# Comparison of the intergenic spacers and 3′ end regions of the large subunit (28S) ribosomal RNA gene from three species of *Schistosoma*

R. A. KANE* *and* D. ROLLINSON

*Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD*

(*Received 10 February 1998; revised 9 April 1998; accepted 10 April 1998*)

SUMMARY

The sequences for the entire intergenic spacer (IGS) and 3′ end region of the large subunit (28S) ribosomal RNA gene are presented for *Schistosoma intercalatum*, *S. haematobium* and *S. mansoni*. The IGS was found to have at least 2 size formats for *S. intercalatum* and *S. haematobium* and the region for all 3 species contains numerous repeats and evidence of recombination. An A+T rich sequence found in *S. intercalatum* and *S. haematobium* is described and its possible function and origins are discussed. The amplification of this region by means of the polymerase chain reaction can be used to discriminate clearly between the 3 species involved in this study. The putative end of the 28S gene is identified.

Key words: intergenic spacer, rRNA, A+T rich sequence, repeats, *Schistosoma intercalatum*, *Schistosoma haematobium*, *Schistosoma mansoni*.

## INTRODUCTION

The work of Walker, Rollinson & Simpson, (1986) on the ribosomal RNA (rRNA) gene complex of *Schistosoma* revealed that the internal transcribed spacer (ITS) and intergenic spacer (IGS) regions displayed variation between species. This was demonstrated by the production of restriction fragment length polymorphisms (RFLPs) through the hybridization of restricted genomic DNA to ribosomal probes. Since then size variation of the ITS has been investigated (Kane & Rollinson, 1994; Kane *et al*. 1996) and found to be due to different numbers of tandem repeat elements within ITS1. Size variation of the IGS has been a key element in the molecular discrimination between the human pathogens *Schistosoma haematobium* and *S. intercalatum* and other species within the terminal spined egg group. RFLP analysis revealed a 0·5 kb deletion in this region for *S. haematobium* (Walker *et al*. 1986; Rollinson *et al*. 1990) which, although used as a discriminatory tool, has never been fully investigated. This paper presents the first complete sequences of the ribosomal intergenic region for *S. haematobium*, *S. intercalatum* and *S. mansoni* and identifies repeats, areas of recombination, possible origins of DNA replication, and the putative end-point of the 28S ribosomal gene.

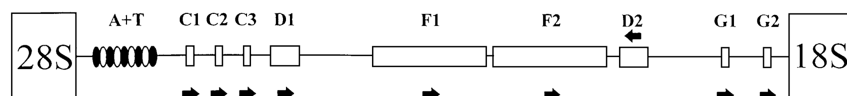
* Corresponding author. Tel: +0171 938 9297. Fax: +0171 938 8754. E-mail: rak@nhm.ac.uk

## MATERIALS AND METHODS

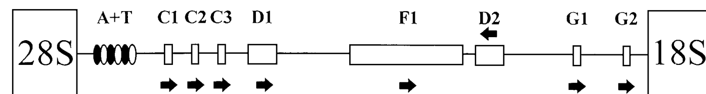
An *S. haematobium* isolate from Mauritius (NHM 2667) and an *S. intercalatum* isolate from São Tomé (NHM 2756) were selected for study. DNA was extracted from adult worms according to the method outlined by Walker *et al*. (1986) with minor modification. The IGS was amplified by PCR using a Hybaid Thermal Cycler and the thermostable polymerase AmpliTaq ‘Gold’ (Applied Biosystems). Cycling conditions were 94 °C for 12 min/58 °C for 2 min/72 °C for 3 min × 1, 94 °C for 1 min/58 °C for 2 min/72 °C for 3 min × 30 and 94 °C for 1 min/58 °C for 2 min/72 °C for 10 min × 1. The primers employed were ET10 (5′ CTGAGACAAGCATATGACTAC 3′) which sits on the 5′ end of the 18S gene and the ‘universal’ primer 28aa (5′ AGGTTAGTTTTACCCTACT 3′), located on the 3′ end of the 28S gene (Hillis & Dixon, 1991). Amplified fragments were cloned into pGEM-T, using a commercial cloning kit (Promega). Transformants were screened by PCR using the primers ET10 and 28aa. Those *Escherichia coli* containing fragments of interest were grown up and the plasmids harvested and purified by a modified mini alkaline-lysis/polyethylene glycol precipitation procedure (Applied Biosystems protocol part no. 401461). The sequence for the *S. mansoni* IGS was generated from the ribosomal DNA clones pSM890 and pSM389 (both in the vector pBR322), which are known to contain the entire region (Simpson *et al*. 1984). These were grown up in *E. coli* and purified in the same manner as mentioned previously. The IGS for *S. intercalatum* and *S. haematobium* was se-

 © 1998 Cambridge University Press
https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

