Update of the Gene Discovery Program in *Schistosoma mansoni* with the Expressed Sequence Tag Approach

Élida ML Rabelo, Glória R Franco*, Vasco AC Azevedo, Heloisa B Pena*, Túlio M Santos*, Wendell SF Meira*, Neuza A Rodrigues*, José Miguel Ortega*, Sérgio DJ Pena*/+**

Departamento de Parasitologia *Departamento de Bioquímica e Imunologia **Departamento de Biologia Geral, ICB, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brasil

Continuing the Schistosoma mansoni *Genome Project 363 new templates were sequenced generating 205 more ESTs corresponding to 91 genes. Seventy four of these genes (81%) had not previously been described in* S. mansoni. *Among the newly discovered genes there are several of significant biological interest such as synaptophysin, NIFs-like and rho-GDP dissociation inhibitor.*

Key words: *Schistosoma mansoni* - genome project - expressed sequence tags

The study of the *Schistosoma mansoni* genome has a high priority in the effort to understand the multiple facets of this complex parasite. Accordingly, a *S. mansoni* genome project was started in 1992 as a Brazilian initiative. From the original cDNA library used, 607 ESTs were generated (Franco et al. 1995a,b) leading to the identification of 154 new genes. This increased considerably the number of known genes in *S. mansoni*.

The *S. mansoni* genome project was adopted by the World Health Organization (WHO) and has been partially funded by this institution in a collaborative international project with the aim of undertaking a co-ordinated gene discovery program in *S. mansoni*. We wish to report the sequencing of 363 further templates by our laboratory, with identification of 91 genes, 74 of which have not been previously described in *S. mansoni*.

MATERIALS AND METHODS

Plasmidial DNA preparation, sequencing and analysis of the sequences were done essentially as described by Franco et al. (1995a).

RESULTS

Tables I an II summarise the data found in the 363 templates sequenced, which produced 205 useful ESTs derived from 185 clones. Some clones were sequenced in both directions using the forward and reverse primers and thus producing two ESTs for the same gene. That is the reason why there are 205 ESTs for 185 clones. All sequences were submitted to homology searches in DNA and protein databanks showing that the 205 ESTs corresponded to 91 different genes, 81% of which had not been previously described in the parasite. The sequences were grouped according to the following criteria: sequences presenting homology to previous identified *S. mansoni* genes (Tables II and III), sequences identified by homology with other organisms (Tables II and IV) and sequences with no homology with any gene deposited in database banks (Table II). Sequences showing low homology with other organisms were identified as partial matches (Tables II and V).

TABLE I

Information about sequencing of the *Schistosoma mansoni* cDNA library

Number of sequenced templates	363
Number of ESTs	205
Number of clones	185
Number of different genes	91
Average EST size	372nt
Average polyA tail length	23nt

TABLE II

EST categories of Sm cDNA library

	No. of clones	(%)
Putatively identified		
Sm match	47	22.7
Non Sm match	62	35.7
Not identified		
Non Sm partial match	4	2.2
Non database match	57	31.3
Mitochondrial	0	0
rRNA	12	6.5
Vector without insert	3	1.6
Total	185	100.0

This investigation received financial support from the following institutions: UNDP/WORLD BANK/WHO Special Programme for Research in Tropical Diseases (TDR) (ID no. 940325, 940751); FAPEMIG Process no. CBS 1190/95; CNPq and PADCT.

