RESEARCH

Telomeric Organization of a Variable and Inducible Toxin Gene Family in the Ancient Eukaryote *Giardia duodenalis*

Peter Upcroft,[1] Nanhua Chen, and Jacqueline A. Upcroft

Queensland Institute of Medical Research, The Bancroft Centre, Brisbane, Queensland 4029, Australia

Giardia duodenalis is the best-characterized example of the most ancient eukaryotes, which are primitively amitochondrial and anaerobic. The surface of *Giardia* is coated with cysteine-rich proteins. One family of these proteins, CRP136, varies among isolates and upon environmental stress. A repeat region within the *CRP136* family is interchangeable by a cassette-like mechanism, generating further diversity in repeat size, copy number, and sequence. Flanking the 5′ region of the *CRP136* family is a novel protein kinase gene and an ankyrin homolog, creating a conserved unit. A short spacer separates the ankyrin gene from the variable, tandem array of rDNA gene units at a common breakpoint within the large subunit gene, which is followed by the $(TAGGG)_n$ telomeric sequence. Transcriptional up-regulation of the *CRP136* family is accompanied by a switch in mRNA length and promoter, or de novo expression, and suggests that *CRP136* mRNA induction is under the control of a telomerically regulated position effect, which evolved very early in the eukaryotic lineage.

[The sequence data of the telomeric gene units TGU1 and TGU2 described in this paper have been submitted to GenBank under accession nos. L49236 and L49298, respectively.]

Giardia duodenalis infects some 300 million people worldwide, and is capable of causing long-term diarrhea, malabsorption, failure to thrive, and death (Warren 1989). In most Western countries, giardiasis is the most commonly reported parasitic infection. *Giardia* is also regarded as representative of the most ancient eukaryotes before they fused with purple nonsulfur bacteria to generate mitochondria (Sogin et al. 1989; Upcroft et al. 1990; Cavalier-Smith 1993). Cavalier-Smith (1993) has argued on morphological, organelle, and 16S rRNA sequence data that the Metamonada (including *Giardia*) and the Archamoebae are so evolutionarily distinct that they should be placed in a separate kingdom, the Archezoa, bridging the gap between the bacteria and the remainder of the eukaryotes. Recent work emphasizes the ancient, bacterial nature of metabolic pathways in *Giardia* (Mertens 1993; Townson et al. 1994, 1996; Brown et al. 1995, 1996).

No introns have yet been found in *Giardia* genes, even those encoding over 1300 amino acids (Chen et al. 1995, 1996a). Intergenic spacers vary from zero to a few hundred bases. Promoter regions and polyadenylation signals are usually quite close to the coding region of each gene (Upcroft et al. 1990; Adam 1991).

We have shown that *Giardia* does not appear to have a classical chromosomal karyotype. Chromosome copy number and size vary dramatically among isolates (Upcroft et al. 1989, 1993, 1995, 1996) and allow a more dynamic flux of genetic information. However, there is sufficient genetic order for each genotype to be viable and to maintain distinct chromosomes plus partial duplications (Upcroft et al. 1993, 1995, 1996; Chen et al. 1994; Le Blancq 1994).

Gene gain and loss in *Giardia* is concomitant with chromosomal rearrangement (Upcroft et al. 1993, 1996; Chen et al. 1994). For example, the rRNA gene repeat unit has been sequenced (Healey et al. 1990) and the copy number can vary from a few tandem repeats to over 100 copies arrayed on single or multiple chromosomes (Upcroft et al. 1990; Adam 1992) and on partially duplicated chromosomes (Chen et al. 1994; Le Blancq 1994; Upcroft et al. 1996). There are two examples where the rRNA gene unit of *Giardia* was found immediately juxtaposed to the telomeric repeat sequence $(TAGGG)_n$

[1]Corresponding author.
E-MAIL peteru@qimr.edu.au; FAX 61 73362 0105.

(Adam et al. 1991), suggesting that the rRNA genes may be clustered at telomeric locations.

The surface of *Giardia* is covered with cysteine-rich proteins (Mowatt et al. 1991) that can vary among different isolates as well as in a single isolate (Nash 1992). The putative roles for these proteins include protease and bile salt resistance and heavy metal resistance (Chen et al. 1996a). Resistance to metronidazole, the most common antigiardial drug, can be induced readily by persistent exposure (Townson et al. 1992). From one such resistant line, two members of a new family of cysteine-rich protein (CRP) genes were found expressed at very high levels and have been cloned recently. The first, *CRP136*, has an internal repeat sequence encoding 23.5 copies of a 40-amino-acid repeat motif related to the snake sarafotoxin gene family (Chen et al. 1995). The second member, *CRP65*, has 100% homology to *CRP136* at the short 5′ terminus and 85% homology at the 936-bp 3′ terminus, but has only four copies of an unrelated 76-amino-acid epidermal growth factor (EGF)-like motif replacing the central repeat motif in CRP136. 5′ remnants of the *CRP65* repeat unit have been left immediately flanking the repeat units in *CRP136* indicating exchange of the central repeated domains by a cassette-like mechanism (Chen et al. 1996a).

Here we analyze the gene arrangement surrounding the variable *CRP* genes and show that there are evolutionarily conserved clusters of membrane-spanning/matrix/cytoskeleton genes with parasite–host signaling and effector molecule domains capable of modulating virulence (Chen et al. 1995; 1996a). These clusters immediately juxtapose the rDNA repeat arrays that in turn are flanked by the telomeric repeats of $(TAGGG)_n$. The gene clusters are mobile within the genome and among telomeres on different chromosomes. Different isolates maintain different subsets of the clusters. Furthermore, transcriptional switching between these subsets occurs under changes in growth conditions. This arrangement is reminiscent of the developmentally regulated position effects at the telomeres of *Trypanosoma brucei* chromosomes (Vanhamme and Pays 1995), and suggests that such mechanisms had evolved in the very earliest eukaryotes.

