DNA RESEARCH 4, 231–240 (1997)

Evaluation of cDNA Libraries from Different Developmental Stages of *Schistosoma mansoni* for Production of Expressed Sequence Tags (ESTs)

Glória R. FRANCO,[1] Élida M. L. RABELO,[2] Vasco AZEVEDO,[3] Heloisa B. PENA,[1] J. Miguel ORTEGA,[1] Túlio M. SANTOS,[1] Wendell S. F. MEIRA,[1] Neuza A. RODRIGUES,[1] Carlos M. M. DIAS,[2] Richard HARROP,[5] Alan WILSON,[5] Mohamed SABER,[6] Hannan ABDEL-HAMID,[6] Michelyne S. C. FARIA,[7] Maria Elizabeth B. MARGUTTI,[4] Juçara C. PARRA,[7] and Sérgio D. J. PENA[1,*]

Departamento de Bioquímica e Imunologia,[1] Departamento de Parasitologia,[2] Departamento de Biologia Geral [3] and Departamento de Microbiologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil 31270-010,[4] Department of Biology, University of York, York, Y015DD, UK,[5] Theodore Bilharz Research Institute, Cairo, 12411, Egypt[6] and Centro de Pesquisas René Rachou, Belo Horizonte, Brazil 30190-002[7]

(Received 7 April 1997)

Abstract

A comparative study of the gene expression profile in different developmental stages of *Schistosoma mansoni* has been initiated based on the expressed sequence tag (EST) approach. A total of 1401 ESTs were generated from seven different cDNA libraries constructed from four distinct stages of the parasite life cycle. The libraries were first evaluated for their quality for a large-scale cDNA sequencing program. Most of them were shown to have less than 20% useless clones and more than 50% new genes. The redundancy of each library was also analyzed, showing that one adult worm cDNA library was composed of a small number of highly frequent genes. When comparing ESTs from distinct libraries, we could detect that most genes were present only in a single library, but others were expressed in more than one developmental stage and may represent housekeeping genes in the parasite. When considering only once the genes present in more than one library, a total of 466 unique genes were obtained, corresponding to 427 new *S. mansoni* genes. From the total of unique genes, 20.2% were identified based on homology with genes from other organisms, 8.3% matched *S. mansoni* characterized genes and 71.5% represent unknown genes.

Key words: Key Words: *Schistosoma mansoni*; developmental stages; cDNA sequencing analysis; expressed sequence tags

1. Introduction

Schistosoma mansoni (*Sm*) is a digenetic trematode worm responsible for schistosomiasis, a parasitic disease that is estimated to affect at least 300 million people in tropical and subtropical areas of the world (WHO, 1985). Despite intense efforts dedicated to eradicating schistosomiasis through sanitary measures, suppression of the intermediate host and drug treatment, the prevalence of the disease has not decreased. No vaccine is yet available and control of the disease is primarily by chemotherapy. However, reinfection of patients is common and we need new approaches to treatment and prevention, since *S. mansoni* is becoming increasingly resistant to drug therapy. It is hoped that detailed information about the genome of *S. mansoni* might uncover key gene products that may constitute new targets for drug and vaccine development.

Accordingly, in 1992 we started a systematic gene discovery program study in *S. mansoni* using the strategy of partial sequencing of cDNA ends to generate expressed sequence tags (EST).[1] Initially, we utilized an adult worm cDNA library, from which 607 ESTs were

Communicated by Kenichi Matsubara

* To whom correspondence should be addressed. Departamento de Bioquímica e Imunologia, ICB/UFMG. Av. Antônio Carlos, 6627, Belo Horizonte, MG 31270-010, Brazil. Tel. +5531-227-3496, Fax. +5531-227-3792, E-mail: spena@dcc.ufmg.br

† EST sequences were deposited in dbEST and GenBank with the following accession numbers: Adult 1 library (T14340→T14651; T18616→T18626; T24126→T24150; W06712→W06824); Adult 2 library (AA185747→AA185837), Adult 3 library (AA218448→AA218524); Adult 4 library (AA125663→AA169943), Egg library (AA140558→AA140638);Cercariae library (AA143808→AA143896); Lung stage schistosomula library (AA125668→AA125734).

obtained, corresponding to 169 different genes, 15 previously known in *S. mansoni* and 154 new genes.[2] This increased considerably the number of genes identified in the parasite. However, we felt that studying only adult worms was insufficient. *S. mansoni* has a complex life cycle with several morphologically very diverse stages (ova, miracidia, cercariae, schistosomula and adult worms), during which different sets of genes are expressed. Obviously, if one considers the acquisition of information about the worm gene expression in the perspective of designing new drugs and vaccines, the young stages can not be overlooked. Actually, the schistosomula stage is increasingly recognized as one of the main targets for the host immune system.[3]

With this in mind, we planned to extend our EST program to the other life stages of *S. mansoni*. For that, stage-specific cDNA libraries were needed, some of which, unfortunately, are very difficult to construct because of difficulties in obtaining the necessary amounts of pure mRNA. Thus, before embarking on large-scale studies, we decided to evaluate the libraries that were already in existence, comparing them with our original adult worm library. We here report our results with seven different cDNA libraries constructed from four distinct stages of the parasite life cycle, from which a total of 1401 ESTs were generated, totaling 466 different genes, 427 of which are newly describe in *S. mansoni*. From the total of identified genes, we can start to outline a pattern of gene expression, with some genes expressed in a stage-specific manner and others, housekeeping ones, in all developmental stages.

