MOLECULAR AND CELLULAR BIOLOGY, Aug. 1988, p. 3008–3016
0270-7306/88/083008-09$02.00/0

Vol. 8, No. 8

Small Gene Family Encoding an Eggshell (Chorion) Protein of the Human Parasite *Schistosoma mansoni*

LIBUSE A. BOBEK,[1]* DAVID M. REKOSH,[1,2] AND PHILIP T. LoVERDE[1]

Departments of Microbiology[1] and Biochemistry,[2] State University of New York at Buffalo, Buffalo, New York 14214

Received 25 January 1988/Accepted 21 April 1988

We have isolated six independent genomic clones encoding schistosome chorion or eggshell proteins from a *Schistosoma mansoni* genomic library. A linkage map of five of the clones spanning 35 kilobase pairs (kbp) of the *S. mansoni* genome was constructed. The region contained two eggshell protein genes closely linked, separated by 7.5 kbp of intergenic DNA. The two genes of the cluster were arranged in the same orientation, that is, they were transcribed from the same strand. The sixth clone probably represents a third copy of the eggshell gene that is not contained within the 35-kbp region. The 5′ end of the mRNA transcribed from these genes was defined by primer extension directly off the RNA. The ATCAT cap site sequence was homologous to a silkmoth chorion PuTCATT cap site sequence, where Pu indicates any purine. DNA sequence analysis showed that there were no introns in these genes. The DNA sequences of the three genes were very homologous to each other and to a cDNA clone, pSMf61-46, differing only in three or four nucleotides. A multiple TATA box was located at positions −23 to −31, and a CAAAT sequence was located at −52 upstream of the eggshell transcription unit. Comparison of sequences in regions further upstream with silkmoth and *Drosophila* sequences revealed several very short elements that were shared. One such element, TCACGT, recently shown to be an essential *cis*-regulatory element for silkmoth chorion gene promoter function, was found at a similar position in all three organisms.

Schistosoma mansoni is a dioecious blood fluke that inhabits the portomesenteric circulation of humans. Schistosomiasis, a chronic and debilitating disease caused by this parasite, afflicts 200 million people throughout the world. The eggs produced by the mature female parasite are responsible for disease. Physical development and sexual maturation of the female worm are dependent on pairing with the male parasite. Accordingly, the details of how the female worm matures and the molecular interactions which take place between the male and female worms provide an interesting and important phenomenon to study and an unusual mechanism of gene regulation.

In our previous studies, we isolated and characterized a cDNA clone (pSMf61-46) which was derived from a mRNA present only in mature female schistosomes. This mRNA appeared to be developmentally regulated, since it was not found in any other parasite stage. The accumulation of the mRNA in schistosome females coincided temporally with the pairing of females with male schistosomes and with subsequent egg production (1, 3).

The DNA sequence of this full-length cDNA clone revealed an open reading frame capable of encoding a 16-kilodalton (kDa) polypeptide, very rich in glycine (44%) and tyrosine (11%). There was a strong correlation between the amino acid composition of this protein and the actual composition of the schistosome eggshell (4). This fact, in addition to the biological context in which the mRNA was found, the significant homology in sequence to the silkmoth chorion gene family (3), and the recognition of an eggshell protein of 14 kDa by an antiserum directed against a peptide from the deduced amino acid sequence of the cDNA clone (unpublished data), led us to conclude that the clone encoded a protein that is a major component of the schistosome eggshell (chorion).

Silkmoth and *Drosophila* chorion genes have been used as model systems to study the organization, coordinate expression, and evolution of multigene families. Like the chorion genes of the silkmoth and *Drosophila melanogaster*, the eggshell protein genes of *S. mansoni* are developmentally regulated, showing a high sex, tissue, and temporal specificity of expression (1, 3, 5, 8, 21).

The silkmoth chorion (eggshell) consists of more than 100 different proteins, encoded by a large number of genes belonging to a gene superfamily with two symmetrical branches, each consisting of three families (7, 12). The majority, if not all, of the chorion genes in the silkmoth *Bombyx mori* are clustered as tandem duplications in a giant locus, probably exceeding 10^6 base pairs (bp) at one end of a single chromosome (7). The very high rates of chorion gene expression required for eggshell synthesis are achieved by the redundancy of the chorion locus. On the other hand, *Drosophila* chorion is composed of approximately 20 structural proteins (14). Chorion genes are located within two clusters, one on the third chromosome and the other on chromosome X (10, 13, 25). The high rates of chorion gene expression in *D. melanogaster* are achieved by specific amplification of chorion genes and their flanking sequences (18, 19, 23).

We have previously determined the copy number of the *S. mansoni* eggshell protein gene. Our data indicated that there were approximately five copies per haploid genome. We also showed that these genes did not amplify or rearrange during schistosome development (1, 3).

In the present study, we examined the genomic organization of the schistosome eggshell genes. To do this, we isolated and characterized genomic sequences related to the cDNA clone pSMf61-46. We present the characterization and sequence analysis of six independent recombinant clones, the organization of the eggshell protein genes in the *S. mansoni* genome, and a comparison of *S. mansoni* eggshell protein genes with those of silkmoth and *D. melanogaster*.

* Corresponding author.

MATERIALS AND METHODS

Parasites. Cercariae of *S. mansoni* (Naval Medical Research Institute) obtained from previously infected snails *Biomphalaria glabrata* were used for infection of hamsters. Schistosome worms were obtained from hamsters by perfusing the hepatic portal system (6).

DNA isolation. Schistosome worm DNA was prepared as described previously (24).

Construction of genomic library. Standard DNA manipulations were performed as described in protocols in the *Molecular Cloning Manual* (15). A genomic library was constructed as described elsewhere (2).

Southern blot analysis. Fragments of restriction enzyme-digested genomic DNA were size-separated on a 1% agarose gel; fragments of digested DNA of lambda recombinant clones were separated on 0.6% agarose gels. DNA fragments were transferred from gels to a Zetabind hybridization membrane (AMF Cuno, Meriden, Conn.) in 20× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate). In general, the hybridizations were performed at 42°C in 50% formamide for 16 h. Other hybridization and washing conditions were followed from the specifications of the Zetabind manufacturer. The hybridization probes were prepared by nick translation.

DNA sequencing. DNA fragments of lambda recombinant clones were subcloned into M13mp19 bacteriophage and then sequenced by the dideoxynucleotide chain termination method (20).