*S. intercalatum*




*S. haematobium*




*S. mansoni*

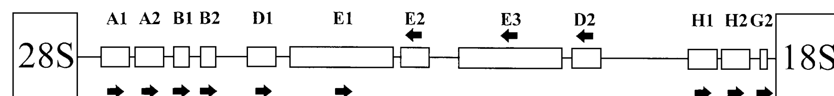



Fig. 1. Schematic diagram of the IGS for the 3 species of schistosome (not to scale). Repeats are labelled A to H and their alphabetical designation indicates the order in which they are encountered in Fig. 2. The arrows define the orientation of a specific repeat sequence i.e. those pointing to the left indicate that the repeat is inverted relative to it's partner. The black and white regions marked as A+T represent the main A+T rich regions of the IGS.

quenced in the main using a pool of 3 cloned PCR products for each species (clones I1D, I2A and I3A for *S. intercalatum* and H1D, H1F and H2A for *S. haematobium*). However, in the A+T rich region where sequence variation occurred between cloned PCR products, individual clones were sequenced (clone I2A for *S. intercalatum* and clone H1D for *S. haematobium*). All sequencing was performed using Fluorescent Dye Terminator Sequencing Kits (Applied Biosystems) and the sequencing reactions run on either an Applied Biosystems 377 or a 373A automated sequencer.

Amplification of the A+T region and adjacent sequence was achieved using primers MNTS1 (5′ CCTATCCCTGCTTGAC 3′) and NTS15 (5′ ATATATCGCGGGCTT 3′). Cycling parameters were the same as those described previously. PCR was performed upon 4 isolates of *S. haematobium*, namely, Cameroon (NHM 1848), Mali (NHM 1804), Mauritius (NHM 2667) and Malawi (Cape Maclear), 4 isolates of *S. mansoni*, Senegal (NHM 2922), Egypt (NHM 2917), Brazil (NHM 3054), and Kenya (NHM 2834) and 2 isolates of *S. intercalatum*, São Tomé (NHM 1784 and NHM 2756).

## RESULTS

### *PCR products*

The IGS of all 3 species proved to have a dynamic and highly complex structure. The first indication of this became apparent upon amplification of the IGS PCR products. *S. haematobium* appeared to produce a main discrete band correlating to the IGS, however, *S. intercalatum* consistently produced at least 2 main PCR fragments and on occasions a further 2 smaller, fainter bands. Despite alteration of annealing temperatures and times and the use of alternate primers, no discrete single IGS band could be obtained for *S. intercalatum*. DNA was then extracted from individual worms in order to check that a particular size of IGS was not associated with specific individuals. Following PCR amplification of individual worm extracts for both *S. intercalatum* and also *S. haematobium*, pattern profiles proved to be identical to those from the DNA of pooled worms. As a consequence, the IGS PCR fragments of both species were cloned directly into pGEM-T plasmids, with a view to sorting out the different sizes of IGS product from their individual clones. Upon screening, it became apparent that a large and a small sized product had been cloned for both species. It is possible that IGSs may exist for both species with lengths different to those described here, however, such were not observed in the clones screened for this study.

### *Sequences*

The IGS region for *S. mansoni* was sequenced along with the small and large cloned spacers for both *S. intercalatum* and *S. haematobium*. Sequences of the largest cloned IGS fragments are presented schematically in Fig. 1 and in detail in Fig. 2, together with the comparable *S. mansoni* sequence. The largest cloned IGSs have EMBL accession numbers AJ223840 (*S. intercalatum*) and AJ223838 (*S. haematobium*), while those for the smaller IGS fragments are AJ223841 (*S. intercalatum*) and AJ223839 (*S. haematobium*). The accession number for the *S. mansoni* IGS is AJ223842. The small *S. haematobium* fragment (537 bp) was considerably

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

smaller than the full sized IGS presented here (2058 bp). It may represent some form of PCR artefact or non-functional relic as only 231 bp of sequence were encountered from the 3′ end of the IGS before the A+T rich sequence was reached i.e. 1521 bp of sequence was missing, in all likelihood containing the promoter and start of transcription.

The IGS sequences for both *S. mansoni* and *S. intercalatum* contained a pair of large non-identical repeats and all 3 species had smaller repetitive regions. Both *S. haematobium* and *S. intercalatum* exhibited long stretches of very A+T rich sequence towards the putative 5′ end of the IGS (see Figs 1 and 2) which were present in both small and large cloned fragments. No consensus sequence was apparent for this region which varied slightly in composition and length between cloned PCR fragments. The A+T region for the sequences displayed in Fig. 2 represents the individual clones which contained the longest A+T sequences. These have a length of 356 bp for *S. intercalatum* and 226 bp for *S. haematobium*. However, despite having no consensus sequence, the overall length of the A+T rich region for *S. haematobium* was always approximately two thirds of the length for that of *S. intercalatum*. *S. mansoni* lacked this long stretch of A+T rich sequence but in common with the other 2 species, exhibited a short 17 bp stretch of A+T rich sequence downstream of the longer A+T regions (see Fig. 2).

### *Inverted repeats*

*S. mansoni* contained 2 large inverted repeats, E1 (285 bp) and E3 (271 bp) with a partial inverted repeat E2 (67 bp) sandwiched between the two (see Figs 1 and 2). All 3 species contained a pair of short inverted repeats D1 (52 bp) and D2 (51 bp) (see Figs 1 and 2) situated respectively 248 bp downstream of the 3′ side of the A+T rich region and 745 bp upstream of the 3′ end of the IGS (*S. intercalatum*).