2. Methodology

2.1. Construction of cDNA libraries and sequencing

The following seven cDNA libraries were used in this study: four libraries (Adult 1-4) from adult worms and one library each from ova (Egg), cercariae and lung stage schistosomula (Lung stage). The construction of the Adult 1 cDNA library, plasmidial DNA preparation and sequencing of clones from this library have been previously described.[2] The other six libraries were constructed in λZapII (Stratagene), according to the manufacturer's instructions. Total RNA was isolated from distinct *S. mansoni* developmental stages by the guanidinium thiocyanate-phenol-chloroform method[4] and poly(A)+ RNA was obtained by chromatography on an oligo (dT) column.[5] Double-stranded cDNA was cloned into *Eco*RI/*Xho*I restriction sites of λZapII. pBluescript SK+ phagemids were obtained by "en masse" *in vivo* excision of λZap clones,[6] by co-infecting *Escherichia coli* XL-1 Blue cells with the ExAssist helper phage (Stratagene). The excised phagemids were used to infect *E. coli* SOLR™ cells (Stratagene) for production of double-stranded DNA (dsDNA) templates. Transformants were plated onto LB agar containing ampicillin, X-gal and isopropyl-β-D(-)-thiogalactopyranoside (IPTG). White colonies were selected and grown for 16 hr in 3 ml of Luria broth (LB) supplemented with ampicillin. Aliquots of the cultures (200 μl) were mixed with the same volume of 30% glycerol in LB and frozen at −70°C in 96-well plates. The rest of the cultures were used for plasmidial DNA preparation using the Wizard Plus Mini Prep DNA Purification System (Promega). dsDNA was sequenced by dideoxy chain-termination sequencing[7] using the Thermo-Sequenase Cycle Sequencing kit (Amersham) and M13 Reverse or M13–40 fluorescent-labeled primers (Pharmacia). Single-pass runs of the sequencing reactions were performed on an A.L.F. automated DNA sequencer (Pharmacia).

2.2. Data analysis

Sequences were manually edited to eliminate vector regions, poly(A) tails and lower quality data at the end of the sequence. ESTs containing less than 150 bp and more than 4% ambiguity were rejected. ESTs were compared to DNA and protein sequences deposited in non-redundant databases using the Basic Local Alignment Search Tool (BLAST) programs[8] at the National Center for Biotechnology Information (NCBI). Alignments scoring more than 200 for BLASTN and 100 for BLASTX were selected and after meticulous visual inspection on the biological significance of the alignment, ESTs were named as putative identification for the gene. ESTs with no significant database matches or showing only partial homology with database sequences were grouped as non-identified genes.

2.3. Clustering analysis:

Sequences sharing local similarities were clustered with the ICATOOLS set of programs[9] (freely available at ftp.ebi.ac.uk). Initially, each library was independently analyzed. The module ICAass was used to create an index of clustered sequences (threshold and ktup set to 25 and 8, respectively). One singular sequence was added to the cluster with ICAass and used to run the module ICAtool, under the same threshold and ktup settings. This was followed by the run of ICAtool with all sequences in the library. ICAprint was used to generate the output file, that was manually inspected since some clones had been sequenced in both orientations and/or led to the same identification when submitted to homology search. A second round of analysis was conducted with all libraries concomitantly in order to join the clusters that had been previously formed, but for this purpose only ICAass followed by ICAtool with a singular sequence was executed.

Table 1. Information about the sequencing of different *S. mansoni* cDNA libraries.

	S. mansoni cDNA libraries							
	Egg	Cercariae	Lung stage	Adult 1	Adult 2	Adult 3	Adult4	Total
Number of ESTs	106	110	107	812	94	101	71	1401
Number of sequenced clones	107	107	107	617	94	101	71	1204
Number of usable ESTs[a)]	80	98	67	657	91	78	52	1123
Number of usable clones[a)]	80	98	67	504	91	78	52	970

[a)] ESTs/clones analyzed by ICATOOLS. These numbers correspond to the total number of ESTs/clones after removing sequences of vector, mitochondrial DNA, rRNA and contaminating sequences from other organisms.

3. Results and Discussion

3.1. Quality control of the cDNA libraries

Since the start of *S. mansoni* genome project, one of our main focuses has been the large-scale sequencing of cDNA to produce ESTs, in an attempt to identify new genes of this organism. Initially, we used an adult worm cDNA library, from which we generated 607 ESTs corresponding to 154 new *S. mansoni* genes.[2] The good quality of this library was attested by the diversity of genes that were isolated, even after the discovery of a significant degree of redundancy (65% of the sequenced clones corresponded to 49 redundant genes).[2] The success of this approach prompted us to extend the sequencing program to include other libraries. We started with eight libraries from distinct developmental stages, all of them constructed using the λZap system (Stratagene): one egg, two cercariae (the human-infecting larvae), three adult worms, one 7-day schistosomula (the lung stage) and one from 25-day old worms. All libraries were excised "en masse" and at least 30 colonies from each library were selected to evaluate the average size of the inserts by polymerase chain reaction (PCR). Most of them had an average insert size greater than 500 bp, except for one cercariae and the 25-day worm libraries. Thus, we decided to use all three adult worm cDNA libraries and the Egg, one cercariae and the 7-day schistosomula libraries in this study.

Table 1 summarizes data obtained from the sequencing of the distinct libraries. A total of 1401 ESTs were produced from one or both ends of 1204 clones. The data from the Adult 1 library are cumulative since the beginning of the program and includes ESTs published by Franco et al., 1995.[2] In the Egg library, the number of clones exceeds the number of ESTs and this is due to the sequencing of a chimeric clone from which two ESTs were generated. Both ESTs were eliminated from subsequent analysis. After homology searches in non-redundant databases using BLAST programs[8] and elimination of ESTs corresponding to useless sequences (vector, mitochondrial DNA, rRNA and contaminating sequences from other organisms), 1123 ESTs derived from 970 clones were submitted to clustering analysis, using ICATOOLS program,[9] resulting in a list of distinct genes.