RESULTS

Cloning and Sequence Analysis of the Upstream Genomic Region Flanking *CRP136* and *CRP65*

The genomic DNA sequences upstream from *CRP136* and *CRP65* were cloned according to the procedure described previously (Chen et al. 1995, 1996a). Briefly, the 460-bp *Pst*I–*Sac*I-cleaved segment from the 5′ terminus of the D1 insert containing *CRP65* (Chen et al. 1996a) was used to select the upstream DNA segment by means of hybridization to genomic DNA from *G. duodenalis* strain WB1B cleaved with *Sac*I. A size-specific genomic DNA library was screened to select a clone (D3) containing a 1-kb segment upstream from the D1 insert (Fig. 1). The DNA segments upstream from D3 (D4, D5, and D6) were obtained by the same procedure (Fig. 1). The DNA segments upstream from *CRP136* (Chen et al. 1995) were cloned in a similar fashion. Systematic deletion subclones of these inserts were constructed from both directions using DNase I in the presence of Mg^{2+} (Lin et al. 1985).

Sequence analysis of the subclones identified two open reading frames (ORFs) upstream from both *CRP136* and *CRP65*. The *CRP* genes are arranged head to head with one of the new ORFs in each case, whereas the second ORF is located downstream from the first in the same orientation (Fig. 1). The two ORFs upstream from *CRP65* are 1329 bp and 2013 bp in length, respectively, whereas those upstream from *CRP136* are 900 bp and 1179 bp, respectively. The putative polyadenylation signal for *G. duodenalis* genes, AGTAA (Adam 1991), was identified downstream from ORF2 in each case, but not downstream from ORF1.

Gene Homologs Upstream from *CRP65* and *CRP136*

Homology comparisons of the translation products of ORFs1 and ORFs2, using the TFASTA program (Lipman and Pearson 1985) implemented in the GCG suite of programs (Devereux et al. 1984) revealed that both ORFs1 are related to protein kinases (Chen et al. 1996b) and have 83% homology to one another. We have designated them *gPK1* and *gPK2*. Similarly, both ORFs2 are homologous and composed of ankyrin-like repeats, and they have been designated *ANK1* and *ANK2*. *ANK1* and *ANK2* are composed entirely of ankyrin repeats, similar to homologs found in bacteria rather than the higher eukaryotes (Dolata et al. 1993). *ANK1* and *ANK2* have 81% homology (70% identity); *ANK1* encodes 19 repeats having 55% homology (31% identity) with the human erythrocyte ankyrin repeat regions, whereas *ANK2* encodes 11 repeats with 58% homology (35% identity).

Nucleotide sequences were compared using the BESTFIT program (Smith and Waterman 1981). *CRP65* and *CRP136* are > 80% identical over their

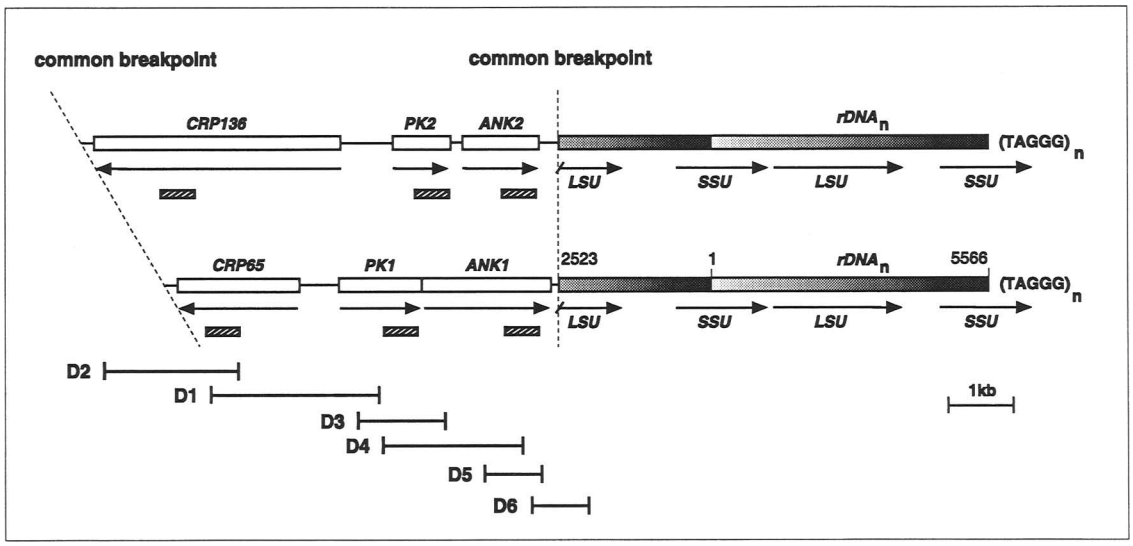

Figure 1 Comparison of the two telomeric regions containing *CRP136* and *CRP65*. The organization of TGU1 and TGU2, and their surrounding gene regions, including the telomeres, has been compiled from the sequence data and hybridization data. The hatched blocks show the regions that were used as probes for each gene region in Fig. 4. The common breakpoints preceding the CRP genes and at the point of insertion into the LSU rDNA genes is also shown. The rDNA genes have not been sequenced beyond the first 47 bp, so the exact copy number of rDNA units in the array prior to addition of the telomere is not certain, as described in the text. The position of the small subunit and LSU rDNA coding regions is marked (Healey et al. 1990). Horizontal bars represent overlapping cloned segments. The clone D1, a 2.7-kb *Pst*I segment, was used as the starting point for walking in the region of *CRP65* in the genome of *G. duodenalis* strain WB1B. D2 was a 1.9-kb *Stu*I segment, D3 a 1-kb *Sac*I segment, D4 a 2-kb *Pst*I segment, D5 a 0.8-kb *Sph*I segment, and D6 a 0.8-kb *Hin*dIII–*Sac*II segment. The sequence of the genes and spacer regions surrounding *CRP136* was determined in a similar fashion. Horizontal arrows represent transcriptional directions in the ORFs. The spacer upstream from *ANK1* and *ANK2* is inserted into the LSU rDNA gene at position 2523 (Healey et al. 1990).

entire length except for the repeat regions, which encode a sarafotoxin-like product in the case of *CRP136* and EGF homologs in the case of *CRP65*. The nonrepeat regions of *CRP136* and *CRP65* are 85% identical at the 3′ end and 95% identical at the 5′ end (Chen et al. 1995, 1996a). The spacer regions between the *CRP* genes and the *gPK* genes are 89% identical, and the equivalent regions between the *gPK* and *ANK* homologs are 82% identical. The two gene blocks, consisting of *CRP*, *gPK*, and *ANK*, and their intergenic spacers are > 80% homologous over their entire lengths. Homology downstream from the *CRP* genes continues for 220 bp, after which there is a clear breakpoint (Fig. 2). The following spacer region and genes are quite different (data not shown).