### *Direct repeats*

Inverted repeats were not the only form of repeat within the *Schistosoma* IGS. Two sets of direct repeats A1, A2 and B1, B2 were found in *S. mansoni* starting at 295 bp and 491 bp respectively from the putative 5′ end of the IGS. Moreover, 3 short direct repeats designated C1 (19 bp) to C3 were located to the 3′ side of the A+T rich region in *S. intercalatum* and also in *S. haematobium*. Two very large direct repeats F1 (379 bp) and F2 (378 bp) occurred in the larger of the 2 cloned *S. intercalatum* IGS sequences (see Figs 1 and 2). The presence of the repeat mainly, though not entirely, accounted for the difference in size between the large and small forms of IGS cloned for this species. Finally, 2 very small direct repeats G1 (12 bp) and G2 lay near the 3′ end of the IGS in *S. intercalatum* and *S. haematobium*. The repeats were also present in *S. mansoni* within the larger repeats designated H1 and H2 and also to the 3′ side of H2.

### *Amplification of the A+T region*

The A+T region and it's immediate adjacent sequence can be amplified from conserved priming sites (MNTS1 and NTS15) to produce a discrete fragment. Following agarose gel electrophoresis, the visible size variation of this fragment allowed easy discrimination of the 3 species, isolates of *S. haematobium* having the smallest fragment and those of *S. intercalatum* the largest. Although *S. mansoni* had no large A+T region, the size of the fragment generated between the 2 priming sites fitted neatly between that of the other 2 species (see Fig. 3). This technique worked equally well on DNA extracted from pooled worms or individual worms and may prove useful for the discrimination of species in areas where *S. haematobium* and *S. intercalatum* overlap.

### *Putative endpoint of the 28S gene*

The 3′ end of the 28S gene for *Schistosoma* has not been determined experimentally in this study. However, through comparison with other 28S sequences from the Ribosomal Database Project (Maidak *et al*. 1997) the putative 3′ end of the 28S gene for *Schistosoma* was identified, allowing the estimation of the length of the IGS sequences shown in Fig. 2, namely 2571 bp for *S. intercalatum*, 2058 bp for *S. haematobium* and 2558 bp for *S. mansoni*. In addition, following a 'blastn' search (Altschul *et al*. 1990) of the EMBL database, the 28S sequences presented here were found to match an *S. mansoni* (Egyptian strain) Expressed Sequence Tag (EST) complete with a short poly A tail (accession no. N20737). It is of interest to note that the junction between the poly A tail of this EST and the rest of the sequence correlates exactly with the end of the 28S gene as estimated by homology alignment.

Whether the end of the 28S gene is also the end of transcription has yet to be determined. Some studies have indicated that the end of the primary rRNA transcript is concomitant with the 3′ end of the 28S gene, while others have found that transcription continues several hundred base pairs beyond to form a 3′ trailer, e.g. *Xenopus laevis* (Sollner-Webb & Tower, 1986).

Whilst the localization of the start of transcription was not an objective of the present study, it should be noted that the rRNA 5′ ends of a number of plants, including *Arabidopsis thaliana* map to a sequence TATAT<u>A</u>GGG (the starting point of transcription being underlined) (Doelling, Gaudino & Pikaard, 1993). This sequence was present in the