Adams et al.[10] proposed criteria to evaluate the quality of the libraries used in large-scale EST analysis. They state that the sequencing of 100–200 clones from a library is sufficient to assess the quality of this library and to detect problems that might have occurred during library construction. A useful library should contain no more than 20% useless sequences, at least 50% new genes and a broad variety of transcripts. We used their criteria to evaluate the seven cDNA libraries used in this study (Fig. 1). The first five parameters are a measure of the proportion of useless clones. In general, the libraries were of good quality with respect to these parameters, except for the Lung stage, Egg and Adult 3 libraries. The Egg library contains 20% clones without an insert, even though a previous blue/white selection of clones had been performed. The Adult 3 library is enriched in clones corresponding to mitochondrial DNA sequences. Most of them correspond to a polymorphic minisatellite sequence of 620 bp,[11] that contains part of an *S. mansoni* nuclear transcript denominated SM750.[12] This transcript is composed of a invariable region that is followed by five copies of a 62-bp polymorphic repeat element (PRE). Interestingly, five or more copies of the 62-bp PRE were seen solely or as part of the mitochondrial minisatellite in all libraries analyzed except the Egg library. This fact implies that PRE is a very frequent element in the genome of the parasite and that it could be part of a nuclear sequence that was incorporated into the mitochondrial genome.[11] None of the libraries contains excessive number of sequences derived from ribosomal RNA. The Lung stage library contains almost 20% contaminating sequences from other organisms. These contaminating sequences are derived either from *E. coli* or other bacteria, probably due to the contamination of the worm samples during the 7-day period of *in vitro* cultivation necessary to mature to lung stage schistosomula.

The quality of the construction of each library was also analyzed. All of them were shown to be unidirectional (most ESTs had matches to database sequences on the expected strand), composed of a high proportion of inserts longer than 500 bp, composed of inserts with short poly(A) tails and containing no chimeric clones. The only exception was the Egg library, where we found a single

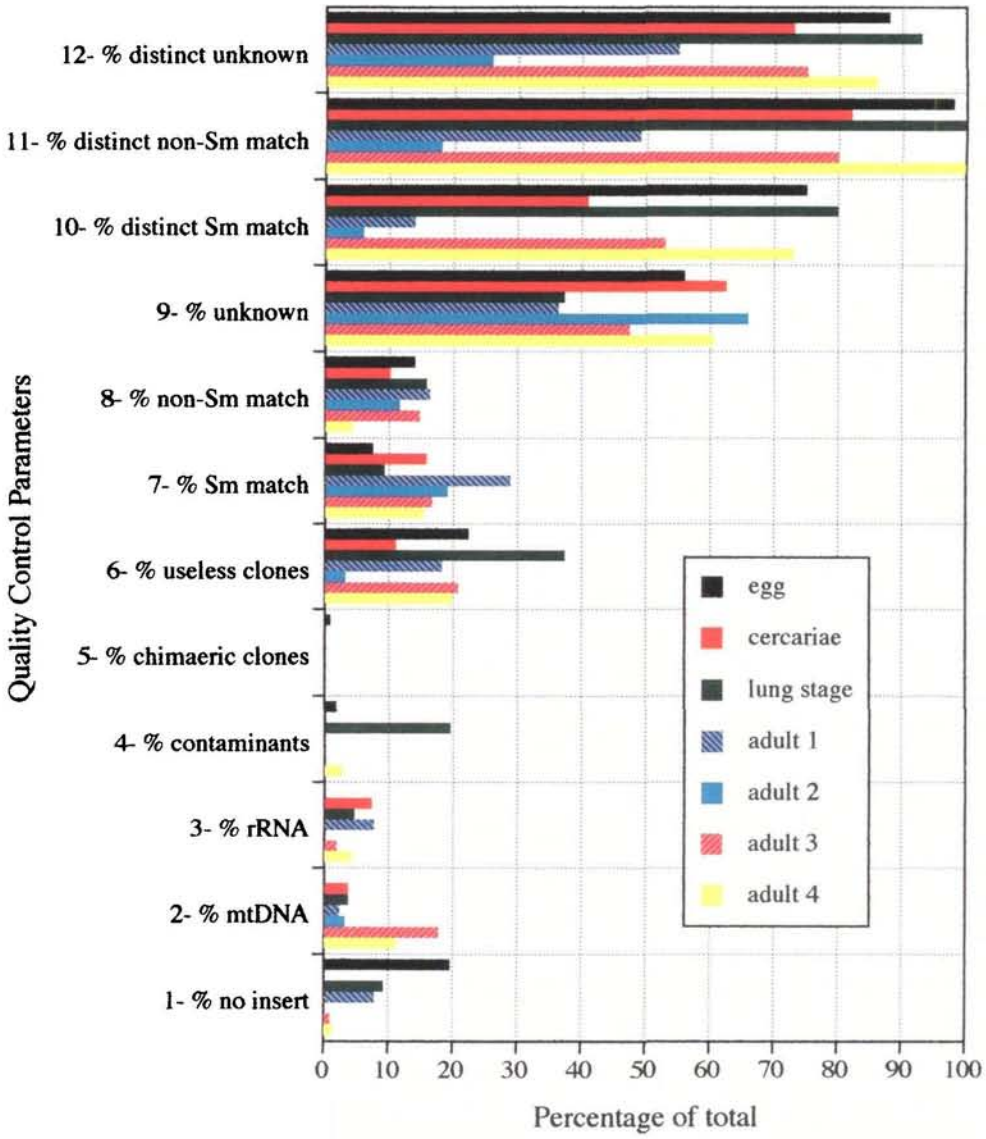

Figure 1. Evaluation of the cDNA libraries according to the criteria of Adams et al.[10] Parameters 1 to 5 indicate the percentage of the total of clones in each library that produced useless ESTs and this set of data is totaled in parameter 6. The percentage of the total of clones that are identified either by homology with previously reported *S. mansoni* genes (*Sm* match), putatively identified by homology with genes from other organisms (non-*Sm* match), or with partial homology with genes from other organisms and non-database match sequences (unknown) is also shown (parameters 7 to 9). The percentage of useful clones that are distinct for each category of genes was determined by clustering analysis and is shown in parameters 10 to 12.

chimeric clone (parameter 5). The sixth parameter is the sum of the first five parameters and totals the frequency of useless clones in each library. Three out of the seven libraries exceed 20% non-useful clones: Lung stage (37%), Egg (22%) and Adult 3 (21%), and this is mainly due to the reasons discussed above. However, when analyzing the gene content in each of these three libraries, we verified that they have a high percentage of distinct genes and a low proportion of redundant genes (see below). This fact justifies the continuation of using of these libraries in the EST sequencing program, but with the inclusion of a previous selection step to eliminate abundant useless clones.