Genome Organization Downstream from the Ankyrin Homologs *ANK1* and *ANK2*

Flanking the two *ANK* homolog genes, on the opposite side to the *CRP* and *gPK* blocks, are intergenic spacers. Four base pairs 3′ to *ANK1* is a polyadenylation signal (AGTAAAC); the same signal is 75 bp 3′ to the *ANK2* gene (Fig. 3). However, these regions

```
 898 TAGACTTAGCTGTGTACTTAGATAGTAAACGCGTCACTGTATAAAGCATA  849
     ||||||||||||||||||||||| |||||||||||| ||| |||  |||
1400 TAGACTTAGCTGTGTACTTAGGTAGTAAACGCGTTACTTTATGTAGCTCG 1351

 848 TTATACTGTGCTGGGGCCCTGCCTGTAGAT........GCGCTTGGAAAT  807
     |      |||||||          ||| || |         ||| |
1350 TGTAGATGTGCTG.........CTGGAGCTATCTGAGCGCGATCCAGGGT 1310

 806 CCCACAAGTAAGCCTCGACAGTTCCTGGCCAGTGCAGAGACGAGGATTGG  757
     |   |  ||||  |||||     ||||||||| || | || |||| |  |||
1309 CATGCGGGTAATCCTCGGTCTTTCCTGGCC.GTCCGGAAACGATGGCTGG 1261

 756 TTGTAGGTTACATACTAGGGGCTCCTCTCTCTATTGGTGCATCCACAGTCTG  707
     | || | | |  ||    ||   |||||| | || ||||||| || |
1260 TCGTGGATAATGTAACGAGG...CCTCTCCAGTGATGCATCCAGAGCCCT 1214

 706 GTACATTCCTCGGCCGTTCT  687
     |  |  ||| ||| ||  ||
1213 GCGCGGTCCCCGGTCGGCCT 1194
```

Figure 2 DNA sequence at breakpoint downstream from CRP genes. The top line is the sequence of *CRP65* at the 3′ terminus of the gene followed by a spacer. The second line is the comparable sequence for the *CRP136* 3′ region. The stop codons for both CRP genes are bold and marked at the start of the sequences. The polyadenylation signals are underlined. The regions of significant homology cease at the end of the sequences.

are still 80% homologous because of the different lengths in the genes and spacers. Following the polyadenylation signals, homology drops to 60% over 27 bp for *ANK1* and 30 bp for *ANK2*, after which there is a region in both spacers of 14 bp with 100% homology. Both spacers are then fused at the same site, position 2523 (Healey et al. 1990), of truncated copies of the large subunit (LSU) rRNA gene. The LSU rRNA gene is part of a very small and compact gene unit containing the small subunit, the LSU, and the 5.8S rRNA gene equivalents (Healey et al. 1990; Upcroft et al. 1990, 1994). Surprisingly, there is 100% homology between the two copies of the truncated and therefore inactivated LSU gene.

Telomeric Location of the Gene Blocks Encoding *CRP*, *gPK*, *ANK*, and the Fused rDNA LSU Gene

The rDNA repeat units are located as tandem arrays of variable numbers of such units on single or multiple chromosomes (Upcroft et al. 1990; Adam 1992) and on partially duplicated chromosomes (Chen et al. 1994; Le Blancq 1994; Upcroft et al. 1996). Telomeric sequences have been located to two rDNA repeat units and it has been suggested that the rRNA genes may be clustered at telomeric locations on *Giardia* chromosomes (Adam et al. 1991).

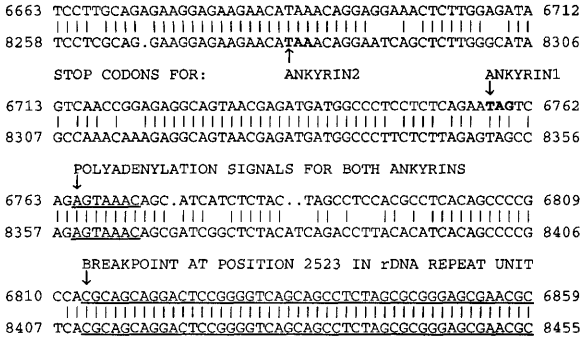

Figure 3 DNA sequence at the breakpoints between ankyrin genes and rDNA repeat units. The top line is the partial sequence of *ANK1* at the 3′ terminus of the gene followed by a spacer that is then fused into a region near the center of the large subunit rDNA gene at position 2523 (Healey et al. 1990). The second line is the comparable sequence for the *ANK2* 3′ region. *ANK1* is located between the *CRP65* gene and the rDNA repeat unit block fused distally with the telomere, and *ANK2* is between *CRP136* and a similar telomeric structure. The stop codons for both ankyrins are bold and marked. The polyadenylation signals are underlined and precede the underlined rDNA large subunit sequences that have a common fusion point.

Sequence analysis of the *CRP*, *gPK*, and *ANK* gene blocks revealed no sites for *Xba*I in either block. Earlier sequence analysis of the rDNA repeat unit (Healey et al. 1990) also showed the lack of *Xba*I sites. Cleavage of *Giardia* genomic DNA with *Xba*I, followed by CHEF gel separation and hybridization with probes representing segments of each of the above genes (Fig. 4) demonstrated that they were retained on the same *Xba*I segments in each case (Fig. 1). Two *Xba*I segments, estimated to be 14 kb and 35 kb, encoded the *CRP65* gene block, and a single *Xba*I segment, estimated to be 20 kb, retained the *CRP136* gene block, whereas a second, longer *Xba*I segment encoded other copies of *CRP136*, but not *ANK2* or *gPK2*. Hybridization of these same Southern transfers with a rDNA probe showed that each *Xba*I segment encoding a *CRP–gPK–ANK* gene block retained the flanking rDNA sequences as expected (Fig. 4). Furthermore, hybridization with a 30-nucleotide oligomer containing six copies of the telomeric repeat sequence TAGGG also demonstrated that this motif was present on each *Xba*I segment encoding the *CRP*, *gPK*, *ANK*, and rDNA genes. Because the DNA downstream from the *CRP* genes did not hybridize with the $(TAGGG)_6$ probe (data not shown) and it was not present internally to the gene block, it must have been present on the *Xba*I gene segment at a position distal to the rDNA repeat unit.