RNA sequencing and primer extension reaction. A procedure published earlier (J. Geliebter, Focus, Spring 1987, p. 5; Bethesda Research Laboratories) was followed except that total RNA (not polyadenylated RNA only) isolated from adult schistosome worm pairs was used. The 5′ end of the mRNA was sequenced by using a synthetic oligonucleotide (22-mer) that hybridizes to the cDNA pSMf61-46 mRNA strand around the ATG start codon (see below). Briefly, 100 ng of oligonucleotide was labeled in 20 μl of reaction mix containing 100 μCi of [γ-^{32}P]ATP (5,000 Ci/mmol) with 7 U of T4 polynucleotide kinase for 30 min at 37°C. The final concentration of oligonucleotide was 5 ng/μl. Unincorporated [γ-^{32}P]ATP was not removed. For one sequencing reaction, 5 ng of the labeled primer was used to anneal to about 20 to 40 μl of total RNA (in 12-μl reaction volume) for 45 min at a temperature dependent on the sequence of the primer. Afterwards, 20 U of avian myeloblastosis virus reverse transcriptase was added, and the annealed mixture was equally split into five tubes, labeled A, C, G, T, and N (N represented a primer extension reaction, a tube without dideoxynucleoside triphosphates). Each tube contained 3.3 μl of reverse transcriptase buffer and 1 μl of either 1 mM dideoxy-ATP (ddATP), ddCTP, or ddGTP, 2 mM ddTTP, or 1 μl of H_2O (tube N). The reaction mix was incubated at 50°C for 45 min, and the reaction was stopped by addition of 4 μl of stop buffer (100% deionized formamide, 0.3% each bromophenol blue and xylene cyanol). Samples were boiled, and half of each reaction mix was loaded onto an 8% polyacrylamide sequencing gel.

RESULTS

Analysis of *S. mansoni* genomic DNA by Southern blotting and hybridization. In order to identify genomic DNA fragments homologous to the eggshell protein cDNA clone pSMf61-46, a Southern blot hybridization analysis (Fig. 1) was performed on genomic DNA isolated from adult schis-

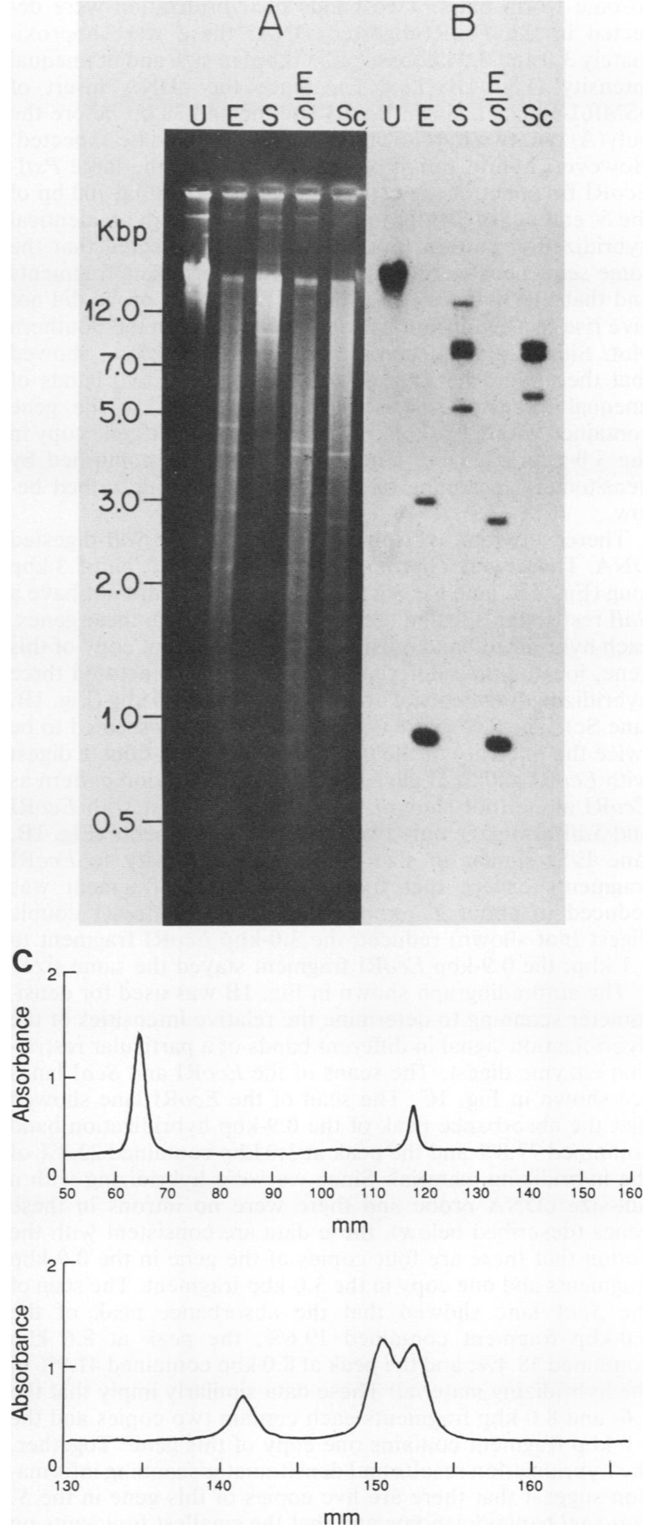

FIG. 1. Southern analysis of *S. mansoni* genomic DNA. (A) Genomic DNA was digested with either *Eco*RI (lane E), *Sal*I (lane S), *Eco*RI-*Sal*I (lane E/S), or *Sca*I (lane Sc). A 10-μg amount of DNA was used for each digest. Lane U, Undigested DNA. DNA was size-separated on a 1% agarose gel and stained with ethidium bromide. (B) The gel in panel A was Southern blotted and hybridized to an insert of cDNA clone pSMf61-46. (C) Densitometer scan of *Eco*RI and *Sca*I lanes (upper and lower panel, respectively) of the autoradiograph shown in panel B.

tosome worm pairs. Two bands of hybridization were detected in the *Eco*RI-digested DNA; these were approximately 3.0 and 0.9 kilobase pairs (kbp) in size and of unequal intensity (Fig. 1B, lane E). Since the cDNA insert of pSMf61-46 has a single *Eco*RI site located 30 bp before the poly(A) tail, two hybridization fragments could be expected. However, hybridization of the same blot to the large *Pst*I-*Eco*RI fragment of the cDNA insert (representing 700 bp of the 5′ end out of 765 bp total length) resulted in an identical hybridization pattern (not shown). This indicated that the same sequences were located in both hybridizing fragments and that the 30 bp of the 3′ end of the cDNA probe did not give rise to a significant hybridization signal on the Southern blot. Since the sequencing data, described below, showed that there were no introns in the genes, the two bands of unequal intensity suggested multiple copies of the gene contained within 0.9-kbp fragments and a single gene copy in the 3.0-kbp fragment. The quantitation was confirmed by densitometer scanning of an autoradiograph described below.