Parameters 7 to 9 of Fig. 1 concern to the analysis of the composition of the libraries after EST homology searches in non-redundant databases. Most libraries showed a low proportion of cDNA clones with exact match to previously described *S. mansoni* genes

Table 2. Gene content of the cDNA libraries after random-sampling of clones.

	S. mansoni cDNA libraries							
	Egg	Cercariae	Lung stage	Adult 1	Adult 2	Adult 3	Adult 4	Total
Distinct genes	73	65	62	198	19	57	48	522
New genes	67	58	54	173	18	48	40	458
% of distinct genes per total of sequenced clones[a)]	68.2	60.7	57.9	32.1	20.2	56.4	67.6	43.4
% of new genes per total of sequenced clones[a)]	62.6	54.2	50.5	28.0	19.1	47.5	56.3	38.0

[a)] for the number of sequenced clones see Table 1.

(less than 20%), except for the Adult 1 library (parameter 7). This can be explained by the fact that this library is enriched in clones corresponding to the *S. mansoni* glycolytic enzyme glyceraldehyde 3-phosphate dehydrogenase (GAPDH),[13] the most redundant gene found in this library. Moreover, as the Adult 1 library was the most sequenced library in this program, it is possible that it better represents the profile of genes expressed in adult worms. Remarkably, all adult worm libraries had, in general, more cDNA matching *S. mansoni* known genes than the libraries constructed from other developmental stages. This is particularly interesting, since it reflects the sort of *S. mansoni* genes that have been deposited in public databases. Most of them are isolated from adult worms. However, the Cercariae library attained the same proportion of clones matching *S. mansoni* genes as the adult libraries. This can be explained by the presence of a very abundant transcript in this category, the calcium-binding protein (CaBP),[14] that corresponds to 10% of the total of useful clones. Most probably this protein is very important for the cercariae metabolism and may be involved in movement. Few clones in all libraries could be putatively identified by significant homology with genes from other organisms (parameter 8) and the great majority of clones in each library (>35%) could not be identified (parameter 9). These last ones correspond to cDNA that had only partial matches to sequences from other organisms or non-database match cDNA.

Parameters 10 to 12 consist of the number of distinct genes divided by the number of useful clones in each category and measure the diversity of transcripts. To obtain the number of distinct genes, each library was submitted to clustering analysis, using the program ICATOOLS. The program grouped together as a single cluster clones with a high degree of identity; each cluster was treated as an independent gene. The veracity of such clusters was attested by the correct grouping of clones that shared the same homology to *S. mansoni* or other organisms database sequences. Considering that one goal of the EST sequencing program is the discovery of new genes, the diversity in the non-*Sm* match and in the unknown categories are particularly relevant. In this respect, in all libraries with exception of the Adult 1 and Adult 2 libraries, more than 70% of the transcripts are distinct in these two categories. This fact counterbalances the low efficiency in obtaining useful clones from the Egg, Lung stage and Adult 3 libraries. An intermediate degree of diversity is observed for the Adult 1 library, while a very low diversity of transcripts is seen in the Adult 2 library. A tendency of decreasing the variety of transcripts in the *Sm* match category is also observed, which can be explained by the presence of very abundant transcripts already characterized in *S. mansoni*. That is the case for the Cercariae, Adult 1 and Adult 2 libraries due to the enrichment of CaBP-, GAPDH- and eggshell protein-encoding[15] transcripts, respectively.

3.2. Gene content and redundancy analysis

The strategy of random-sampling of cDNA libraries always produces a series of clones corresponding to a single transcript; either because abundant mRNA will be more represented in the library, or because each library has an inherent bias that was introduced during its construction. Thus, clones obtained from a such library will reflect its cDNA composition. For this reason, we decided to analyze each library according to its gene content and to evaluate its quality based on the extent of redundancy. This was only possible after performing clustering analysis by ICATOOLS.

Table 2 shows the number of distinct genes, as well as the number of new genes obtained from each library. This last class includes genes homologous to genes from other organisms (non-*Sm* match category) and genes either partially homologous to genes from other organisms or non-database match genes (unknown category). A total of 522 distinct genes were obtained from the seven libraries, 458 of which (88%) were newly identified in *S. mansoni*. This corresponds to three times the number of new genes obtained in the beginning of the sequencing program.[2]

Considering the effort to get distinct or new genes from random selection of clones in each library, it is important to consider the percentage of genes in the total of sequenced clones. This is a measure of the library quality regarding both its redundancy and content of useless clones. It can be seen that, in all libraries with the exception of the Adult 1 and Adult 2 libraries, more than 50% of the sequenced clones were found to be distinct genes. It is important to note that the Adult 1 library was se-