The 14-kb *Xba*I segment, containing the *CRP65* gene, can only encode less than two copies of the 5566-bp rDNA unit because of the limited coding potential remaining after the *CRP65*, *pGK*, and *ANK* genes. This suggests that there is only one complete copy of the entire rRNA gene unit at this telomere. Using a similar argument, there can be only a single complete unit in the 20-kb *Xba*I segment encoding the *CRP136* gene, although there could be as many as four encoded on the 35-kb *Xba*I segment.

We have called the gene block encoding *CRP65*, *pGK1*, *ANK1*, and their flanking spacers TGU1, for telomeric gene unit 1, and the corresponding block for *CRP136*, TGU2. There is some cross-hybridization between the *ANK* homologs and among the different kinase family members. The slight smearing in the upper row labeled TGU1 in Figure 4 is a result of the changes in copy number of the telomeric repeat $(TAGGG)_n$ because changes in copy number of the rDNA units would be in discrete blocks of 5.6 kb and this is not observed. The increased smear seen in the lane hybridized with the rDNA probe is attributable to changes in array of

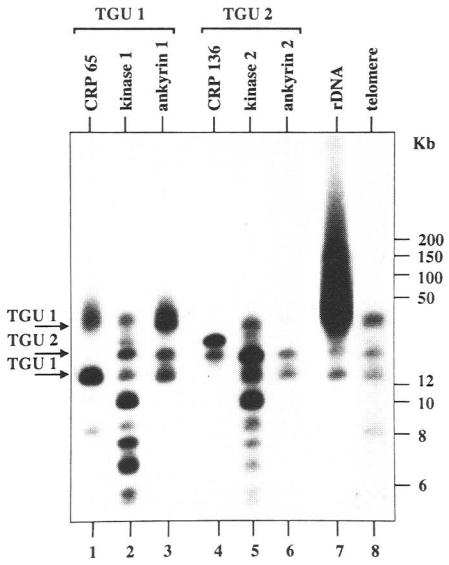

Figure 4 Hybridization of *Xba*I-cleaved genomic *Giardia* DNA with probes for the CRP and flanking genes, and the telomere. Intact chromosomal DNA prepared in agarose blocks was cleaved in situ with *Xba*I, which does not cleave the *CRP–gPK–ANK* TGUs or the rDNA unit, and separated by CHEF gel analysis. After Southern transfer, individual lanes were hybridized with radiolabeled probes as shown across the *top* of the figure. Fig. 1 shows the portion of each gene that was used as the probe; the rDNA probe was the complete rDNA unit (Healey et al. 1990) and the probe for the telomere was the end-labeled oligonucleotide $(TAGGG)_6$. TGU1 and TGU2 represent the units containing *CRP65–gPK1–ANK1* and *CRP136–gPK2–ANK2*, respectively.

rDNA units in low copy number variant chromosomes in the parasite population, because these are not labeled with the telomere probe, nor specific gene probes, but are detectable with the very strong, GC-rich rDNA probe. These may contribute to recombination events at the telomeres.

Chromosomal Localization of the *CRP* Gene Blocks

There are only two major *Xba*I segments encoding the *CRP136* gene. These two segments have also been located to only two *Not*I segments of 40 kb and 150 kb (data not shown), confirming the limited subset of *CRP136* locations. Because *CRP136* is present on every chromosome (five major bands on a pulsed field gel, numbered 3/4, 5, 6, 7, and 8 from the smallest) in this WB1B strain (Chen et al. 1995), at least some of the telomeres appear to be common.

We have also shown that *CRP65* is a member of a family of genes, with each member having EGF-like repeats differing by a single unit encoding 76 amino acids (Chen et al. 1996a). The *CRP65* genes are found located on two major *Xba*I segments on single-dimension gels (Fig. 4). On two dimensional pulsed-field gels, after cleavage with *Xba*I prior to electrophoresis in the second dimension, *CRP65* is found on seven *Xba*I segments (data not shown). Three of these hybridized with the telomeric probe; two of these are on the chromosome 3/4 cluster; and one is located on chromosome 5. Two more of the *Xba*I segments localized to the chromosome 3/4 cluster, but are not telomeric. There is also a non-telomeric copy of *CRP65* on chromosomes 7 and 8. On the same gels, the telomeric probe hybridized to a total of 10 bands, accounting for both telomeres of the major chromosomes. Two of these *Xba*I bands hybridizing to the telomere probe were identical to a second pair of bands in size but were located on different chromosomes (3/4 and 5), suggesting a common source; only one of each of these hybridized to *CRP65*. The three copies of TGU1 are therefore located at the telomeres of chromosomes 3/4 and 5.

Transcriptional Switching

Upon selection of *Giardia* WB1B for resistance to metronidazole, a large band (~6 kb) of new mRNA was found induced, sufficient to be readily detected on agarose gels after staining with ethidium bromide (Chen et al. 1995, 1996a). Reverse-transcribed mRNA from this band was used to clone *CRP65* and *CRP136*. *CRP136* was not expressed in the parent line, but induced in the resistant line, WB1B-M3. On the other hand, *CRP65* was expressed at a low level in the parent line as a 2-kb mRNA, sufficient in length to encode the gene, whereas in the drug-resistant line, WB1B-M3, *CRP65* was expressed as two longer transcripts of 5.5 and 7 kb (Chen et al. 1996a). These two transcripts also hybridized to both the *ANK1* and *gPK1* probes (data not shown), indicating a change to a stronger promoter upstream, reading through the *ANK1* and *gPK1* genes from the reverse direction (see Fig. 1). The lengths of these mRNAs suggest that they were initiated from promoters 3′ to *ANK1* and encoded in the spacer and/or the rDNA repeat unit.