There were four hybridizing fragments in the *Sal*I-digested DNA. These were approximately 14.0, 9.0, 8.2, and 5.3 kbp long (Fig. 1B, lane S). Since the cDNA clone did not have a *Sal*I restriction site and there were no introns in these genes, each hybridized band must represent a different copy of this gene, located on a different fragment. *Sca*I generated three hybridizing fragments of about 8.6, 8.0, and 6.0 kbp (Fig. 1B, lane Sc). The 8.6- and 8.0-kbp fragments each seemed to be twice the intensity of the 6.0-kbp fragment. A double digest with *Eco*RI and *Sca*I gave the same hybridization pattern as *Eco*RI alone (not shown), and a double digest with *Eco*RI and *Sal*I also gave only two hybridizing fragments (Fig. 1B, lane E/S) similar in size and relative intensity to *Eco*RI fragments except that the 3.0-kbp *Eco*RI fragment was reduced to about 2.7 kbp. A *Hin*dIII and *Eco*RI double digest (not shown) reduced the 3.0-kbp *Eco*RI fragment to 1.1 kbp; the 0.9-kbp *Eco*RI fragment stayed the same size.

The autoradiograph shown in Fig. 1B was used for densitometer scanning to determine the relative intensities of the hybridization signal in different bands of a particular restriction enzyme digest. The scans of the *Eco*RI and *Sca*I lanes are shown in Fig. 1C. The scan of the *Eco*RI lane showed that the absorbance peak of the 0.9-kbp hybridization band contained 77.8% and the peak at 3.0 kbp contained 22.8% of the hybridizing material. Since we were hybridizing with a full-size cDNA probe and there were no introns in these genes (described below), these data are consistent with the notion that there are four copies of the gene in the 0.9-kbp fragments and one copy in the 3.0-kbp fragment. The scan of the *Sca*I lane showed that the absorbance peak of the 6.0-kbp fragment contained 19.6%, the peak at 8.6 kbp contained 38.4%, and the peak at 8.0 kbp contained 41.9% of the hybridizing material. These data similarly imply that the 8.6- and 8.0-kbp fragments each contain two copies and the 6.0-kbp fragment contains one copy of this gene. Together, the hybridization results and densitometer scanning information suggest that there are five copies of this gene in the *S. mansoni* haploid genome and that the smallest fragments on which these are located are 0.9-kbp *Eco*RI fragments. These results are in accord with our previous data, which led us to conclude that there were about two to five copies per haploid genome.

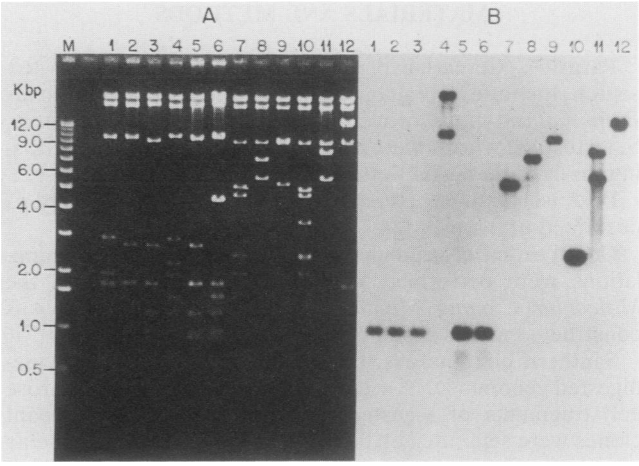

FIG. 2. Southern analysis of six indepedent lambda recombinant clones. (A) DNA (2 μg) from each clone was digested with either *Eco*RI (lanes 1 to 6) or *Sal*I (lanes 7 to 12), size-separated on a 0.6% agarose gel, and stained with ethidium bromide, Lane M, DNA size markers. (B) Hybridization fragments after Southern blotting of the gel and hybridization to an insert of cDNA clone pSMf61-46. Lanes 1 and 7, clone 6-1-3; lanes 2 and 8, clone 6-2-4; lanes 3 and 9, clone 7-1; lanes 4 and 10, clone 8-1-4; lanes 5 and 11, clone 10-1; lanes 6 and 12, clone 10-2.

Isolation and analysis of recombinant genomic clones. Genomic DNA isolated from *S. mansoni* adult worm pairs, partially digested with *Sau*3A, was ligated into the *Bam*HI site of the bacteriophage lambda EMBL3 replacement vector as described elsewhere (2). About 100,000 bacteriophage were screened, without prior library amplification, by plaque hybridization with the female-specific cDNA clone pSMf61-46 as a hybridization probe. Six independent hybridizing recombinants were identified, purified, and analyzed further.

DNA from each of the six genomic clones was separately digested with *Eco*RI and *Sal*I restriction enzymes (Fig. 2A). The resulting fragments were subjected to Southern blot analysis with the cDNA clone pSMf61-46 as a hybridization probe. This allowed the gene to be localized to a particular restriction fragment(s) (Fig. 2B). In five of the six clones, the hybridizing *Eco*RI fragments were internal fragments, 0.9 kbp in length—the same size as the fragments in the genomic DNA Southern blot (Fig. 2B, lanes 1 to 3, 5, 6). One clone (8-1-4, Fig. 2B, lane 4) did not contain the internal hybridizing *Eco*RI fragment, since it was not excised by *Eco*RI but stayed attached to the right arm of the lambda vector. (The larger hybridizing fragment seen in this lane represents the two phage arms annealed by their cohesive ends). Since there was a *Sau*3A site in the cDNA clone (located about 90 bp upstream from the 3′-end *Eco*RI site) and since the cloned genomic DNA were *Sau*3A fragments, it was predicted that this clone might contain the part of the gene up to this *Sau*3A site. This was confirmed by mapping (Fig. 3) and DNA sequencing. The hybridization pattern of the *Sal*I-digested clones (Fig. 2B, lanes 7 to 12) showed that five of the six clones had one hybridizing fragment, while one clone (10-1, Fig. 2B, lane 11) had two fragments. This suggested that this clone spanned two closely linked copies of the gene. The two copies were confirmed by restriction enzyme mapping, described in the next section.

In the collection of six genomic clones, we have not selected a genomic clone which would represent the gene contained within the 3.0-kbp *Eco*RI fragment (seen in the Southern blot of genomic DNA, Fig. 1, lane E). From the *Eco*RI fragment hybridization patterns, it was evident that five of six clones contained the 0.9-kbp hybridizing fragment. The only clone which could possibly contain a part of