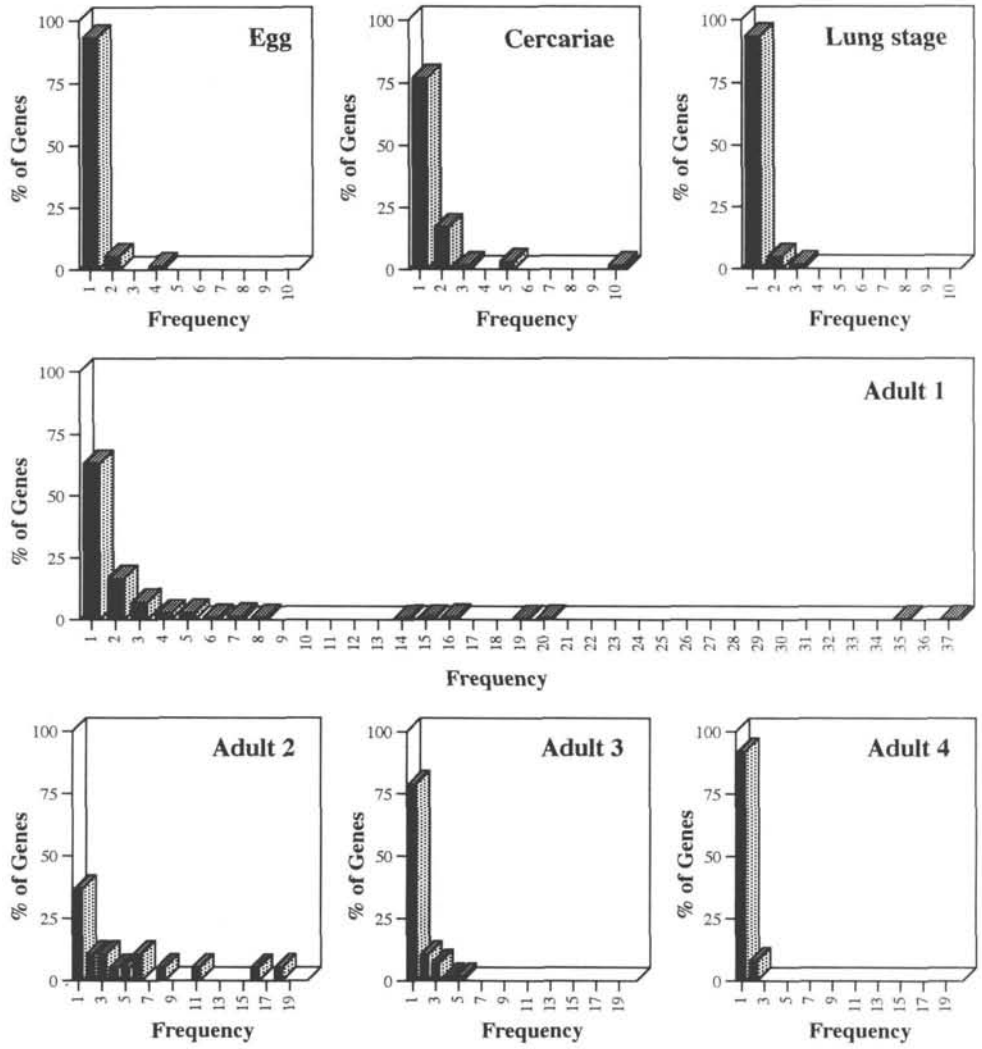

Figure 2. Redundancy in EST sequencing of the *S. mansoni* cDNA libraries. On the abcissa we show the number of times that each gene was sampled and on the ordinate we depict the fraction of genes sharing a given sampling frequency.

quenced close to six times more than the other libraries (Table 1), and this might explain the rate of 32% of new genes. The same tendency was seen for the ratio of new genes per total of sequenced clones. Again, the Adult 1 and Adult 2 libraries provided the lowest efficiencies. Rates of 50% in acquirement of new genes as observed for the *S. mansoni* libraries met the criteria established for the human EST program.[10]

A direct representation of the extent of redundancy in each library is seen in Fig. 2, that shows the percentage of genes that appear in the library under a given frequency. As random sampling of a cDNA library should follow a Poisson distribution for rare events, the unexpected presence of genes under classes of high frequency of isolation reveals a bias in the library. This is evident for the Adult 2 library, where the profile of frequency distribution clearly escapes a typical Poisson distribution, which strongly supports our decision not to use this library for large-scale EST production. The high proportion of redundant genes in this library might have resulted from errors introduced during library construction and amplification, "en masse" excision or clone sampling for EST generation. The occurrence of genes under classes of high frequency of isolation is also seen in the Cercariae and Adult 1 libraries. Nevertheless, it would be possible to eliminate the most redundant genes (8 genes

Table 3. Putatively identified genes homologous to *S. mansoni* genes.[a)]

Gene	Library	EST accession[b)]
Enzyme		
Aspartic proteinase	Adult 4	AA169900
Carbonyl reductase	Egg	AA140589
Cathepsin B	Cercariae	AA143823
Cyclophilin B	Lung stage	AA125705
Enolase	Adult 1	T14396
ER-luminal cysteine protease (ER-60)	Adult 4	AA169915
Fructose-1,6-bisphosphate aldolase	Lung stage	AA125670
Glutathione peroxidase	Cercariae	AA143892
Glutathione S-transferase	Adult 1	T14549
Glyceraldehyde-3-phosphate dehydrogenase	Adult 1	T14434
Hemoglobinase (Sm32)	Adult 1	T14348
Hexokinase	Adult 1	T14603
Triose phosphate isomerase	Egg	AA140583
Cytoskeletal/structural protein		
Actin	Cercariae	AA143846
Alpha-tubulin	Egg	AA140633
Eggshell protein	Adult 4	AA169905
Female-specific polypeptide	Adult 3	AA218489
Myosin heavy chain	Lung satge	AA125688
P48 eggshell protein	Adult 3	AA218479
Sm23 integral membrane protein	Adult 1	T14382
Tropomyosin (GB:SCMTPM)	Adult 1	W06805
Tropomyosin (GB:SCMTROPO)	Adult 1	W06761
Antigen		
Antigen 10-3	Adult 1	T14386
Antigen Sm21.7	Adult 3	AA218508
Major egg Antigen (P40)	Egg	AA140559
Sm13 tegumental antigen	Adult 4	AA169901
Transport/storage protein		
Calcium binding protein (CABP)	Cercariae	AA143886
Calcium-calmodulin binding protein	Cercariae	AA143883
Calreticulin	Adult 1	W06720
Fatty-acid binding protein (Sm14)	Adult 1	T14374
Ferritin	Adult 3	AA218482
Glucose transporter	Adult 1	T14364
Other		
Breast basic conserved protein/ ribosomal protein L13	Adult 1	T14585
Calnexin homolog SmIrV1	Adult 3	AA218511
Elongation factor 1 alpha	Lung stage	AA125724
Heat shock protein 86	Adult 1	T14407
S. mansoni mRNA for tandem repeat	Egg	AA140585
S. mansoni (Liberia) zinc finger protein	Lung stage	AA125700
Y-box-binding protein	Adult 1	T14571

[a)] Genes putatively identified by homology with *S. mansoni* database sequences. Only one representative EST matching the respective gene is shown, together with the name of the library it was isolated from.
[b)] EST accession corresponds to the GenBank accession number.

for the Adult 1 library and 3 genes for the Cercariae library) from these libraries by filter screening, using the abundant transcripts as probes, and this should result in a profile compatible with a Poisson distribution.