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

```
S. intercalatum  TGAAGTACAATCTGGCCGTTGCTATGGTAATCCTGTTTAGTACGAGAGGAACCGCAGGTTCAGATAATTGGTATATGTGCCTGGTCGAAAGGCCAATGGTGCGAAGCTACCATC  114
S. haematobium   ..................................................................................................................
S. mansoni       ..................................................................................................................

int   TGAGGGATTAAGACTGAACGCCTCTAAGTCTGAATCCCGTCCAAATTTGCAACGGTACATTCTGTATCCGACACGTGGCTGGGCAAGCTTACGTATTCCTTTTGGTGTCTGGACTGGACCATGTG  239
haem  .............................................................................................................................
man   .............................................................................................................................

int   TTCGGCTAGGCCCATTAGGTCTGGTGTTATGCCAGAGCGATGACCATAAACCGTGCTCCGGATTTTGGGGTACGGGGGTCGACACGATATCTAATTGTGGTATGAGCGTGAGGATACTAAATCGT  364
haem  .............................................................................................................................
man   .............................................................................................................................

                                                                                                                          ↓
int   TTGTAGACGACTTAATGTTTTGTCGGGGAGACGTACGTGCTCGAGCAGCTTTCACACTGCGAGGCATTGAATCCTATCCCTGCTTGACGTAATGATTTGAGGACATGTTCCTCTATCATTTCCCT  489
haem  .............................................................................................................................
man   ..................................A..........................................................................................

int   TATTGTGCTGGACTTCGGTCTGGTTAAAGCATTTATACCTATGTGTATACTATGGGATGCATGGGTCTCGGC--TTGTGCATAA-CATATATCCATATATTGTATACTTGCGATGTTCTGCATAT  614
haem  .....C...A.................................................A........--.........-..............C...........C..........
man   ...G..A....G...........C.G.....C...........................AA....GT.GAC.....T..TCT...G.G.T...G...G..........A...A.......C.

int   GTGCATATATGTGCTATACATCTGTATCATATGTATATATGTATGTATATATATATATATATGTGCATATATATATATATGTATATATATATGTATATATATATGCATGTATATGCATATATATG  739
haem  ...............................AC....G.AC..A...........G.......AT....G...G.....A.......G.GCA.......G...ATG.A.....AT......GCA
man   A...GA.--.....ATATATGT..GTGT.ACGTATGTGTATATGTGT.G.C.---------.......TACAT.TGTATCA...GC......A.GA....CATATATGTGTATATGT....AG..

int   TGCATATATATGTATATATATATGTATATATATATATGCATGTATATGTATATATATGTGCATATATATGTATATGTATATATATATATGTATATATATATATGTATATATATATATGCATGTAT  864
haem  ..T....G...A.....G.GC..A.....G.....G.AT..A.....A...GC..G.A.ATG.......A.G.GCA.......G.....A.............A....-----------------
man   .ATG.TC..G.TA.C.C....GCACG.GCT...GCTATT.CTAG....C....CA.C.CATGG.CGCGCT...CAT..CC...TC.ATA.ACACGCGTGTGT.TAT.G.G.TGT.CATATATG..
                                 A1

int   ATGTATATATATGTGCATATCTATGTATATGTATATATATGT-GCATATCTATGTATATGTATATATATATATATATATGCATGTATATATATAAATATGTATATATATATACATGCATATATGT  989
haem  ---------------------------------------------------------------------------------------------------------------------.....G...A.
man   .GAGTCCGGGTATA.TCCGCAGTG.ATATGC...TAC..GTAT.......A.C.CT.GCACG..CT...CAT..CC...T.TCTA.A.A.ATATTG...TGTGCAGATC....G..AGACC.GG.
                              A2                                                                                  B1

                                                                          C1                       C2                      C3
int   ATATATATATGTGCATATGAACGTGCGTATGCACATATATCGCATTCTGCAAAATTAAGCTGATTCAGGTGTAGTCCGCGGCAAATCCAGGTGTAATCCGCGGCAAGTTCAGGTGTAGTCCGCGC  1114
haem  G..................T....T...........................................................A..G..........G..........A.C................
man   ...GTCCGCA..TA..CCAGGT.GTAA.CC...G..TATA..TGAGACCAGGTG.AGTC.GC.ACGGA-----------------.........G......CGT.AAATTTT.----------..
                                 B2

int   GATTTTCGCGCAAAAATCGGCCGGCCCGCGGCGCGGCTGGTCGCTAAAAACTTTCTAAGCCCGCGATATATATATATATATATCCCCCTTATTTTAGGGTGCACGGAAAAAAA-GCGCACCATT-  1239
haem  ..............................................................................................................C..A.......C.-
man   .-......GCTCGTT.AGAC----------------------.................G...................-C.......TCA.ATTTT..ACTAA..T.GCTA.TTTCTAGC.A

                                D1
int   ---------------GGGAGACCACGAAGTTAGGGGTAGTCTCCCTATATAACATGGCGAGTAAGACCATGTGTAAAGCGCGTCAAAAATTTGACTTAGCCGAAAAACACTGATTTCATAGTAAG  1364
haem  ---------------..............................................................................................................
man   GTTTAGTGCATTATA..........CG....TA...............AT....A..C.......G..AC.......A..AAACTTTGGAGTCTCG.TGTC..T.TACA.TT.A.GCATCATTT
                                                                                                                         E1

int   ACCAATTGTAATGCGCAATTGAAAAGG--------------------------------------------------------------------------------------------------  1489
haem  .....G...................T--------------------------------------------------------------------------------------------------
man   .AT.GC.C.T.ATTAG.TAA.T.CTACAAAGAAAACCCCATTGAAAAATATTAATCCGTTCAAAAGTTATAAAGAAAAGTGTGACCCAAAAAAATTCCTATAGAAAGGTGCGGACGGCAAAATCCT

int   -----------------------------------------------------------------------------------------------------------------------------  1614
haem  -----------------------------------------------------------------------------------------------------------------------------
man   ATAGAAAAGTCCGAGATAGCGAAATTTTACGAACACACTTTTCCGCATAACTTTTGAACGGCTCAACCGATTTTGTTCATTCTTTTTTTAATCGATTTATCTTGCGAATATCTATCTTTTGATGT

int   -----------------------------------------------------------------------------------------------------------------------------  1739
haem  -----------------------------------------------------------------------------------------------------------------------------
man   ATAACATATAACCGTGTATCATCTAAAAAAATAATTGCGCACTACACCTGGTTCACTATGGTATTTGGCCGTCCGCACCTTTCTATAGGAATTTTTCGAGTCACATTTTCCTTCATAACTTTTGAA
                                                                   E2
```

Fig. 2. For legend see opposite.