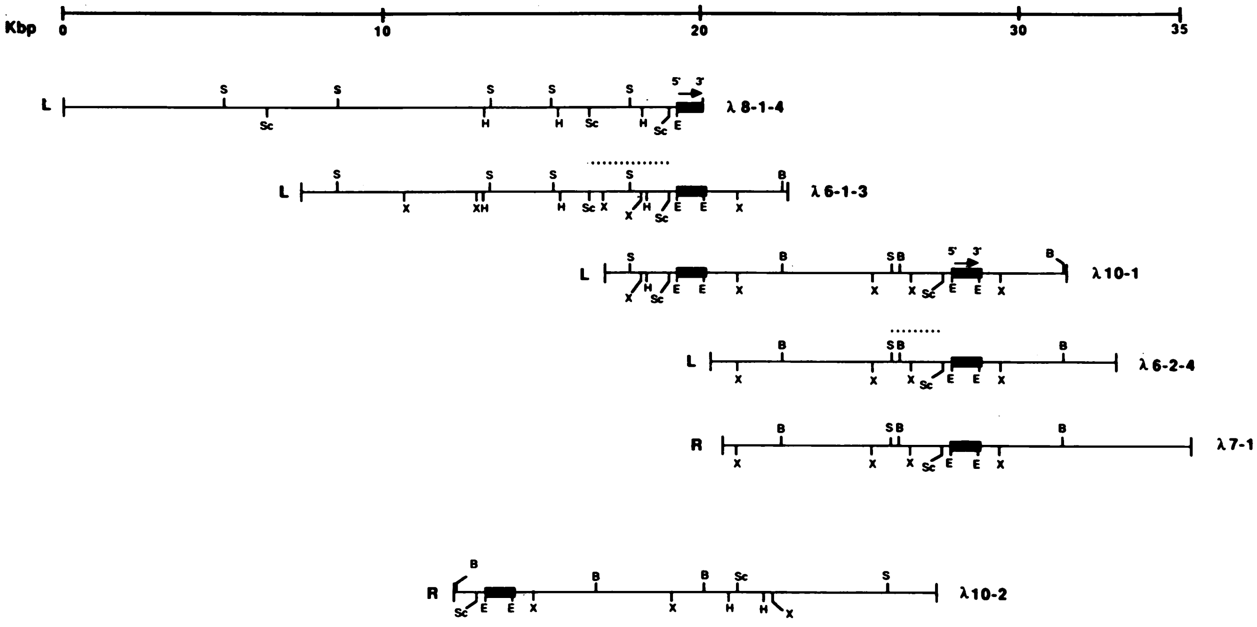

FIG. 3. Restriction enzyme maps of the inserts of six lambda recombinant clones and their linkage arrangement. Endonuclease sites: S, *Sal*I; Sc, *Sca*I; H, *Hin*dIII; E, *Eco*RI; B, *Bgl*II; X, *Xba*I. *Bgl*II and *Xba*I maps were not determined for clone 8-1-4. *Hin*dIII maps were not determined for clones 6-2-4 and 7-1. For *Eco*RI, only the sites confining the gene sequences are shown, although there were many others in each clone. Arrows indicate gene orientations with respect to transcription. Solid boxes indicate the gene sequences including 226 bp upstream of the transcriptional start (up to the 5' *Eco*RI site). The dashed lines indicate the additional hybridization probes used for generation of the linkage map. L and R represent where the left and right arms, respectively, of the lambda vector are located with respect to the schistosome DNA inserts.

the 3.0-kbp *Eco*RI fragment was clone 8-1-4, but upon further analysis this clone was found to overlap with clones 6-1-3 and 10-1, both of which contained the internal 0.9-kbp hybridizing fragment.

Construction of restriction enzyme maps and linkage map of cloned eggshell protein genes. Detailed restriction enzyme maps and the position of each gene within each clone are shown in Fig. 3. These maps were derived form restriction patterns of single and double enzyme digests of all clones and the hybridization patterns to the gene probe (the cDNA clone pSMf61-46) and to the lambda vector probes (the right and left arms of lambda EMBL3).

Restriction maps showed that five clones contained one copy of the eggshell protein gene, and one clone (10-1) contained two closely linked copies, separated from each other by 7.5 kbp of intergenic DNA. The precise order of the two genes within the clone 10-1 was important for further work with this clone and was determined in the following way (data not shown). There were two *Xba*I-*Xba*I fragments of very similar size, 3.0 kbp and 2.8 kbp, hybridizing to the gene probe. *Xba*I-*Hin*dIII double digestion reduced the larger (3.0-kbp) hybridizing fragment to 2.6 kbp, but the smaller of the two fragments (2.8 kbp) stayed the same size. Since there was only one *Hin*dIII site in the insert of clone 10-1, the location of which was already known, this determined that the larger of the two fragments was located on the left side of the insert and the smaller of the two on the right side, as drawn in Fig. 3. On the other hand, there was a *Bgl*II site in front of the gene on the right side but not in front of the gene on the left side. Thus, the *Hin*dIII and the *Bgl*II sites distinguished the two genes in their left-side flanking sequences.

It was important to relate the restriction maps of the genomic clones to the genomic Southern data. Of the bands hybridizing to the cDNA pSMf61-46 probe on the genomic Southern blot (Fig. 1B), we identified the 14.0- and 8.2-kbp *Sal*I fragments and the 8.0- and 8.6-kbp *Sca*I fragments in the genomic clones. Clone 10-2 had a 13.0-kbp *Sal*I fragment which must have been derived from the genomic *Sal*I fragment of about 14.0 kbp (Fig. 1, lane S). This *Sal*I fragment gave rise to the hybridizing internal 8.0-kbp *Sca*I fragment when digested with *Sca*I. Clone 10-1 contained an internal 8.2-kbp *Sal*I fragment and an internal 8.6-kbp *Sca*I fragment hybridizing with pSMf61-46.

The arrangement of genomic clones in the linkage map given in Fig. 3 was derived as follows. A Southern blot of the restriction fragments of all six genomic clones digested with *Sal*I, *Sal*I-*Sca*I, and *Sca*I enzymes was first hybridized with the 1.65-kbp *Sal*I-*Sca*I subcloned fragment of clone 6-2-4. This fragment did not include eggshell protein coding sequences and is indicated by the dashed line in Fig. 3. Besides itself, this probe (clone 6-2-4) hybridized very strongly to the same-size *Sal*I-*Sca*I fragment in clones 7-1 and 10-1, suggesting that these three clones overlap. However, the probe also weakly cross-hybridized with *Sal*I-*Sca*I fragments of about 1.3 kbp in clones 10-1, 6-1-3, and 8-1-4, suggesting possible overlaps with two additional clones. These weaker 1.3-kbp *Sal*I-*Sca*I fragments corresponded to the sequences located immediately upstream of the left gene of the cluster.