Genetic Rearrangements Involving the CRP Genes

After selection for metronidazole resistance in WB1B, *CRP136* is also found located on a new 1.1-Mb chromosome. Chromosome 6 also appeared to have lost its copy of *CRP136* (Chen et al. 1995). A probe for the repeat unit in *CRP65* hybridized only

to the chromosome 3/4 cluster and chromosome 5 in the parent line, WB1B (Chen et al. 1996a). After drug selection, hybridization was also detected to the 1.1-Mb chromosome in the resistant line, WB1B-M3 (Chen et al. 1996a). Chromosome mapping probes (Chen et al. 1994, 1995; Upcroft et al. 1996) showed that the 1.1-Mb chromosome was derived from chromosomes 3/4, by partial duplication (Upcroft et al. 1993; Chen et al. 1994, 1995, 1996a). This minor variant was selected during the drug-resistance pressure. Aneuploidy in *Giardia* appears to be a common event under environmental stress, but selection for discrete, stable, minor chromosomes is less frequent (Upcroft et al. 1996).

We examined a number of *Giardia* isolates (Upcroft et al. 1995) for the presence of the *CRP136* and *CRP65* gene family by hybridizing with the appropriate isolated segments containing the sarafotoxin-like repeats from *CRP136*, or the EGF-like repeats from *CRP65*. In summary, isolates BRIS/83/HEPU/106, OAS1, and BAC2 were similar to WB1B and have copies of both *CRP65* and *CRP136*; BRIS/91/HEPU/1279 and BRIS/91/HEPU/1229 have only a *CRP136*-like gene present in the genome, and no *CRP65*; BRIS/93/HEPU/1709 retains only a very low copy number of each; and BRIS/92/HEPU/1487 retains no gene with sufficient homology to hybridize to any of the probes (data not shown). The CRP gene repertoire varies considerably among isolates both in copy number, location, and sequence.

Sequence analysis of the two CRP gene family members showed that they have internal repeated regions, but maintain conserved 5′ and 3′ flanking regions. The two termini have drifted in sequence over evolutionary time as have the spacer regions that flank the complete *CRP* genes; the rates of drift appear similar. However, the internal repeat regions have no detectable homology, *CRP136* having 23.5 copies of a 120-bp sarafotoxin gene homolog and *CRP65* encoding four copies of a 228-bp EGF-related domain. These are unlikely to have evolved by a simple genetic drift mechanism at a rate faster than the flanking gene and spacer sequences. Furthermore, the two repeat regions appear to have been replaced recently by a cassette-like mechanism that left a region encoding six amino acids at the amino terminus of the repeat in *CRP65* at the start of the unrelated repeat region in *CRP136* (Chen et al. 1996).

Rearrangements and Recombination within the rDNA Units

From the conservation of the TGUs, it is apparent that the *CRP–gPK–ANK* gene block is also conserved at the telomeric locations, although there is less conservation in the sequence of the spacers flanking the *ANK* gene prior to the last 14 bp at the site of fusion with the LSU rRNA gene. Although no other members of this CRP gene family have been analyzed as yet, other CRP genes, so far described, have little homology with those described here (Chen et al. 1995, 1996a) and may insert at different sites within the rDNA unit or remain independent. Large changes in copy number and chromosomal location of the rDNA arrays have been reported (Upcroft et al. 1990, 1996; Adam 1992; Chen et al. 1994; Le Blancq 1994) and the two different sequences that have been described for the location of the telomeric repeat $(TAGGG)_n$ demonstrate that the site of telomere addition can vary within the rDNA unit (Adam et al. 1991).

Our data also show that some telomeres do not contain the rDNA unit subtelomerically and therefore have formed at different subtelomeric locations (Fig. 4). We have no evidence for internal sites of telomeric sequences as found in chromosome III of yeast (Oliver et al. 1992). However, cleavage of intact *Giardia* chromosomes with the intron-specific endonuclease, I–*Ppo*I (Muscarella et al. 1990), which also cleaves a 15-bp recognition site in the *Saccharomyces cerevisiae* rDNA sequence (Lowery et al. 1992) and a number of other rDNA sequences, including *Giardia* (P. Upcroft, unpubl.), removes all rDNA arrays. There are no internal I–*Ppo*I cleavage sites in *Giardia* chromosomes, except for chromosome 6 for which the internal sites are not located in rDNA units (P. Upcroft, unpubl.). All rDNA sequences in the isolates that we have examined so far are therefore located at telomeres.

DISCUSSION

Chromosome walking and sequencing from two members of a new family of genes encoding cysteine-rich proteins, which have the typical transmembrane domains of *Giardia* cysteine-rich surface proteins, have identified their localization to the telomeres of *Giardia* chromosomes. Both these CRP genes were overexpressed after induction of resistance to the antigiardial drug metronidazole (Chen et al. 1995, 1996a). *CRP136* has 57% homology with a class of snake toxins, the sarafotoxins, from the burrowing adder *Atractaspis engaddensis* (Chen et al. 1995). Our motivating interest in following the expression and transcriptional activation of this gene is that the symptoms of giardiasis are the same as those following injection of the sarafotoxins; nausea, vomiting, stomach cramps, and diarrhea. *CRP136* is the first

cloned gene from *Giardia* with the potential of causing the major symptoms of the disease.

The CRP genes described here are parts of conserved gene blocks (TGUs) encoding novel protein kinases and ankyrin homologs. This unit organization suggests a functional coupling of the membrane–spanning CRP with the membrane-cytoskeleton anchor protein ankyrin (Dolata et al. 1993), perhaps via a two-component sensory system using a regulatory kinase as described for *Salmonella* (Véscovi et al. 1996). Ankyrin repeats are also found in a number of eukaryotic regulatory proteins, and perhaps of significance to the clustering of *ANK2* and *CRP136*, which is related to the sarafotoxins, is that the highest number of ankyrin repeats is found in the neurotoxin from the black widow spider venom, latrotoxin (Dolata et al. 1993). The sequences of the genes in the TGUs, including the spacers separating them, are over 80% homologous, except for the interchangeable cassette repeat units in the CRP genes. Although the two gene blocks have drifted in sequence, their organization has been conserved and probably reflects functional and regulatory requirements for the retention of this organization. The next gene downstream from *CRP136* is a tenascin homolog (P. Upcroft, N. Chen, and J.A. Upcroft, unpubl.), which is not present downstream from *CRP65*; a clear breakpoint occurs in the spacer separating *CRP136* and tenascin, where homology with the same downstream spacer region in *CRP65* is lost (Fig. 2). However, at the other end of the TGUs, within the spacer region following the *ANK* genes, homology has dropped to 60%, suggesting that this region does not encode information as essential as the remainder of the TGUs. Because these regions still have 60% homology they do not appear to have been exchanged by recombination at random with other regions of the *Giardia* genome, but may reflect the overall genetic drift in nonconserved sequences.