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

```
     F1
int  ----TGAATGGATTTCAAAGAGTTTTAATTTATTTTATGTAAAATTGCACGATCTATTGAATAAAAAAATAATCAAGGGATTTTATCCATCAGAAAAAAAAATAATTAAATTTTTAATTTTTCAAT  1864
haem ----.........................................................................................................................
man  CGGC..........TGCT..T.C........C..G..C..CT.....T...........G.G.T.........AA.C....G.T..C...G........TT.C......C.T.......T..

int  AGAAAGTCCGAGAGAATTTCAGTTTTTTTATCAAAATAATTTTTCCAGTAAATATATAGCAATATGTTA-TACTAATTCAGATAGATATTTGCAAGAGAAATCGAATAAAAAAAGAATGAACGAA  1989
haem .....................................T...............................-..................................................
man  .....................C....GG....G.-........GTAGCC.ATACACAA..ATAAG.TG..ATCA...................TG...T..T................A..
                                                     E3

int  ATCGGTGTASCCGTTCAAAAGTTATGGTCAAAAGTGTGTTCAAGAAATTTGCCGTCTCGGGCATTTCTATAGGATTTTGCCGCTCGGAC-TTTCTATAGGAATTTTTGGGTCACATTTTCCTTCA  2114
haem ......T..G.G...................................C...................................-.....................................
man  ......TG.G...............CGG............GTA.....C..TA......A.T..................TC..C..C......................C...T...T.

                     F2
int  TAACTTTTGAATGGCTGGAAGGATTTCGAAGAGGTTTAATTTATTTTATGTAAAATTGCACGATCTATKGAATAAAAAAATAATCAAGGGATTTTATCCATCAGAAAAAAAAATAATTAAATTTTT  2239
haem ...........---------------------------------------------------------------------------------------------------------------
man  ...........---------------------------------------------------------------------------------------------------------------

int  AATTTTTCAATAGAAAGTCCGAGAGAATTTCAGTTTTTTTATCAAAATAATTTTTCCAGTAAATATATAGCAATATGTTATACTAATTCAGATAGATATTTGCAAGAGAAATCGAATAAAAAAAG  2364
haem ----------------------------------------------------------------------------------------------------------------------------
man  ----------------------------------------------------------------------------------------------------------------------------

int  AATGAACGAAATCGGTGTAGCCGTTCAAAAGTTATGGTCAAAAGTGTGTTCAAGAAATTTGCCGTCTCGGGCATTTCTATAGGATTTTGCCGCTCGGACTTTCTATAGGAATTTTTGGGTCACAT  2489
haem ----------------------------------------------------------------------------------------------------------------------------
man  ----------------------------------------------------------------------------------------------------------------------------

int  TTTCCTTCATAACTTTTGAACGAATTATTATTTTTCAATGGGGTTTTTTTATTTTTCTTATCTCATTGAGTTCTATTCAATGATGTATAATATGTACACATACATTGAGATTTGTGTTTGACGA  2614
haem -------------------.......................................................................................................G
man  -------------------..G....A..................C...G.AG.A.......A......AG.....A...........C.......TG...T.A..A...AAAA...TG.TC
                          E3 (continued)
D2
int  TTTACACATGCTCTTACTCG-CCATGTTATATAGGGATACTACCCCTAACTCAGGGTCTGAAAAGTCGGAAAAATGAAATTTTCGCAAATGCGATAAAAACGATCTTGGTGGTCTAATTATAGTA  2739
haem ...................T.......................................................................................................
man  ......GT...........-......AT.......................................G.................A...................T..............TG.ATC.C..

int  TTCCGGCATCAGACAATGGGTTTTAAGCGTTATTTAGGCCCTAACGCTTCAGGCTGGGCCATTTCGCTTACTACCCTATACATGTGTTGTATGCACTATTAGCTATATGGTTATATAGAGCTGAA  2864
haem .....A..................................................T..............................A.................................
man  .....A..C....................C-...........T.............TG.............-.....TG.....G....A..G..A..........CC..G..........

int  GGTCTCATATGAGGCTGGAGGTTCGCCATTGTTATTGTGGACGGTTCGGCATTGCTTTGGCAAGTTTTGCTGACTGTTCATGCAAAAA-TTGGTTTGCTGGTTAACCGATTTCATTCATACTGGT  2989
haem ...........................................-...............C....................-.......................................
man  ......C.....A.................A.........T...........................G....C....TG....G......G..A..........................

int  GGTTGCTTTTCAGTGTGCCCGAGGGCCATTGTCATTGGGACATTGTAAGGCATGCCGCCATTGCTAATATTAACGAGTTTGACTGTCGTGTAAGCGGAATGGGTCAACTTTGGCCTGTTCGTGCT  3114
haem ...............T.........................C...........A.....................................................................C
man  ........C.......T.T.A......TAG........AG.............AT.........C.........T.......G..........................................

int  GTGCATAATTGATTCACGCGATCATTGATCGGAATAACTAT-ACATGTGATGTGAGTGTGTATATGAATGGTGCGAATGGTCCACCAGGGTTCTTGGTACCTGTGTGGTTGCGCCACAGGATTGA  3239
haem ...........A..........................-......................................................................................
man  ..............T.........C..........TT...A...C.G.....A.A......C.....C.....................GT..........A...................G

                                            G1
int  GCTGTACCATTCAGATTGTAGCGTTAAGGCCTGTACGGTGTCTGGTTCCTCTGCTAAA------------------------------------------------------------CGGGGTTGA  3364
haem ...............................................................------------------------------------------------------.........
man  ..C.AG.TG...G.......A....G.A....T..TA...........TC...TAGGGCGGGATTGAGCTGTACTGCTTGGGCTAAGCCTGGGTAGTGTCTGGCTCTCCTATCCAG....A....
                               H1                                                    H2
                                            G2
int  GCTGCACTGTTTGGGTTTTTACGTGAAAGTCCGGATGGTGTCTGGTTCTCCTGAAAACCAATGTAAAATTACCCTGGCCCACAAGGGTATATTATATAACTGTGAAC  3471
haem .........................................................................................................
man  ....T............GCCT..........T.................A...........G..C..G.........C.....T........................
```