To verify the overlap of the clones analyzed, the same Southern blot was then hybridized with the 2.6-kbp *Sca*I-*Sca*I fragment of clone 6-1-3, located to the immediate left of the eggshell protein gene (representing the left gene of the cluster) and depicted by the dashed line in Fig. 3. This fragment contained the weakly hybridizing 1.3-kbp *Sca*I-*Sal*I fragment described above. Only two other clones cross-hybridized very strongly with this 2.6-kbp *Sca*I-*Sca*I probe, clone 8-1-4 and clone 10-1. The hybridizing *Sca*I fragments

```
                    -480                -460                -440
  CCCACTAGTGTCTCGTCTACTATCTTAATCAGTATTGTCATGGATGAAATATGACAATATGGAGAATGGATTGTTGGCTG

                    -400                -380                -360
  CCAGATATCCCTTATTTAAGGTCATTGCAACATCGTCTATGAGGTGAACAAGTTAAATTTAAATAAACGTTTATTTAACC

                    -320                -300                -280
  TGTTTAGTTATAACTTGTATGTCCAGAGcAAACATCACTTGTCAGTGTTATCAAAAAGAGTTTCGTCTCATTCAaTGAAG

                    -240              ↓ -220                -200
  ACTTAATCAACAGTTTCGATGAACCACATGCTAACGAATTCTAACGGTAGAATCAAAATAGTGAGTAATTCGTGAATTAA

                    -160                -140                -120
  CTGTTTCATTGACATTTTTAACTATCACGCTCAACTATCATTAACGATTACACAAAAGACAGTTCGACAGTAGAGTGCAC

                    -80                 -60                 -40
  AAGCGTAATATGTATTTAATAATGATGCACTTAGTGAGGCACAACTCTTCAAATCTATAATCAAAAACAATATATAAATG

                       1                 20                  40
  ATAAATCACACTAGTCTACACATCATCACACCCAGTACAACAACACCAACAATTTGAAAA ATG AAA CAG TCA CTC
                                                               MET Lys Gln Ser Leu
cDNA tcaacatctgagcataaag

       60                      80                         100
  ACA CTC GTC TTC TTA GTA GCC ATT GGT TAC GCC ACC GCC CAC ACC ACA TCA CAT GAC TAT
  Thr Leu Val Phe Leu Val Ala Ile Gly Tyr Ala Thr Ala His Thr Thr Ser His Asp Tyr
  10-2                                            C
  cDNA                                            T(Tyr)

      120                     140                         160
  TCG GGT GGG TAC GGT GGC GGT TGC TAT GGT AGC GAT TGT GAT AGC GGT TAT GGC GAT AGT
  Ser Gly Gly Tyr Gly Gly Gly Cys Tyr Gly Ser Asp Cys Asp Ser Gly Tyr Gly Asp Ser

      180                     200                         220
  GGA TAT GGT GGA GGC TGT ACT GGT GGT GAC TGT GGC GGC GGC TAT GGT GGT GGC TAT GGT
  Gly Tyr Gly Gly Gly Cys Thr Gly Gly Asp Cys Gly Gly Gly Tyr Gly Gly Gly Tyr Gly
  10-2                        C A(Ser)            A(Ser)
  cDNA                        C A(Ser)

      240                     260                         280
  GGA GGT TGC AGT GGT GGA GAT TGT GGT AAT TAC GGT GGT GGC TAT GGT GGT GAT TGC AAT
  Gly Gly Cys Ser Gly Gly Asp Cys Gly Asn Tyr Gly Gly Gly Tyr Gly Gly Asp Cys Asn

      300                     320                         340
  GGT GGA GAT TGT GGT AAT TAC GGT GGT GGC TAT GGT GGT GGG AAT GGT GGT GGT TGC AGT
  Gly Gly Asp Cys Gly Asn Tyr Gly Gly Gly Tyr Gly Gly Gly Asn Gly Gly Gly Cys Ser

      360                     380                         400
  GGT GGC AAT TGT GGA GGT GGC TTC GAT GAG GCC TTC CCT GCC CCC TAT GGC GGT GAT TAT
  Gly Gly Asn Cys Gly Gly Gly Phe Asp Glu Ala Phe Pro Ala Pro Tyr Gly Gly Asp Tyr
  10-2                                        C
  cDNA                                        A(Leu)

      420                     440                         460
  GGT AAC GGT GGC AAC GGC TTT GGA AAA GGT GGT AGT AAA GGC AAC AAT TAT GGA AAG GGT
  Gly Asn Gly Gly Asn Gly Phe Gly Lys Gly Gly Ser Lys Gly Asn Asn Tyr Gly Lys Gly

       480                     500                         520
  TAT GGC GGT GGT AGC GGT AAG GGT AAG GGT GGT GGC AAA GGT GGC AAA GGC GGC AAA GGT
  Tyr Gly Gly Gly Ser Gly Lys Gly Lys Gly Gly Gly Lys Gly Gly Lys Gly Gly Lys Gly

      540                      560                        580                 600
  GGC ACT TAC AAA CCC AGC CAT TAT GGA GGC GGT TAC TGA GGCACCAGTTGAGTTGTGGATCATTCT
  Gly Thr Tyr Lys Pro Ser His Tyr Gly Gly Gly Tyr ***

                    620                 640                 660
  AATTTGTTTGTGTCACACTCTCCACTGTCCTATTTTTCTACACACCTCTCAATTCAACTCACTGTAATATAGTCGTGTT

    ↓               700
  TGAATTCGAGATGAATAAAACCTATTCATTCTAAA
  cDNA                               (A)30
```

of other clones were much less intense and corresponded to the fragments containing sequences located in front of the right gene of the cluster. These results confirmed the overlap of the clones, as presented in the linkage map in Fig. 3. The map shows a stretch of 35 kbp of *S. mansoni* DNA with two eggshell protein genes separated by 7.5 kbp of intergenic DNA and flanked by at least 20 kbp of DNA on the left side and 6.0 kbp of DNA on the right side. Since clone 10-2 cross-hybridized only very weakly with both probes and the restriction data did not support the overlap of this clone with the other five clones, clone 10-2 was not included in the linkage map shown in Fig. 3.

Orientation of the genes with respect to transcription. In order to determine the direction of transcription of the two genes within the cluster, we subcloned sequences which included the *Sca*I site located to the left of both genes into the vector M13mp19 and carried out DNA sequence analysis. From clone 10-1, digested with *Sca*I-*Sal*I, we subcloned the 4.0-kbp *Sca*I-*Sal*I fragment containing the right gene (the *Sal*I site in this case was the *Sal*I site in the right arm of the lambda polylinker). The DNA sequence of the relevant regions of this subclone revealed that the *Eco*RI restriction site closest to the *Sca*I restriction site represented the 5′ upstream sequences of this gene. This gene was therefore oriented in the 5′ to 3′ direction with respect to the *Sca*I site, as indicated by an arrow in Fig. 3. The *Sca*I site was 275 bp upstream of the *Eco*RI site. The 3′ end of the gene was also analyzed and agreed with the sequence of the cDNA pSMf61-46. For the left gene, we subcloned the *Sca*I-*Sal*I fragment (*Sal*I from the right arm of lambda) of clone 8-1-4 of about 1.1 kbp (Fig. 3). Sequencing of this subclone showed that this gene was transcribed in the same orientation as the right gene of the cluster, as also indicated by an arrow in Fig. 3.