+Corresponding author. Fax: +55-31-441.5409

Received 16 April 1997

Accepted 30 June 1997

TABLE III

Characterization of ESTs: database match of ESTs to Sm genes

EST name	dbEST Acces	Homology with (accession)	Identity (%)	Length (nt)	Number of clones	Program	Score	Probability
SMPBE13F	W06737	alpha-tubulin (GB:S98950)	98.5	274	3	FASTA	1068	
SMPBD26F	W06720	Calreticulin (GB:SCMCALRET)	98.3	460	1	BLASTN	2230	5.7e-183
SMPBE11F	W06736	Fructose 1,6 bisphosphate aldolase (GP:SCMALDO_1)	98.0	432	7	FASTA	1156	
SMPBG05F	W06784	GAPDH (GB:SCMGAPDH)	99.6	252	11	FASTA	989	
SMPBE62Fwazzu	W06751	GST (GB:SCMANT28K)	98.6	291	9	FASTA	1148	
SMPBF02F	W06765	HSP 86 (GB:SCMHSP86)	99.0	387	1	FASTA	1532	
SMPBG92F	W06807	Myosin heavy chain (GB:SCMMYH)	99.6	262	2	BLASTN	1301	1.2e-100
SMPBF01F	W06764	ER -luminal cysteine protease ER 60 gene (GB:SMF1PSPCX)	97.9	380	1	BLASTN	1830	9.9e-145
SMPBE24F	W06741	Sm750 gene (GB:SCMSM750)	96.2	240	2	FASTA	902	
SMPBE51F	W06747	Triose phosphate isomerase (GB:SCMSGTPI05)	100.0	256	2	BLASTN	1280	1.6e-101
SMPBE92F	W06761	Tropomyosin (GB:SCMTROPO)	99.7	369	2	BLASTN	1836	1.2e-145
SMPBE03F	W06734	Actin (GB: M80334)	96.6	304	14	BLASTN	1138	3.4e-87
SMPBF30F	W06772	Fibrillin 2 mRNA	89.2	337	1	BLASTN	673	1.7e-99
SMPBG29F	W06788	Elongation factor 1-alpha (EMB:SMF1ALPH)	97.5	405	3	BLASTN	1902	5.5e-151
SMPBD40R	W06733	Y-box binding protein (GB: U398831)	97.8	358	10	BLASTN	1536	4.9e-125
SMPBE59F	W06749	BBC-1 (GB:U57003)	98.9	373	2	BLASTN	1400	6.8e-147
SMPBG90F	W06805	Tropomyosin (GB:SCMTPM)	100.0	200	1	BLASTN	1000	9.0e-76

TABLE IV

Characterization of ESTs: database match of EST to non-Sm genes

EST name	dbEST	Homology with (accession)	Similary (%)	Identity (%)	Length (aa)	Number of clones	Program	Score	Probability
SMPBF59F	W06782	Aldose reductase(PDB:1DLA)	72.5	53.8	80	1	BLASTX	219	1.4e-23
SMPBF07F	W06768	Asp-tRNA synthetase (SP:SYD_CAEEL)	85.8	74.2	155	1	BLASTX	625	1.0e-80
SMPBE38F	W06746	*C. elegans clone* C16C10.10 (GP:CEC16C10_4)	75.0	57.4	68	1	BLASTX	210	2.3e-22
SMPBG74F	W06795	Dihydrolipoamide Acetyltransferase (GP:RATPDCE2_1)	83.3	72.9	48	2	BLASTX	192	5.1e-39
SMPBD30R	W06722	DNAJ homolog (PIR:S42031)	79.2	69.8	53	2	BLASTX	194	4.1e-38
SMPBE17F	W06739	Enolase (SP:ENO_SCHJA)	95.9	87.6	73	2	BLASTX	348	1.6e-41
SMPBE97R	W06821	Glutamine Synthetase (SP:GLNA_HUMAN)	78.6	66.0	103	1	BLASTX	413	5.0e-51
SMPBG67F	W06794	H+-transporting ATP synthase alpha-chain (PIR:S14516)	80.2	64.8	91	1	BLASTX	303	3.3e-37
SMPBD33R	W06725	Homo sapiens 9G8 splicing factor (GP:HUM9G8SF_1)	70.7	63.8	58	2	BLASTX	184	4.9e-37
SMPBE61F	W06750	Human Alu subfamily (SP:ALU7_HUMAN)	76.7	67.4	43	1	BLASTX	154	1.2e-14
SMPBF15F	W06771	Hypothetical protein 5 Xanthobacter sp (PIR:S47055)	60.0	44.0	125	1	BLASTX	271	2.6e-31
SMPBE28F	W06744	Lactate dehydrogenase (GP:MUSLDHB_1)	75.4	45.9	61	3	BLASTX	158	1.8e-15
SMPBE82F	W06757	NIFS-like 54.5KD protein (SP:NFS1_YEAST)	65.6	53.6	125	1	BLASTX	328	1.2e-38
SMPBE27F	W06743	Phosphoglycerate mutase (SP:PMG1_ECOLI)	1 77.6	67.2	67	1	BLASTX	239	1.3e-26
SMPBD34R	W06727	Polyadenylate binding protein (DBJ:HUMPOLYABP)	76.3	74.8	127	1	BLASTX	400	2.3e-23
SMPBD23F	W06714	Purine nucleoside phosphorylase (PDB:1ULA)	35.5	55.5	45	2	BLASTX	76	2.9e-15
SMPBF65F	W06783	Ribosomal protein L5 (SP:RL5A_XENLA)	73.4	52.1	94	3	BLASTX	252	5.0e-28
SMPBH15F	W06814	Ribosomal protein S4 (SP:RS4_HUMAN)	81.1	72.6	107	1	BLASTX	441	3.5e-56
SMPBD32R	W06723	rho-GDP dissociation inhibitor (GP:MUSGDPDI_1)	75.0	52.1	48	1	BLASTX	138	4.4e-19
SMPBE16R	W06818	Synaptophysin(PIR:A60548)	60.5	39.5	43	1	BLASTX	87	9.3e-08
SMPBE57F	W06748	Tubulin beta chain (PIR:S18457)	98.1	94.4	107	1	BLASTX	543	6.8e-70
SMPBF13F	W06770	Polyubiquitin (X60390)	99.0	99.0	105	2	BLASTX	517	6.1e-67
SMPBF52R	W06824	Vacuolar ATP synthase - subunit B(SP:VAT_DROME)	97.0	94.0	68	1	BLASTX	335	1.4e-39
SMPBE65R	W06819	Yeast hypothetical 103.7KD Protein (SP:YBM7_YEAST)	74.1	51.9	27	1	BLASTX	71	4.4e-11