Although some libraries presented problems detected by quality analysis, they all contributed to the list of putatively identified genes, as well as 333 distinct unknown genes (see below). Genes identified by homology with previously described *S. mansoni* genes are distributed amongst various classes, such as genes coding for enzymes, structural proteins, antigens, proteins involved in transport and storage, etc (Table 3). Two genes in this list have been the subject of a more extensive study and were characterized in detail in our laboratory. They are the *S. mansoni* homologues of the Y-box-binding protein (Franco et al., submitted) and the breast basic conserved protein, or the 60S ribosomal protein L13 (Franco et al., submitted). Table 4 lists distinct genes putatively identified by homology with genes from other organisms. They code for enzymes of different metabolic pathways, a great variety of ribosomal proteins, several constituents of transcriptional/translational machinery, and regulatory cytoplasmic and membrane proteins, among others. Three of these genes were selected for further studies. One is the homologue of mago nashi gene from *Drosophila*. This gene is necessary for proper germ plasm assembly and mutations in it result in sterility of F1 progeny.[16] The *S. mansoni* purine nucleoside phosphorylase was selected for presenting a high similarity with the human counterpart, the 3D structure of which has already been resolved and deposited in the Protein Data Bank. Modeling studies with this protein have led to the identification of powerful inhibitors of this enzyme, whose activity is crucial in T cell guanosine metabolism.[17] The third gene is the homologue of the human HLA-DR-associated protein I, a protein which may be involved in signal transduction in B cells.[18] We are interested in the selection of proteins that can interact with it, which may help to define its biological function in the parasite.

3.3. Gene expression profile in S. mansoni

To obtain an initial profile of gene diversity in the parasite and a preliminary pattern of gene expression in distinct stages of the development of *S. mansoni*, we performed a clustering analysis, joining sequences from all libraries. This resulted in a total of 466 unique genes (considering only once the genes present in more than one library), corresponding to 427 new *S. mansoni* genes. From the total of unique genes, 39 (8.3%) matched previously characterized *S. mansoni* genes, 94 (20.2%) matched genes from other organisms and 333 (71.5%) represent unknown genes. From the clustering analysis, most genes (433 of 466) were present only in a single library (e.g. CaBP was found only in the Cercariae library). Other genes were expressed in more than one developmental stage and are listed in Table 5. They may represent housekeeping genes in the parasite and, curiously, ten of them were unknown. The antigenic potential of such genes should be investigated, since they might be specific to this parasite.

At this point of the sequencing program, only three genes were found to be expressed in all developmental stages analyzed: the cytochrome oxidase chain I, the fructose-1,6-bisphosphate aldolase and unknown gene 10. Somewhat unexpectedly, actin and GAPDH, the most frequent genes in the collection, were not isolated from all stages, perhaps because the number of transcripts sequenced in each library was not very large. Five genes

Table 4. Identified genes homologous to non-*S. mansoni* genes.[a)]

Gene	Library	EST accession[b)]
Enzymes		
Alcohol dehydrogenase class III	Adult 3	AA218449
Aldehyde dehydrogenase	Adult 1	T24129
Aldose reductase	Adult 1	W06782
ATP synthase, vacuolar	Adult 1	T24142
Cytochrome Oxidase chain I	Egg	AA140564
Cytocrome oxidase II	Adult 3	AA218486
Dakt1 serine/threonine protein kinase	Adult 4	AA169931
Dihydrolipoamide acetyltransferase	Adult 1	W06795
Enoyl-CoA Hydratase	Lung stage	AA125690
Glutamine Synthetase	Adult 1	W06821
Glycerol 3-phosphate dehydrogenase	Cercariae	AA143842
H+-transporting ATP synthase alpha-chain	Adult 1	W06794
Lactate dehydrogenase	Adult 1	W06744
Oligosaccharyl transferase 48 KD	Adult 3	AA218463
Ornithine aminotransferase	Adult 1	T14588
Phosphoenolpyruvate Carboxykinase	Egg	AA140576
Phosphoglycerate kinase	Adult 1	T14620
Phosphoglycerate mutase	Adult 1	W06743
20S proteasoma subunit RC7-I=PRE1 homolog	Lung stage	AA125733
Proteasome zeta chain	Adult 1	T14568
Purine nucleoside phosphorylase	Adult 1	W06714
Pyruvato kinase	Adult 1	T24140
Ribonuclease- phosphate 3-epimerase (pentose-5-phosphate 3-epimerase)	Adult 3	AA218494
Vacuolar ATP synthase subunit B	Adult 1	W06824
Transcriptional/ Translational Machinery		
40S ribosomal protein S3	Adult 3	AA218471
40S ribosomal protein S4	Adult 1	W06814
40S ribosomal protein S7	Egg	AA140626
40S ribosomal protein S11	Adult 4	AA169892
40S ribosomal protein S12	Lung stage	AA125707
40S ribosomal protein S14	Adult 1	T14564
40S ribosomal protein S17	Lung stage	AA125727
40S ribosomal protein S20	Egg	AA140581
40S ribosomal protein S21	Egg	AA140582
40S ribosomal protein S26	Lung stage	AA125695
60S ribosomal protein L5	Lung stage	AA125687
60S ribosomal protein L7	Egg	AA140600
60S ribosomal protein L7a	Adult 1	T14431
60S ribosomal protein L10a	Egg	AA140612
60S ribosomal protein L25	Lung stage	AA125723
60S ribosomal protein L30	Adult 3	AA218468
Asp-tRNA synthetase	Adult 1	W06768
Elongation factor 1 gamma	Adult 1	T14422
Homo sapiens 9G8 splicing factor	Adult 1	W06725
Jun-binding protein	Adult 4	AA125664
Lys-tRNA synthetase	Adult 1	T14484
Polyadenylate binding protein	Adult 1	W06727
Putative transcriptional regulator	Lung stage	AA125694
Reverse transcriptase	Egg	AA140605
Rho-GDP dissociation inhibitor	Adult 1	W06723
RNA polymerase II subunit	Lung stage	AA125704
RNA-binding protein X-16	Adult 1	T14358
Small nuclear ribonucleoprotein	Adult 1	T14459