DNA sequence of the genomic clones and determination of the mRNA cap site. The coding, 5′ upstream, and 3′ downstream regions from three genomic clones were sequenced and are presented as a composite in Fig. 4 together with the sequence of the cDNA clone pSMf61-46, determined previously. Nucleotide 1 was shown to be the mRNA start site by the experiments described in the next section (see Fig. 5). Except for the first 19 nucleotides of the cDNA, which did not coincide with the genomic sequence (Fig. 4), it is clear that the cDNA sequence differed from the sequence of the genomic subclone 6-1-3 in only four nucleotides, indicating that there were no introns in the coding region of the gene. The changes occurred at nucleotide positions 94, 198, 199, and 390, as depicted along with the corresponding amino acid changes in Fig. 4. The DNA sequence of the right gene of clone 10-1 agreed 100% with that of 6-1-3, even in the 5′-flanking sequences. The DNA sequence of genomic subclone 10-2 differed in three nucleotides from those of 6-1-3 and 10-1 (positions 198, 199, and 214). These changes all occurred in the coding region. Significantly, the 5′ and 3′ untranslated sequences and 5′-flanking sequences up to the *Eco*RI site were identical. The sequence of clone 10-2 also differed in three nucleotides from that of the cDNA sequence (positions 94, 214, and 390) (Fig. 4). Reanalysis of the previously published (1, 3) cDNA sequence indicated that nucleotide A at position 390 in the cDNA sequence was an error in DNA sequencing and should be a C.

Analysis of the sequences upstream from the coding region of the gene revealed elements typical of eucaryotic promoters. For example, a multiple TATA box was located at positions −23 to −31 and a CAAAT sequence was present at −52. In addition, other sequences specific to regulatory regions of *D. melanogaster* and silkmoth chorion genes were also present further upstream (see Discussion). The DNA sequence data thus suggest that the region directly upstream from the coding sequence of the eggshell protein contains the eggshell gene promoter. However, there were 19 nucleotides at the 5′ end of the cDNA sequence that did not correspond to any genomic sequence presented in Fig. 4. These 19 nucleotides could be derived from another exon of the gene not contained within the sequenced region or could be a cloning artifact, if the sequenced region represents the eggshell gene promoter. To distinguish between these possibilities, the 5′ end of the mRNA was determined by primer extension and direct dideoxynucleotide sequencing. To do this, total RNA of *S. mansoni* adult worm pairs was used as a template. The primer was a [γ-^{32}P]ATP-labeled synthetic oligonucleotide (22-mer) complementary to the mRNA strand at the ATG (Met) codon (Fig. 4, positions 38 to 60). The results of this experiment are depicted in Fig. 5C. Figure 5A shows the relevant portion of the DNA sequence of the genomic subclone 6-1-3, and Fig. 5B shows the DNA sequence of cDNA clone pSMf61-46. In both panels A and B, sequencing was performed with the same oligonucleotide primer as in panel C. Comparison of the three sequences indicated that an adenine (A) residue was the last common residue in all three sequences (indicated by arrows in Fig. 5). From that point the sequence of the genomic DNA differed from the cDNA sequence, and the sequence of the mRNA ended with specific termination bands. In lane N, in addition to the major band corresponding to a full-length cDNA transcript, several shorter bands, corresponding to premature terminations, were seen. These results clearly show that the 5′ end of the mRNA does not include the 19 nucleotides present at the 5′ end of the cDNA clone pSMf61-46. Thus, the mRNA molecule is 714 bases long, without a poly(A) tail, and is not spliced. The cDNA contained within pSMf61-46 is therefore a full-size cDNA clone.

FIG. 4. Compiled DNA sequences and deduced amino acid sequences of the eggshell protein of several genomic subclones compared with the cDNA clone pSMf61-46. The nucleotide sequence from −501 to −226 (up to the 5′ *Eco*RI site) was determined from the genomic subclone 8-1-4; nucleotides −226 to +686 (the 912-bp *Eco*RI-*Eco*RI fragments) from genomic subclones 6-1-3, 10-1, and 10-2; and nucleotides 681 to the end from clone 10-1. Only the nucleotides which differ from the sequence of the genomic subclone 6-1-3 (within the *Eco*RI-*Eco*RI fragment) are indicated; the rest of the sequence was identical. The changes occurred at nucleotide positions 94, 198, 199, 214, and 390. The nucleotides represented by lowercase letters are the 19 nucleotides at the 5′ end of cDNA pSMf61-46 except for positions −267 and −313, which represent ambiguous bases in sequence autoradiograms. Nucleotide 1 corresponds to the transcription start of this mRNA, as defined by the primer extension experiment shown in Fig. 5. *Eco*RI cuts at positions −225 and +681, indicated by arrows. The multiple TATA box and CAT sequence are boxed. The ATG translational start codon and TGA translation stop codon are both underlined once; the polyadenylation signal is underlined twice. Sequences indicated by dotted lines are the elements conserved among *S. mansoni*, silkmoth, and *D. melanogaster*.

DISCUSSION

The six *S. mansoni* genomic clones that we isolated represent three copies of the eggshell protein gene, each containing a 0.9-kbp *Eco*RI fragment of genomic DNA. The gene copy that is encoded by a 3.0-kbp *Eco*RI genomic fragment has not been isolated. One clone (10-1) contained

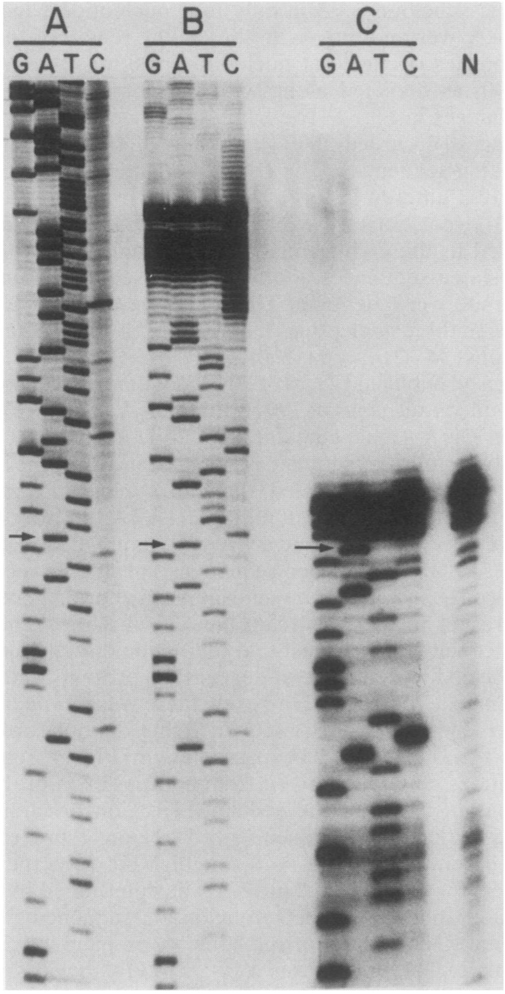

FIG. 5. Dideoxynucleotide sequence of genomic DNA, cDNA pSMf61-46, and mRNA. (A) Part of the DNA sequence of genomic subclone derived from clone 6-1-3. (B) Part of the cDNA pSMf61-46 sequence. (C) Part of the mRNA sequence and a primer extension reaction (lane N). In all panels, a synthetic oligonucleotide (22-mer) complementary to the mRNA strand at the ATG translational start codon (see Fig. 4) was used as a primer. In panels A and B, the primer was not radioactively labeled; in panel C, the primer was end labeled with [γ-^{32}P]ATP. Arrows indicate the last common residue in all three sequences.

two copies of the gene. Four of the five remaining clones overlapped with either the left or right gene in clone 10-1, so that the two eggshell protein genes were found within 35.0 kbp of the *S. mansoni* genome, separated by 7.5 kbp of intergenic DNA. The sixth clone (10-2) differed in restriction pattern (Fig. 3), hybridization of flanking sequences, and DNA sequence (Fig. 4) from the other five clones and thus probably represents a third copy of the eggshell protein gene. However, because these genes and some of the flanking sequences appear to be conserved, we cannot rule out that the differences observed in clone 10-2 are due to polymorphism in the schistosome worm population.