The spacers flanking the *ANK* genes terminate in a highly conserved 14-bp motif before fusion with the LSU rRNA gene at position 2523 of the rDNA unit sequence (Healey et al. 1990). Considering the drift in sequence between the two TGUs and particularly the drop to 60% homology in the *ANK* spacer, it is noteworthy that the breakpoints in the two genes are identical and that there is no drift in the LSU sequence for the next 47 bp. The high levels of conservation surrounding the insertion sites, compared with more distal sites, suggest that there is a common and evolutionarily frequent mechanism of exchanging these regions among rDNA sequences at the telomeres of *Giardia* chromosomes, even though each TGU can drift in overall sequence.

Consistent with recombinational events in the exchange of TGUs at telomeres is the variability seen between rDNA unit copy number and location among different isolates and on partially duplicated, aneuploid chromosomes (Upcroft et al. 1990, 1996; Adam 1992; Chen et al. 1994; Le Blancq 1994) which indicates that the rDNA complement in the arrays is mobile. Because our data show that all the rDNA units are telomeric in the strains so far analyzed, significant recombination is occurring at immediately subtelomeric regions that are juxtaposed to the TGUs. Furthermore, the TGUs analyzed here are flanked by only one or a few copies of the rDNA unit, suggesting a minimal size recently acquired by such recombination events. Recombination mechanisms must also be involved in the exchange of repeat unit cassettes in the generation of *CRP* family members, described above. We have also presented evidence that the spacer between the LSU and small subunit rDNA genes has been the recipient of genes for amplification purposes, because an ORF with appropriate promoter and signal sequences was found encoded on the opposite strand to the LSU rDNA in the first rDNA repeat unit that was sequenced (Upcroft et al. 1990). The spacer in a recently sequenced, clinical *G. duodenalis* isolate that appears to be a new species is devoid of any ORF in the rDNA spacer region (Upcroft et al. 1994).

Although all rDNA unit arrays appear to be telomeric, all telomeres detected by the repeat sequence $(TAGGG)_n$ are not flanked by rDNA arrays. Nor are all *CRP* family members located at telomeres, with some TGUs at telomeres being occupied by *CRP136* and perhaps by other CRP genes. The kinase gene family found in the TGUs has more members than can be accounted for by telomeric locations (Fig. 2), and there are many *ANK* homologs in the *Giardia* genome (P. Upcroft, N. Chen, and J.A. Upcroft, unpubl.). However, the ankyrins described here coupled to the CRP genes appear to be a limited subset localized to the TGUs (Fig. 4).

Although no cysteine-rich surface proteins in *Giardia* have been detected previously at telomeric sites (Yang and Adam 1995), the requirements to generate such a large overexpression of mRNA for *CRP136* and *CRP65* when induced by environmental stress such as metronidazole exposure appear to invoke alternative expression systems at *Giardia* telomeres, and rearrangement mechanisms comparable, but differing in detail, to the trypanosomes (Vanhamme and Pays 1995). Furthermore, at least two different *CRPs* are induced at the same time.

CRP136 was induced de novo during selection for drug resistance in *G. duodenalis* WB1B, whereas *CRP65* had low-level transcription in the parent line. Transcription switched from a gene-length mRNA of 2 kb to two very highly expressed mRNAs of 5.5 and 7 kb (Chen et al. 1996a). The new promoters have overidden the promoter just 5′ of the CRP genes, which normally transcribes an mRNA of the same length as the gene. Because the 5.5- and 7-kb mRNAs also hybridized to the *ANK* and *gPK* genes, and the 3′ region encoded *CRP65*, these mRNAs were initiated from promoters distal to the TGU, either in the spacer or from rDNA promoters. These promoters can be provided only in the TGUs in the case of *CRP65* and *CRP136* because these units are the only ones detectable with the clustering of *CRP*, *gPK*, and *ANK* seen in the large mRNAs. Because the TGUs are telomeric, the mRNA induction occurred at telomeric expression sites.

Few *Giardia* promoters have been defined, although repeated motifs have been described in the *Giardia* rDNA spacer (Upcroft et al. 1994) analogous to the elements found in the same region of the *Trypanosoma cruzi* rDNA spacer and other eukaryotes, with promoter and enhancer functions (Dietrich et al. 1993). This is consistent with the transcription of an ORF encoded in the spacer by the opposite strand to the LSU rRNA gene, where promotion is bidirectional (Healey et al. 1990). A promoter region was defined for this ORF that had the characteristics of a fused *pol*II–*pol*III promoter (Upcroft et al. 1990). The common site of insertion of the TGUs in the truncated LSU rRNA gene also suggests an essential site for transcriptional control that can override rRNA transcription. As in the multifactorial regulation of transcriptional switching, silencing, and gene regulation at the telomeres of *S. cerevisiae* (Blackburn and Greider 1995; Kipling 1995) and *T. brucei* (Vanhamme and Pays 1995), *G. duodenalis*, one of the earliest eukaryotes, had already exploited the unusual environment of the telomere for regulating high levels of variable gene expression.

METHODS

Materials

Restriction endonucleases were from New England Biolabs (Beverly, MA). I–*Ppo*I was from Promega (Madison, WI).

