRl, since the available sequence of the latter begins and terminates with partial copies of the 11-residue repeat (. . .SKPR and SEPSKGS, . . ., respectively). In BRb DNA, the short subunits have diversified through deletions or insertions, as well as replacements (Fig. 5). Tandemly repetitive sequences are known to be hot spots for segmental mutations (17). We predict that the apparent difference in length between the BRb and BRl units may result from deletions/reduplications in the short-subunit region.

The remainder of the sequence (39 codons in BRl) is apparently conservative and not easily subject to segmental mutations: in the overlap of BRb and BRl, 19 amino acid residues are represented in both sequences, with 5 replacements but without deletions or insertions. Furthermore, this 39-residue section also shows unmistakable homology with part of the BRC sequence

(Bäumlein et al., in preparation). No obvious repeat subunits are evident in this portion, although a short polypeptide, Arginine.Cysteine.Glycine (Alanine).Serine.Alanine, is repeated twice, in a non-tandem manner. The extent to which the tandem short-subunit repeats and the apparently non-repeated remainder of the sequence may be related, and the evolutionary relationships of all available BR sequences, are currently under investigation by computer-assisted analysis (J. Pustell, personal communication). Additional Southern analysis using double digests of C. thummi DNA, and ultimately recovery of a larger BRb clone will be necessary for confirming the model of the BRb/BRl unit presented in this report.

In recent years, it has become evident that many structural proteins have an "accordion" structure (terms proposed by M.A.T. Muskavitch, Ref. 18): they consist to a major extent of tandem internal repeats which are subject to change in number, as well as to internal diversification. Examples would include fibroin (19), chorion proteins (17,20), keratins (21), Drosophila salivary glue proteins (22) and Chironomus BR-encoded proteins (Refs. 10,11, Bäumlein et al., in preparation; this report). At least in the case of the insect proteins listed above, no introns exist between the internal repeats. Continued sequence analysis of BR genes should help illuminate the evolutionary mechanisms of "accordion" protein-encoding genes. To date, the BR sequence comparisons suggest that the evolutionary mechanisms of satellite DNA diversification may not be as specialized as previously thought, but may operate in the generation and diversification of important types of structural genes, as well.

ACKNOWLEDGMENTS

We thank M. Hartmann and F. Walter (Jena) for continuous supply with restriction enzymes, and G. Hesse (Jena) for the fluorometric estimation of the deoxyribose content of DNA preparations used in the gene titration experiments. H.B. wishes to thank W. Ross and R. Gourse for fruitful discussions and help in sequencing; his stay in S.A.G.'s lab was supported by grant PHS-GM 26070. H.B. and U.W. acknowledge the continuous stimulating discussions on the subject with R. Panitz and E. Serfling.

REFERENCES

1. Case, S.T. and Daneholt, B. (1978) J. molec. Biol. 124, 223-241.
2. Serfling, E. and Huth, A. (1978) Chromosoma 66, 205-223.
3. Hardy, P.A. and Pelling, C. (1980) Chromosoma 81, 403-417.
4. Rydlander, L., Pignon, A. and Edstrom, J.-E. (1980) Chromosoma 81, 101-113.
5. Grossbach, U. (1977) Res. Prob. Cell Different. 8, 147-196.
6. Lambert, B. (1972) J. molec. Biol. 72, 65-75.
7. Hollenberg, C.P. (1976) Chromosoma 57, 185-197.
8. Wobus, U. and Serfling, E. (1977) Chromosoma 64, 279-286.
9. Wieslander, L. (1979) J. molec. Biol. 134, 347-367.
10. Wobus, U., Bäumlein, H., Panitz, R., Serfling, E., and Kafatos, F.C. (1980) Cell 22, 127-135.
11. Degelmann, A. and Hollenberg, C. (1981) Chromosoma 83, 295-313.
12. Bäumlein, H., Wobus, U. and Panitz, R. (1980) In DNA-Recombination Interactions and Repair, S. Zadravil and J. Sponar, eds., pp. 73-75, Pergamon Press, Oxford and New York.
13. Gourse, R.L. and Gerbi, S.A. (1980) J. molec. Biol. 140, 321-339.
14. Maxam, A. and Gilbert, W. (1980) Methods Enzymology 65, 499-580.
15. Bäumlein, H., Serfling, E. and Wobus, U. (1981) Biol. Zbl. 100, 285-294.
16. Santa-Cruz, M.C., Villanueva, A. and Diez, J.L. (1978), Chromosoma 69, 92-100.
17. Jones, C.W. and Kafatos, F.C. (1982), J. molec. Evolution, submitted.
18. Muskavitch, M.A.T. (1980) Ph.D. Diss., Stanford Univ. School of Medicine, Stanford, Calif., USA.
19. Sprague, K.U., Roth, M.B., Manning, R.F. and Gage, L.P. (1979) Cell 17, 407-413.
20. Jones, C.W., Rosenthal, N., Rodakis, G.C. and Kafatos, F.C. (1979) Cell 18, 1317-1332.
21. Rogers, G.E. (1978) Trends in Biochem. Sci. 3, 131-133.
22. Muskavitch, M.A.T. and Hogness, D.S. (1981) Proc. Natl. Acad. USA 77, 7362-7366.
23. Wobus, U., (1975) Eur. J. Biochem. 59, 287-293.