The organization and DNA sequence of the *S. mansoni* chorion genes can be compared with the chorion genes of silkmoth and *D. melanogaster*. Silkmoth genes (of which there are over 100 copies; see the Introduction) are arranged in tightly linked pairs, each member of the pair belonging to a different gene family. They are arranged head to head, divergently transcribed from different strands, and coordinately expressed. The 5′-flanking region between the two genes in each pair is short (ca. 300 bp), and the pairs are separated by long segments of 3′-flanking DNA (for review, see reference 7). In contrast, *Drosophila* chorion genes are arranged tail to head, are transcribed from the same strand, are not found in pairs, and are separated form each other by about 1 kbp. Six of the proposed 20 genes have been identified. They are each present in a single copy and code for a different protein (for review, see reference 10).

The two *S. mansoni* eggshell protein genes contained within genomic clone 10-1 were identical throughout the sequenced region, including the flanking sequences. They were arranged tail to head and were transcribed from the same strand, as is the case for *Drosophila* chorion genes, but were separated from each other by 7.5 kbp. They were flanked by at least 20 kbp of DNA upstream of the left gene of the cluster (which did not contain another related chorion gene) and at least 6.0 kbp of DNA downstream of the right gene. The third *S. mansoni* chorion gene copy (contained within clone 10-2) differed from the other two copies in three nucleotides, all located in the coding region. However, the 5′ and 3′ untranslated sequences and 5′-flanking sequences up to the *Eco*RI site were identical. There were no introns in the *S. mansoni* eggshell protein genes, but the silkmoth chorion genes have a single intron that is invariably located at the same position within the signal peptide-encoding region of each gene. The *Drosophila* chorion genes also contain a small intron(s) near the 5′ end. There is a strong amino acid homology between the schistosome eggshell protein and the silkmoth chorion A and B protein gene families (1, 3). However, there is no homology between silkmoth and *Drosophila* proteins.

The 5′ regulatory regions of the silkmoth and *Drosophila* chorion genes have been studied in detail. When cloned DNA fragments bearing silkmoth chorion genes were introduced into the *Drosophila* germ line by P-element-mediated transformation, transformed lines accumulated abundant silkmoth chorion RNA transcripts with correct sex, tissue, and temporal specificity (16). This suggests that regulatory elements responsible for chorion gene expression are conserved between *D. melanogaster* and silkmoths. In fact, DNA sequence comparisons between *Drosophila* and silkmoth 5′-flanking regions reveal numerous, very short sequence elements that are shared (22). One such element, TCACGT, found in silkmoths is also associated with all five *Drosophila* chorion genes sequenced to date. This element is found at positions −58 to −65 upstream of the cap site of *Drosophila* chorion genes and at positions −67, −87, and −87 from the cap site of B.L12, B.L11, and HcB12, respectively, in *B. mori*. Interestingly, this element is not found in upstream sequences of other *Drosophila* genes (25). Mitsialis et al. (17), using the bacterial gene chloramphenicol acetyltransferase as an expression marker followed by P-element-mediated germ line transformation in *D. melanogaster*, reported that the hexanucleotide TCACGT is essential for silkmoth chorion gene promoter function. We found a related element (the same sequence in a backward orientation) upstream of all *S. mansoni* eggshell protein genes sequenced to date, located at position −71 (indicated by a dotted line, as are the other conserved elements, in Fig. 4). A very similar element, TCACGCT, was located at −157 from the cap site. Another homologous element between the silkmoth A/B.L12 gene and *Drosophila* s15-1 gene has the sequence GTAGAAT (22). This exact element was found in the *S. mansoni* sequence at position −215. The *Drosophila* s15-1 upstream sequence also shares one further upstream element with A/B.L12. This sequence reads AGTGTATTC and begins at −279 for s15-1 and −239 for B.L12 (22). A similar element, reading AGTGTTATC, was also found in the upstream sequence of *S. mansoni* chorion genes, starting at position −298 from the cap site (22).

Another similarity found between the chorion genes of silkmoth and *S. mansoni* are the cap site sequences. In the first eight silkmoth chorion genes sequenced, that sequence was PuTCATT (8), where Pu indicates a purine. The cap site sequence in *S. mansoni* chorion genes that have been sequenced to date reads ATCAT.

Recently, Kunz et al. (11) reported the sequences of two genomic fragments coding for a putative eggshell protein of a Liberian strain of *S. mansoni*. A 901-bp *Eco*RI-*Eco*RI fragment is homologous in the coding region with our *S. mansoni* (Puerto Rican strain) genomic clones 6-1-3 and 10-1, except for a 12-bp deletion in this fragment, spanning nucleotides 223 to 234 in Fig. 4. The two sequences are also entirely homologous in the 5′ and 3′ untranslated regions of the mRNA, and they differ only in three nucleotides in the 5′ upstream sequences. A 1,144-bp *Hin*dIII-*Eco*RI fragment showed the same homology with our clones in the coding region as the *Eco*RI-*Eco*RI fragment, but the two diverge completely, starting about 220 bp upstream of the translation start. Since the 5′-flanking sequences of these two fragments are completely different, these must be different gene copies. Presumably, the *Hin*dIII-*Eco*RI fragment represents a gene encoded in the 3.0-kbp *Eco*RI fragment.

Simpson et al. (21) reported a partial nucleotide sequence of an *S. mansoni* (Puerto Rican strain) female-specific cDNA clone, pSF10. The 365-bp sequence, which encompasses the 3′ end of the corresponding mRNA, is homologous with pSMf61-46 except that pSF10 has a deletion of 24 nucleotides, corresponding to nucleotide numbers 469 to 492 of pSMf61-46 (Fig. 4).

As previously estimated by us (3) and substantiated in this work, there are five copies of the eggshell gene in the *S. mansoni* haploid genome. Simpson et al. (21) estimated three copies of this gene per haploid genome. Thus, only a few copies of this gene must facilitate the high levels of protein products necessary for the eggshell synthesis. The very high level of expression of these genes must be controlled by another mechanism than that seen in *D. melanogaster*, since no specific amplification of these genes has been detected in *S. mansoni* (3). Since the accumulation of mRNA (expression of these genes) coincides temporally with the pairing of female schistosomes with males, the induction of these genes may be hormonal or pheromonal.