Fig. 2. Alignment of IGS sequences for *Schistosoma intercalatum*, *S. haematobium* and *S. mansoni*. Sequences for the 2 former species represent the largest of the cloned IGS PCR products. The sequences for *S. mansoni* are derived from the ribosomal probes pSM389 and pSM890. A dot (.) indicates that the sequence at that point is the same as *S. intercalatum*. Hyphens (-) indicate alignment spaces. Repeats are underlined and given a letter in order of their appearance within the figure. The main A+T region and it's smaller satellite are in bold lettering while chi-like sites are delineated in bold, italic script. The inverted arrow (↓) indicates the putative end of the large subunit (28S) gene.

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

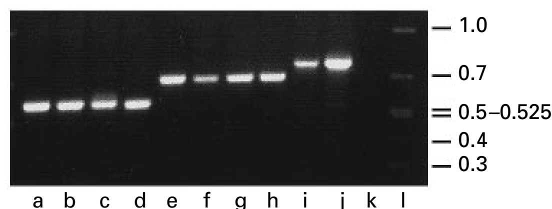



Fig. 3. Amplification of the fragment containing the A+T rich region using primers MNTS1 and NTS15. In all cases DNA was extracted from isolates of pooled worms. Letters a–d are isolates of *Schistosoma haematobium*, respectively, Cameroon, Mali, Mauritius and Malawi (NHM codes 1848, 1804, 2667, and Cape Maclear), e–h isolates of *S. mansoni*, respectively, Senegal, Egypt, Brazil and Kenya (NHM codes 2922, 2917, 3054, and 2834) whilst i and j are *S. intercalatum* from 2 locations in São Tomé, Chacara and San Antonio (NHM codes 1784 and 2756). The negative control and reference marker – sizes given in kilobases (Cambio) are labelled k and l.

inverted repeat D2 of *S. intercalatum*, *S. haematobium* and *S. mansoni* with the exception that in the latter species, the positions of the first T and A were reversed.

### *3′ end region of the large ribosomal (28S) subunit*

The final 464 bp of the 28S gene prior to the putative start of the IGS was identical in all 3 species with the exception of a substitution of a C for an A in *S. mansoni* 65 bp from the proposed 28S/IGS junction.

## DISCUSSION

A+T rich regions of sequence are often associated with origins of DNA replication and are known as DNA Unwinding Elements (DUEs). These have been identified in both prokaryotes and also eukaryotes and usually range from 30 to > 100 bp in length (Sinden, 1994). Replication origins have been located within the rDNA of the sea urchin *Lytechinus variegatus* where the initiation of replication has been found to be confined to a region within the IGS (Botchan & Dayton, 1982). A similar situation has been described concerning the extrachromosomal rRNA genes within *Physarum polycephalum* where up to 4 site-specific origins of replication within the IGS have been identified (Bénard, Lagnel & Pierron, 1995). Gruendler *et al.* (1991) whilst studying the ribosomal IGS of *Brassicaceae* (common wall cress, radish and white mustard) noted that the regions around putative promoters were A+T rich and suggested that the consequent lowering of duplex stability might affect the kinetics of DNA melting during transcription.