Table 4. Continued.

Gene	Library	EST accession[b)]
Membrane/cytoplasm		
ADP/ATP carrier protein	Adult 1	T14447
Annexin family	Adult 1	T14511
Beta-1 tubulin	Egg	AA140634
Chaperonin-like protein	Adult 1	T14632
Cytochrome *c*	Cercariae	AA143872
DNAJ homolog	Adult 1	W06722
GTP-binding protein	Adult 3	AA218450
Heat shock protein 108	Adult 1	T18621
HLA-DR associated protein I	Adult 3	AA218460
Polyubiquitin	Egg	AA140632
Possible membrane protein	Lung stage	AA125728
Protein kinase C inhibitor protein	Adult 1	T14595
Nonerythroid alpha-spectrin	Adult 1	T14622
UDP-galactose translocador	Adult 3	AA218519
Other		
52k active chromatin boundary protein	Adult 1	W06740
Alpha-collagen	Adult 1	T14493
Apoptosis-inducible	Cercariae	AA143880
Arginine-rich gene	Adult 1	T14555
C. elegans hypothetical 272 KD protein *C50C3.6 in chromosome III*	Adult 3	AA218465
C. elegans clone C16C10.10	Adult 1	W06746
Coded for by *C. elegans* cDNA	Adult 3	AA218495
Cysteine-rich intestinal protein	Adult 3	AA218481
E. coli hypothetical 53.1 KD protein in LYSU-CADA intergenic region	Lung stage	AA125683
Fibrillin 2	Cercariae	AA143891
GATA-3 gene	Egg	AA140598
Golden Syrian Hamster repetitive DNA	Egg	AA140590
Histone H3.3	Cercariae	AA143814
H. sapiens mRNA for *Sm* protein F	Lung stage	AA125673
Human Alu subfamily	Adult 1	W06750
Hypothetical protein - *D. melanogaster*	Cercariae	AA143820
Hypothetical protein 5 *Xanthobacter* sp	Adult 1	W06771
Hypothetical 30.5 KD protein of *C. elegans*	Egg	AA140628
Liver regeneration factor augmenter	Adult 2	AA185826
Mago nashi protein	Lung stage	AA125719
MER5 Protein	Adult 1	T14649
NIFS-like 54.5 KD protein	Adult 1	W06757
Proliferation-associated protein	Lung stage	AA125729
Retrovirus-related GAG polyprotein	Egg	AA140602
Synaptophysin	Adult 1	W06818
Yeast hypothetical 103.7 KD Protein	Adult 1	W06819
Valosin-containing protein homologue	Adult 1	T14640

[a)] Genes putatively identified by homology with genes from other organisms. Only one representative EST matching the respective gene is shown, together with the name of the library it was isolated from.
[b)] EST accession corresponds to GenBank accession number.

were present in two or more adult libraries, but absent in other stages. This is the case for the eggshell protein, that is recognized to be expressed in mature females, and also unknown gene 2.

Clustering analysis also included formation of contigs of sequences. As an example, the cDNA sequence of an unknown gene, that is abundant in the Adult 1 library, was obtained after assembling ESTs from both cDNA ends that were clustered together by ICATOOLS. This gene is currently being characterized in more detail. We expect that, with the advance of the sequencing program, a higher number of partial cDNA sequences will be assembled as full-length contigs, increasing the ability to identify unknown genes and more precisely define the real number of distinct genes in each library and in each developmental stage.

Acknowledgments: The authors thank Kátia Barroso for carrying out automated DNA sequencing. This investigation received financial support from the following sources: PADCT, CNPq, UNDP/ WORLD BANK/WHO Special Program for Research and Training in Tropical Diseases (TDR N°: 940325 and 940751), USAID/HOH (N° 264.01.01.04), FAPEMIG, PAPES/ FIOCRUZ.

Table 5. Frequence of genes present in multiple *S. mansoni* cDNA libraries.[a)]