A nucleotide sequence of a different female-specific cDNA clone (F4) of *S. mansoni* has been reported by Johnson et al. (8). This cDNA represents 640 bp of the 3′ end of a corresponding mRNA estimated to be 1.5 kb long and does not cross-hybridize with pSMf61-46. The deduced amino acid sequence of this clone showed considerable homology with silkmoth chorion proteins and therefore strongly suggested that it also encodes a schistosome eggshell component. This polypeptide has a region very rich in histidine and has much less glycine than the polypeptide encoded by pSMf61-46. This is, therefore, another female-specific gene product, distinct from the products of pSMf61-46 and pSF10. Together, these clones and their gene products may represent the majority of the eggshell protein in *S. mansoni* (8). We have examined the possibility that F4 is closely linked to the eggshell protein genes contained within the six genomic clones described in this study. Using F4 as a hybridization probe in a Southern analysis of six genomic clones, we have not found any cross-hybridizing fragments (unpublished data).

ACKNOWLEDGMENTS

We thank Patricia Cotter for expert technical assistance and Theresa Wnuk for excellent secretary service.

This work was funded by Public Health Service grant R22 AI-18867 from the National Institutes of Health and by the Edna McConnell Clark Foundation. D.M.R. is the recipient of an RCDA (CA-000905) from the National Cancer Institute. This work was conducted as part of the Center for Applied Molecular Biology and Immunology, supported by the State University of New York at Buffalo.

LITERATURE CITED

1. **Bobek, L. A., P. T. LoVerde, H. van Keulen, and D. M. Rekosh.** 1987. A developmentally regulated gene in *Schistosoma mansoni* that may encode an eggshell protein. UCLA Symp. Mol. Cell. Biol. New Series **42**:133–143.
2. **Bobek, L. A., D. M. Rekosh, and P. T. LoVerde.** 1987. Isolation and analysis of adult-female-specific genes from three species of human schistosome parasites. UCLA Symp. Mol. Cell. Biol. New Series **60**:149–158.
3. **Bobek, L. A., D. M. Rekosh, H. van Keulen, and P. T. LoVerde.** 1986. Characterization of female-specific cDNA derived from a developmentally regulated mRNA in the human blood fluke *Schistosoma mansoni*. Proc. Natl. Acad. Sci. USA **83**:5544–5549.
4. **Byram, J. E., and A. W. Senft.** 1979. Structure of the schistosome eggshell: amino acid analysis and incorporation of labelled amino acids. Am. J. Trop. Med. Hyg. **28**:539–547.
5. **Cordingley, J. S.** 1987. Trematode eggshells: novel protein biopolymers. Parasitol. Today **3**:341–344.
6. **Duvall, R. H., and W. D. Dewitt.** 1967. An improved perfusion technique for recovering adult schistosomes from laboratory animals. Am. J. Trop. Med. Hyg. **16**:483–486.
7. **Goldsmith, M. R., and F. C. Kafatos.** 1984. Developmentally regulated genes in silkmoths. Annu. Rev. Genet. **18**:443–487.
8. **Johnson, K. S., D. W. Taylor, and J. S. Cordingley.** 1987. Possible eggshell protein gene from *Schistosoma mansoni*. Mol. Biochem. Parasitol. **22**:89–100.
9. **Jones, C. W., and F. C. Kafatos.** 1980. Structure, organization and evolution of developmentally regulated chorion genes in a silkmoth. Cell **22**:855–867.
10. **Kafatos, F. C.** 1983. Structure, evolution and developmental expression of the chorion multigene families in silkmoths and Drosophila, p. 33–61. *In* S. Subtelny and F. C. Kafatos (ed.), Gene structure and regulation in development. Alan R. Liss, Inc., New York.
11. **Kunz, W., K. Opatz, M. Finken, and P. Symmons.** 1987. Sequences of two genomic fragments containing identical coding region for a putative eggshell precursor protein of *Schistosoma mansoni*. Nucleic Acids Res. **15**:5894.
12. **Lecanidou, R., G. C. Rodakis, T. H. Eickbush, and F. C. Kafatos.** 1986. Evolution of the silk moth chorion gene superfamily: gene families CA and CB. Proc. Natl. Acad. Sci. USA **83**:6514–6518.
13. **Levine, J., and A. Spradling.** 1985. DNA sequencing of a 3.8 kb region controlling Drosophila chorion gene amplification. Chromosoma **92**:136–142.
14. **Mahowald, A. P., and M. P. Kambysellis.** 1980. Oogenesis, p. 141–224. *In* M. Ashbuner and T. R. F. Wright (ed.), The genetics and biology of Drosophila. Academic Press, New York.
15. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
16. **Mitsialis, S. A., and F. C. Kafatos.** 1985. Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature (London) **317**:453–456.
17. **Mitsialis, S. A., N. Spoerel, M. Leviten, and F. C. Kafatos.** 1987.

A short 5'-flanking DNA region is sufficient for developmentally correct expression of moth chorion genes in Drosophila. Proc. Natl. Acad. Sci. USA **84**:7987–7991.

18. **Orr, W., K. Komitopoulou, and F. Kafatos.** 1984. Mutants suppressing in trans chorion gene amplification in Drosophila. Proc. Natl. Acad. Sci. USA **81**:3773–3777.
19. **Osheim, Y. N., and O. L. Miller.** 1983. Novel amplification and transcriptional activity of chorion genes in *Drosophila melanogaster* follicle cells. Cell **33**:543–553.
20. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
21. **Simpson, A. J. G., M. Chaudri, M. Knight, C. Kelly, F. Rumjanek, S. Martin, and S. R. Smithers.** 1987. Characterization of the structure and expression of the gene encoding a major female specific polypeptide of *Schistosoma mansoni*. Mol. Biochem. Parasitol. **22**:169–176.
22. **Spoerel, N., H. T. Nguyen, and F. C. Kafatos.** 1986. Gene regulation and evolution in the chorion locus of *Bombyx mori*. Structural and developmental characterization of four eggshell genes and their flanking DNA regions. J. Mol. Biol. **190**:23–35.
23. **Spradling, A., and A. P. Mahowald.** 1980. Amplification of genes for chorion proteins during oogenesis in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **17**:1096–1100.
24. **van Keulen, H., P. T. LoVerde, L. A. Bobek, and D. M. Rekosh.** 1985. Organization of the ribosomal DNA genes in *Schistosoma mansoni*. Mol. Biochem. Parasitol. **15**:215–230.
25. **Wong, Y. C., J. Pustell, N. Spoerel, and F. C. Kafatos.** 1985. Coding and potential regulatory sequences of a cluster of chorion genes in *Drosophila melanogaster*. Chromosoma **92**: 124–135.