A+T rich sequences in general appear to be common across the entire taxonomic spectrum. The origin of such sequences in *S. haematobium* and *S. intercalatum* must be open to speculation, however, the lack of direct repeats on both sides of the region would suggest that their origins are not due to transposition. Jeffreys, Wilson & Thein, (1985) faced a similar situation with minisatellites from human DNA. None of these satellites were flanked by direct repeats and they proposed that the presence of a chi-like sequence was probably promoting recombination. The position within the *Schistosoma* IGS appears to be comparable. The true chi site in *E. coli* has the sequence 5′ GCTGGTGG 3′ (Cheng & Smith, 1987). At a site 151 bp 3′ of the A+T region lies a sequence which deviates from the true chi sequence by only 1 base, namely 5′ GCTGGTCG 3′ and it may be of significance that *S. mansoni* lacks both this sequence and a large A+T rich region. The area of sequence from the putative start of the IGS to just beyond the 3′ end of the A+T region, as defined within this study, contains 36 copies of the sequence 5′ TATG 3′ and 16 copies of the sequence 5′ TGCAT 3′. Thirty copies of the former and 11 copies of the latter lie within the A+T region itself. Whether these short sequences are merely a consequence of repeated recombination and duplication or if they are actively involved in inducing this process is unknown. It is of interest to note that the 3′/5′ junction of the tandem repeats within the ITS1 are all straddled by the sequence 5′ TATG 3′ and, in addition, all contain the sequence 5′ TGCAT 3′ as well as a chi-like site (Kane *et al.* 1996). Whatever the origin of these sequences, there seems little doubt that crossing over has occurred through time resulting in the A+T structure. Should this act as an origin of replication, aid to duplex melting during transcription, or have no defined function at all, it's effect is most likely to destabilize the duplex in so far as denaturation or melting will occur first within this region.

Recombination may also be responsible for the existence of the numerous other repeats present within the IGS of all 3 species. Although the repeats do not contain chi-like sites, 2 such sites do exist 53 bp (5′ CTTGGTGG 3′) and 317 bp (5′ ACTGGTGG 3′) 3′ of the final inverted repeat D2. This is similar to the format found within *Opechona bacillaris* (EMBL accession number Z29504) and an undescribed *Dolichosaccus* species (Luton, Walker & Blair, 1992) both of which possess large tandem repeats within the ITS and a chi-like site 3′ of the tandem repeat block (Kane *et al.* 1996).

Inverted repeats frequently occur at origins of DNA replication and putative control regions of genes. If the interstrand hydrogen bonds of an inverted repeat are melted then the repeat has the potential to form a hairpin (snapback) or even a cruciform structure. Sinden (1994) has indicated, that the positioning of such structures near origins of replication and control regions suggests that alternate DNA conformations may 'provide a molecular switch for controlling transcription or replication'.

The IGS of *Schistosoma* contains many repeats

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

and the indications are that recombination is a relatively frequent event. In general, the sequences of *S. haematobium* and *S. intercalatum* are well conserved within the IGS (96·5 %). When compared with the other 2 species, the sequence of *S. mansoni* has a good though lesser degree of similarity for most of the IGS (75·9 % with *S. intercalatum* and 77·7 % with *S. haematobium*) with the exception of a region starting approximately 150 bp on the 3′ side of the putative starting point of the IGS and which continues for roughly 500 bp. This area loosely corresponds to the A+T rich region of the other 2 species.

Most of the rRNA unit for the genus *Schistosoma* has now been sequenced in one species or another (Rollinson *et al*. 1997), namely the 18S gene (Omer Ali *et al*. 1991; Johnston, Kane & Rollinson, 1993), internal transcribed spacer (Després *et al*. 1992; Bowles *et al*. 1993; Kane & Rollinson, 1994; Bowles, Blair & McManus, 1995; Després *et al*. 1995; Kane *et al*. 1996), and approximately the first quarter and small central 'gap' region of the 28S gene (Mertz *et al*. 1991; Littlewood & Johnston, 1995). With the current study taken into consideration, this only leaves approximately 2·0 kb of the 28S gene to be analysed for the entire ribosomal gene complex to have been sequenced. When, the whole sequence is available for study, a comprehensive analysis of structure and function within this region can be undertaken. For example, a natural extension of the current study would be the determination of the start of transcription by RNA Polymerase I within the IGS along with the delineation of promoter regions. Recognition sites for protein attachment can be sought and the dynamics of their involvement with the regulatory mechanisms of transcription and ribosomal maturation inferred.

One immediate benefit of the current study is the ability to identify pure strains of *S. haematobium* from *S. intercalatum* and *S. mansoni* through the PCR amplification of the A+T region. This provides a much quicker and less involved procedure for identification than blotting and probe hybridization methods and allows the testing of individual worms.

## REFERENCES


ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

BÉNARD, M., LAGNEL, C. & PIERRON, G. (1995). Site-specific initiation of DNA replication within the non-transcribed spacer of *Physarum* rDNA. *Nucleic Acids Research* **23**, 1447–1453.

BOTCHAN, P. M. & DAYTON, A. I. (1982). A specific replication origin in the chromosomal rDNA of *Lytechinus variegatus*. *Nature, London* **299**, 453–456.

BOWLES, J., HOPE, M., TIU, W. U., LIU, X. & McMANUS, D. P. (1993). Nuclear and mitochondrial genetic markers highly conserved between Chinese and Philippine *Schistosoma japonicum*. *Acta Tropica* **55**, 217–229.

BOWLES, J., BLAIR, D. & McMANUS, D. P. (1995). A molecular phylogeny of the human schistosomes. *Molecular Phylogenetics and Evolution* **4**, 103–109.

CHENG, K. C. & SMITH, G. R. (1987). Cutting of chi-like sequences by the RecBCD enzyme of *Escherichia coli*. *Journal of Molecular Biology* **194**, 747–750.