G. duodenalis Isolates

G. duodenalis BRIS/83/HEPU/106 is our standard laboratory strain (Upcroft et al. 1993, 1995). For the origin of OAS1, WB1B, and BAC2, see Capon et al. (1989). *G. duodenalis* strains BRIS/91/HEPU/1229, BRIS/91/HEPU/1279, BRIS/92/HEPU/1487, and BRIS/93/HEPU/1709 have been described recently (Upcroft et al. 1995). The generation of the metronidazole-resistant line WB1B-M3 has been described (Townson et al. 1992; Chen et al. 1995). All *G. duodenalis* isolates were grown in modified TYI-S-33 medium containing 1 mg/ml of bile (Upcroft et al. 1995).

Construction and Screening of Size-Specific Genomic DNA Libraries

G. duodenalis trophozoites from cultures expanded in Nunc single-layer Cell Factories (Nunclon, Denmark) were lysed in pronase–SDS, and DNA was extracted with phenol as described previously (Upcroft et al. 1987). DNA samples were cleaved with selected restriction endonucleases (New England Biolabs) according to the manufacturer's instructions and electrophoretically separated in 0.9% SeaPlaque agarose (Marine Colloids, Rockville, MN). The areas of the gel corresponding to the size of the required DNA segments were excised and the DNA purified (Upcroft and Healey 1987). The DNA segments were then cloned into Bluescript (Stratagene, La Jolla, CA) using T4 DNA ligase (Upcroft et al. 1993). Transformation was carried out according to the procedure of Hanahan (1983) (Upcroft and Healey 1987) using *Escherichia coli* DH5α (Gibco-BRL, Gaithersburg, MD) as a host. The size-specific genomic DNA library was screened by colony hybridization (Hanahan and Meselson 1980). Plasmids from the positive colonies that contained the required DNA segments were purified with Qiagen-tip 20 columns according to the manufacturer's instructions (Qiagen, Hilden, Germany) and confirmed by Southern transfer and hybridization (Southern 1975).

DNA Sequencing and Comparison

DNA sequencing was carried out using the Taq dye primer cycle sequencing kit and automated sequencer from Applied Biosystems (Foster City, CA), which is based on the chain termination sequencing procedure (Sanger et al. 1977). Subclones were constructed using DNase in the presence of Mn^{2+} (Lin et al. 1985) or restriction endonucleases followed by ligation and transformation as described above.

DNA sequences were analyzed and compared with the FASTA (Lipman and Pearson 1985) and BESTFIT (Smith and Waterman 1981) programs implemented in the Genetics Computer Group suite of programs (Devereux et al. 1984) on a Sun Sparcstation 10 computer.

Southern Transfer and Hybridization

*Xba*I-cleaved chromosomal DNA in agarose blocks was separated by contour-clamped homogeneous electric field (CHEF) gel electrophoresis using a Bio-Rad (Hercules, CA) CHEF-DR II gel electrophoresis apparatus as described (Upcroft et al. 1993, 1996). Restriction endonuclease-cleaved genomic DNA was separated by agarose gel electrophoresis (Upcroft and Upcroft 1994; Upcroft et al. 1995). DNA was transferred (Southern 1975) to nylon membrane (Amersham Hybond-N, Buckinghamshire, UK) and hybridized with ^{32}P-labeled probes. Radioactive genomic DNA probes were prepared from recombinant

plasmid DNA or purified inserts with [α^{32}P]dCTP (DuPont, Sydney) by random priming (Feinberg and Vogelstein 1984). The telomeric probe, $(TAGGG)_6$, was end-labeled as described (Adam et al. 1991). The filters were stripped between hybridization with 0.2 M NaOH for 2 hr at 37°C.

ACKNOWLEDGMENTS

We thank Ray Campbell for technical assistance, and the National Health and Medical Research Council of Australia and the Australian Research Council for support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adam, R.D. 1991. The biology of *Giardia* spp. *Microbiol. Rev.* **55:** 706–732.

———. 1992. Chromosome-size variation in *Giardia lamblia*: The role of rDNA repeats. *Nucleic Acids Res.* **20:** 3057–3061.

Adam, R.D., T.E. Nash, and T.E. Wellems. 1991. Telomeric location of *Giardia* rDNA genes. *Mol. Cell. Biol.* **11:** 3326–3330.

Blackburn, E.H. and C.W. Greider. 1995. *Telomeres*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Brown, D.M., J.A. Upcroft, and P. Upcroft. 1995. Free radical detoxification in *Giardia duodenalis*. *Mol. Biochem. Parasitol.* **72:** 47–56.

———. 1996. A H_2O-producing NADH oxidase from the protozoan parasite *Giardia duodenalis*. *Eur. J. Biochem.* **241:** 155–161.

Capon, A.G., J.A. Upcroft, P.F.L. Boreham, L.E. Cottis, and P.G. Bundesen. 1989. Similarities of *Giardia* antigens derived from human and animal sources. *Int. J. Parasitol.* **19:** 91–98.

Cavalier-Smith, T. 1993. Kingdom protozoa and its 18 phyla. *Microbiol. Rev.* **57:** 953–994.

Chen, N., J.A. Upcroft, and P. Upcroft. 1994. Physical map of a 2Mb chromosome of the intestinal parasite *Giardia duodenalis*. *Chromosome Res.* **2:** 307–313.

———. 1995. A *Giardia duodenalis* gene encoding a protein with multiple repeats of a toxin homologue. *Parasitology* **111:** 423–431.

———. 1996a. A new cysteine-rich protein-encoding gene family in *Giardia duodenalis*. *Gene* **169:** 33–38.

———. 1996b. A novel protein kinase gene family in *Giardia duodenalis*. *Gene* **177:** 191–194.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387–395.

Dietrich, P., M.B. Soares, M.H.T. Affonso, and L.M. Floeter-Winter. 1993. The *Trypanosoma cruzi* ribosomal RNA-encoding gene: Analysis of promoter and upstream intergenic spacer sequences. *Gene* **125:** 103–107.

Dolata, M.M., J.J. Van Beeumen, R.P. Ambler, T.E. Meyer, and M.A. Cusanovich. 1993. Nucleotide sequence of the heme subunit of flavocytochrome *c* from the purple phototropic bacterium, *Chromatium vinosum*. *J. Biol. Chem.* **268:** 14426–14431.

Feinberg, A.P. and B. Vogelstein. 1984. Addendum. A technique for radio-labelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137:** 266–267.

Hanahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166:** 557–580.

Hanahan, D. and M. Meselson. 1980. Plasmid screening at high colony density. *Gene* **10:** 63–67.

Healey, A., R. Mitchell, J.A. Upcroft, P.F.L. Boreham, and P. Upcroft. 1990. Complete nucleotide sequence of the ribosomal RNA repeat unit of *Giardia intestinalis*. *Nucleic Acids Res.* **18:** 4006.

Kipling, D. 1995. *The telomere*. Oxford University Press, New York, NY.

Le Blancq, S.M. 1994. Chromosome rearrangements in *Giardia lamblia*. *Parasitol. Today* **10:** 177–179.

Lin, H., S. Lei, and G. Wilcox. 1985. An improved DNA sequencing strategy. *Anal. Biochem.* **147:** 114–119.

Lipman, D.J. and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* **227:** 1435–1441.

Lowery, R., L. Hung, K. Knoche, and R. Bandziulis. 1992. Properties of I-*Ppo*I: A rare-cutting intron-encoded endonuclease. *Promega Notes* **38:** 8–12.

Mertens, E. 1993. ATP versus pyrophosphate: Glycolysis revisited in parasitic protozoa. *Parasitol. Today* **9:** 122–126.

Mowatt, M.R., A. Aggarwal, and T.E. Nash. 1991. C-terminal sequence conservation among variant-specific surface proteins of *Giardia lamblia*. *Mol. Biochem. Parasitol.* **49:** 215–228.

Muscarella, D.E., E.L. Ellison, B.M. Ruoff, and V.M. Vogt. 1990. Characterization of I-*Ppo*, an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of *Physarum polycephalum*. *Mol. Cell. Biol.* **10:** 3386–3396.

Nash, T.E. 1992. Surface antigen variability and variation in *Giardia lamblia*. *Parasitol. Today* **8:** 229–234.

Oliver, S.G., Q.J. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P. Ballesta, P. Benit, et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38–46.

Sanger, F., S. Nicklen, and A.R. Coulsen. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **90:** 5463–5467.

Smith, T.F. and M.S. Waterman. 1981. Comparison of biosequences. *Adv. Appl. Math.* **2:** 482–489.

Sogin, M.L., J.H. Gunderson, H.J. Elwood, R.A. Alonso, and D.A. Peattie. 1989. Phylogenetic meaning of the kingdom concept: An unusual ribosomal RNA from *Giardia lamblia*. *Science* **243:** 75–77.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98:** 503–517.

Townson, S.M., H. Laqua, P. Upcroft, P.F.L. Boreham, and J. Upcroft. 1992. Induction of metronidazole and furazolidone resistance in *Giardia*. *Trans. R. Soc. Trop. Med. Hyg.* **86:** 521–522.

Townson, S.M., G.R. Hanson, J.A. Upcroft, and P. Upcroft. 1994. A purified ferredoxin from *Giardia duodenalis*. *Eur. J. Biochem.* **220:** 439–446.

Townson, S.M., J.A. Upcroft, and P. Upcroft. 1996. Characterisation and purification of pyruvate:ferredoxin oxidoreductase from *Giardia*. *Mol. Biochem. Parasitol.* **79:** 183–193.

Upcroft, J.A. and P. Upcroft. 1994. Two distinct varieties of *Giardia* in a mixed infection from a single human patient. *J. Euk. Microbiol.* **41:** 189–194.

Upcroft, J.A., A.G. Capon, A. Dharmkrong-At, A. Healey, P.F.L. Boreham, and P. Upcroft. 1987. *Giardia intestinalis* antigens expressed in *Escherichia coli*. *Mol. Biochem. Parasitol.* **26:** 267–276.

Upcroft, J.A., P.F.L. Boreham, and P. Upcroft. 1989. Geographic variation in *Giardia* karyotypes. *Int. J. Parasitol.* **19:** 519–527.

Upcroft, J.A., A. Healey, R. Mitchell, P.F.L. Boreham, and P. Upcroft. 1990. Antigen expression from the ribosomal DNA repeat unit of *Giardia intestinalis*. *Nucleic Acids Res.* **18:** 7077–7081.

Upcroft, J.A., A. Healey, and P. Upcroft. 1993. Chromosomal duplication in *Giardia duodenalis*. *Int. J. Parasitol.* **23:** 609–615.

Upcroft, J.A., A. Healey, R.W. Mitchell, and P. Upcroft. 1994. A new rDNA repeat unit in human *Giardia*. *J. Euk. Microbiol.* **41:** 639–642.

Upcroft, J.A., P.F.L. Boreham, R.W. Campbell, R.W. Shepherd, and P. Upcroft. 1995. Biological and genetic analysis of a longitudinal collection of *Giardia* samples derived from humans. *Acta Tropica* **60:** 35–46.

Upcroft, J.A., N. Chen, and P. Upcroft. 1996. Mapping variation in chromosome homologues of different *Giardia* strains. *Mol. Biochem. Parasitol.* **76:** 135–143.

Upcroft, P. and A. Healey. 1987. Rapid and efficient method for cloning of blunt-ended DNA fragments. *Gene* **51:** 69–75.

Vanhamme, L. and E. Pays. 1995. Control of gene expression in trypanosomes. *Microbiol. Rev.* **59:** 223–240.

Véscovi, E.G., F.C. Soncini, and E.A. Groisman. 1996. Mg^{2+} as an extracellular signal: Environmental regulation of Salmonella virulence. *Cell* **84:** 165–174.

Warren, K.S. 1989. Selective primary health care and parasitic diseases. In *New strategies in parasitology* (ed. K.P.W.J. McAdam), pp. 217–231. Churchill Livingstone, Edinburgh, UK.

Yang, Y.-M. and R.D. Adam. 1995. Analysis of a repeat-containing family of *Giardia lamblia* variant-specific surface protein genes: Diversity through gene duplication and divergence. *J. Euk. Microbiol.* **42:** 439–444.

Received September 17, 1996; accepted in revised form November 12, 1996.

Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote Giardia duodenalis.

P Upcroft, N Chen and J A Upcroft

Genome Res. 1997 7: 37-46
Access the most recent version at doi:10.1101/gr.7.1.37