Genes	Egg (%)	Cercariae (%)	Lung stage (%)	Adult 1 (%)	Adult 2 (%)	Adult 3 (%)	Adult 4 (%)	Total (%)
1- Actin	—	1.0	—	6.9	—	3.9	1.9	4.1
2- Alpha tubulin	2.5	—	1.5	0.8	—	2.6	—	0.9
3- ATP synthase	—	—	1.5	0.2	—	—	—	0.2
4- Beta tubulin	1.3	—	—	0.4	—	—	—	0.3
5- Carbonyl reductase	1.3	—	—	0.2	—	—	—	0.2
6- Cathepsin	—	1.0	1.5	0.4	—	—	—	0.4
7- Cyclophilin B	—	—	1.5	—	—	2.6	—	0.3
8- Cysteine-rich intestinal protein	—	1.0	—	—	—	1.3	—	0.2
9- Cytochrome oxidase chain I	1.3	2.0	1.5	0.4	8.8	—	—	1.4
10- EF1alpha	—	—	1.5	3.0	—	—	—	1.6
11- Eggshell protein	—	—	—	0.2	19.8	—	1.9	2.1
12- Enolase	—	—	—	0.8	—	—	1.9	0.5
13- ER-luminal cysteine protease	—	—	—	0.4	—	—	1.9	0.3
14- Fibrillin	—	1.0	—	0.6	—	—	—	0.4
15- Fructose-1,6-BP aldolase	1.3	1.0	1.5	3.2	—	—	3.8	2.2
16- GAPDH	—	2.0	4.5	7.3	—	1.3	—	4.4
17- Major egg Antigen (P40)	1.3	—	—	0.2	—	—	—	0.2
18- Myosin heavy chain	—	—	1.5	0.4	—	—	—	0.3
19- Oligosaccharyl transferase 48 KD	—	—	—	0.2	—	1.3	—	0.2
20- Triose phosphate isomerase	1.3	—	—	0.4	—	—	—	0.3
21- Ubiquitin	1.3	—	—	1.0	—	—	—	0.6
22- 60S ribosomal protein L5	—	—	1.5	1.0	—	—	—	0.6
23- 60S ribosomal protein L30	—	—	1.5	—	—	1.3	—	0.2
24- Gene 1[b)]	—	2.0	—	—	—	3.8	—	0.5
25- Gene 2	—	—	—	0.4	2.2	5.1	—	0.8
26- Gene 3	—	—	1.5	—	—	1.3	—	0.2
27- Gene 4	—	—	1.5	0.2	—	—	—	0.2
28- Gene 5	—	1.0	—	0.2	—	—	—	0.2
29- Gene 6	—	1.0	—	0.2	—	—	—	0.2
30- Gene 7	—	—	1.5	—	—	—	1.9	0.2
31- Gene 8	—	—	1.5	0.2	—	—	—	0.2
32- Gene 9	1.3	2.0	—	0.2	4.4	6.4	1.9	1.4
33- Gene 10	1.3	5.1	1.5	1.6	—	1.3	1.9	1.8

[a)] Percentage of clones matching the corresponding gene in the total of usable clones analyzed by ICATOOLS. For the total of usable clones see Table 1. [b)] unknown genes are numbered 1–10.

References

1. Adams, M. D., Kelley, J. M., Gocayne, J. D. et al. 1991, Complementary DNA sequencing: expressed sequence tags and human genome project, *Science*, **252**, 1651–1656.
2. Franco, G. R., Adams, M. D., Soares, M. B., Simpson, A. J. G., Venter, J. C., and Pena, S. D. J. 1995, Sequencing and Identification of expressed *Schistosoma mansoni* genes by random selection of cDNA clones from a directional library, *Gene*, **152**, 141–147.
3. Smithers, S. and Terry, R. J. 1965, The infection of laboratory hosts with cercarial of *S. mansoni* and the recovery of adult worms, *Parasitology*, **55**, 695–700.
4. Chomczynski, P. and Sacchi, N. 1987, Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction, *Anal. Biochem.*, **162**, 156–159.
5. Aviv, H. and Leder, P. 1972, Purification of biologically active globin messenger RNA by chromatography on oligo-thymidylic acid-cellulose, *Proc. Natl. Acad. Sci. USA*, **69**, 1408.
6. Short, J. M., Fernandez, J. M., Sorge, J. A., and Huse, W. D. 1988, λZAP: A bacteriophage λ expression vector with *in vivo* excision properties, *Nucleic Acids Res.*, **16**, 7583–7600.
7. Sanger, F. 1981, Determination of nucleotide sequences in DNA, *Science*, **214**, 1205–1210.
8. Altschul, S. F., Gish, W. Miller, W. Myers, E. W., and Lipman, D. 1990, Basic local alignment search tool, *J. Molec. Biol.*, **215**, 403–410.
9. Parsons, J. D., Brenner, S., and Bishop, M. J. 1992, Clustering cDNA sequences, *Comput. Appl. Biosci.*, **8**, 461–466.
10. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D. et al. 1995, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, *Nature*, **377** (supp), 3–174.
11. Pena, H. B., Souza, C. P., Simpson, A. J. G., and Pena, S. D. J. 1995, Intracellular promiscuity in *Schistosoma mansoni*: nuclear transcribed DNA sequences are part of a mitochondrial minisatellite region, *Proc. Natl. Acad. Sci. USA*, **92**, 915–919.
12. Spotila, L. D., Rekosh, D. M., and LoVerde, P. T. 1991, Polymorphic repeated DNA element in the genome of *Schistosoma mansoni*, *Mol. Biochem. Parasitol.*, **48**,

117–120.

13. Goudot-Crouzel, V., Caillol, D., Djabali, M., and Dessein, A. J. 1989, The major parasite surface antigen associated with human resistance to schistosomiasis is a 37 kDa glyceraldehyde-3P-dehydrogenase, *J. Exp. Med.*, **170**, 2065–2080.
14. Ram, D., Grossman, Z., Markovics, A. et al. 1989, Rapid changes in the expression of a gene encoding a calcium-binding protein in *Schistosoma mansoni*, *Mol. Biochem. Parasitol.*, **34**, 167–175.
15. Menrath, M., Michel, A., and Kunz, W. 1995, A female-specific sequence of *Schistosoma mansoni* encoding a mucin-like protein that is expressed in the epithelial cells of the reproductive duct, *Parasitology*, **111**, 477–483.
16. Boswell, R. E., Prout, M. E., and Steichen, J. C. 1991, Mutations in a newly identified Drosophila melanogaster gene, mago nashi, disrupt germ cell formation and result in the formation of mirror-image symmetrical double abdomen embryos, *Development*, **113**, 373–384.
17. Ealick, S. E., Babu, Y. S., Bugg, C. E. et al. 1991, Application of the crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors, *Proc. Natl. Acad. Sci. USA*, **88**, 11540–11544.
18. Vaesen, M., Barnikol-Watanable, S., Gotz, H. et al. 1994, Purification and characterization of two putative HLA class II associatedd proteins: PHAPI and PHAPII, *Biol. Chem. Hoppe-Seyler*, **375**, 113–126.