DESPRÉS, L., IMBERT-ESTABLET, D., COMBES, C. & BONHOMME, F. (1992). Molecular evidence linking hominid evolution to recent radiation of schistosomes (Platyhelminths: Trematoda). *Molecular Phylogenetics and Evolution* **4**, 295–304.

DEPRÉS, L., KRUGER, F. J., IMBERT-ESTABLET, D. & ADAMSON, M. L. (1995). ITS2 ribosomal RNA indicates *Schistosoma hippopotami* is a distinct species. *International Journal for Parasitology* **25**, 1509–1514.

DOELLING, J. H., GAUDINO, R. J. & PIKAARD, C. S. (1993). Functional analysis of *Arabidopsis thaliana* rRNA gene and spacer promoters *in vivo* and by transient expression. *Proceedings of the National Academy of Sciences, USA* **90**, 7528–7532.

GRUENDLER, P., UNFRIED, I., PASCHER, K. & SCHWEIZER, D. (1991). rDNA intergenic region from *Arabidopsis thaliana* structural analysis, intraspecific variation and functional implications. *Journal of Molecular Biology* **221**, 1209–1222.

HILLIS, D. M. & DIXON, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology* **4**, 411–446.

JEFFREYS, A. J., WILSON, V. & THEIN, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature, London* **314**, 67–73.

JOHNSTON, D. A., KANE, R. A. & ROLLINSON, D. (1993). Small subunit (18S) ribosomal RNA gene divergence in the genus *Schistosoma*. *Parasitology* **107**, 147–156.

KANE, R. A. & ROLLINSON, D. (1994). Repetitive sequences in the ribosomal DNA internal transcribed spacer of *Schistosoma haematobium*, *Schistosoma intercalatum* and *Schistosoma mattheei*. *Molecular and Biochemical Parasitology* **63**, 153–156.

KANE, R. A., RIDGERS, I. L., JOHNSTON, D. A. & ROLLINSON, D. (1996). Repetitive sequences within the first internal transcribed spacer of ribosomal DNA in schistosomes contain a chi-like site. *Molecular and Biochemical Parasitology* **75**, 265–269.

LITTLEWOOD, D. T. J. & JOHNSTON, D. A. (1995). Molecular phylogenetics of the four *Schistosoma* species groups determined with partial 28S ribosomal RNA gene sequences. *Parasitology* **111**, 167–175.

LUTON, K., WALKER, D. & BLAIR, D. (1992). Comparisons of ribosomal internal transcribed spacers from two congeneric species of flukes (Platyhelminths: Trematoda: Digenea). *Molecular and Biochemical Parasitology* **56**, 323–328.

MAIDAK, B. L., OLSEN, G. J., LARSEN, N., OVERBEEK, R., McCAUGHEY, M. J. & WOESE, C. R. (1997). The RDP (Ribosomal Database Project). *Nucleic Acids Research* **25**, 109–111.

MERTZ, P. M., BOBEK, L. A., REKOSH, D. M. & LOVERDE, P. T. (1991). *Schistosoma haematobium* and *S. japonicum*: analysis of the ribosomal RNA genes and determination of the 'gap' boundaries and sequences. *Experimental Parasitology* **73**, 137–149.

OMER ALI, P., SIMPSON, A. J. G., ALLEN, R., WATERS, A. P.,


https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press

HUMPHRIES, C. J., JOHNSTON, D. A. & ROLLINSON, D. (1991). Sequence of a small subunit rRNA gene of *Schistosoma mansoni* and its use in phylogenetic analysis. *Molecular and Biochemical Parasitology* **46**, 201–208.

ROLLINSON, D., WALKER, T. K., KNOWLES, R. J. & SIMPSON, A. J. G. (1990). Identification of schistosome hybrids and larval parasites using rRNA probes. *Systematic Parasitology* **15**, 65–73.

ROLLINSON, D., KAUKAS, A., JOHNSTON, D. A., SIMPSON, A. J. D. & TANAKA, M. (1997). Some molecular insights into schistosome evolution. *International Journal for Parasitology* **27**, 11–28.

SIMPSON, A. J. G., DAME, J. B., LEWIS, F. A. & McCUTCHAN, T. F. (1984). The arrangement of ribosomal RNA genes in *Schistosoma mansoni*. Identification of polymorphic structural variants. *European Journal of Biochemistry* **139**, 41–45.

SINDEN, R. R. (1994). *DNA Structure and Function*. Academic Press, London.

SOLLNER-WEBB, B. & TOWER, J. (1986). Transcription of cloned eukaryotic ribosomal RNA genes. *Annual Review of Biochemistry* **55**, 801–830.

WALKER, T. K, ROLLINSON, D. & SIMPSON, A. J. G. (1986). Differentiation of *Schistosoma haematobium* from related species using cloned ribosomal RNA gene probes. *Molecular and Biochemical Parasitology* **20**, 123–131.

https://doi.org/10.1017/S0031182098003059 Published online by Cambridge University Press