TABLE V

Characterization of ESTs: database partial match of EST to non-Sm genes

EST No	dbEST	Homology with (accession)	Similarity (%)	Identity (%)	Length aa	nt	No. of clones	Program	Score	Probability
SMPBF32F	W06774	14 ORF YJR83.9 gene product [*S. cerevisiae*](GP:IX87611)	65.4	47.3	26		1	BLASTX	126	3.2e-12
SMPBG91F	W06806	RNA binding protein (PIR:S53050)	45.6	43.4	46		2	BLASTX	95	3.1e-08
SMPBH26F	W06816	*D. discoideum* plasmid Ddp2 trans-acting factor gene (GB:DDIDDP2)		68.1		62	1	BLASTN	194	4.7e-6

DISCUSSION

Tables IV and V show that some of the genes with homology either to *S. mansoni* or with other organism were sequenced more than twice. This is the case of GAPDH (11 times selected from the library), GST (9 X), frutose 1,6 biphosphate aldolase (7 X), actin (14 X) and Y-box binding protein (10 X) showing a certain degree of redundance of the library as had been previously reported (Franco et al. 1995a). Adams et al. (1995) defined some parameters to define whether a library has sufficient quality for the purpose of generating ESTs. These parameters include the proportions of: vectors without inserts, contaminant of the library with others cDNAs (host or bacteria), presence of mitochondrial DNA or rRNA, number of new genes, number of genes matching other organism genes and number of genes with homology to the original organism *(S. mansoni* in our case). Although some redundancy was found, the library is still considered of very good quality, especially when we take in account the fact that 81% of the genes identified were new. Thus, we believe that it is worth pursuing further work with this library for the generation of new ESTs.

Among the new genes identified by homology with other organisms several stand out for having significant biological interest. Thus, an EST with homology to synaptophysin gene was found. In mammals, synaptophysin is one of the major integral membrane proteins of synaptic vesicles (McMahon et al. 1996) it is speculated that synaptophysin may function as a gap junction-like pore or channel (Calakos & Scheller 1994). Another homology that called our attention is the one with a NIFS-like protein from yeast. The NIFS-like protein is supposed to be involved in both tRNA-processing and mitochondrial metabolism (Kolman & Soll 1993), two interesting targets for design of new drugs.

The EST approach is also contributing for adding new members into gene families. Members of the family of GDP dissociation inhibitors (GDI) for the ras-related rho-subtype proteins appear to take part in the regulation of a number of biological processes, including cell growth and differentiation. We have identified an EST with a high similarity (75%) with a murine D4 cDNA, a new member of the GDP-GDI family (Adra et al. 1993).

From these initial identifications, one cannot ascertain to the *S. mansoni* genes the same functions of those founds to the homologue genes in other organisms. However, these identifications open new avenues to further characterise these genes and through functional studies obtain a correlation between gene function and homology with the most diverse organisms.

Through the sequencing of this adult cDNA library, a great number of new genes were identified in *S. mansoni,* showing the high efficiency of the EST approach. However, this parasite presents a complex life cycle with enormous changes in its morphology. Obviously, one would expect that such great changes are accompanied by changes at the gene expression level. Furthermore, if one considers the acquisition of information about the worm gene expression in the perspective of designing new drugs and/or vaccines, the young stages cannot be overlooked. Actually, the schistosomula stage is recognized as the main target to the host immune system attack (Smithers & Terry 1965). With the foregoing in mind it is our future aim to study the expression pattern of the different life stages of *S. mansoni* by sequencing cDNA libraries for the distinct stages.

ACKNOWLEDGEMENTS

To Kátia Barroso for carrying out automated DNA sequencing.

REFERENCES

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JA, Brandon RC, Chiu MW, Clayton RA, Cline RT, Cotton MD, Earle-Hughes J, Fine LD, FitzGerald LM, FitzHugh WM, Fritchman JL, Geoghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle Jr. PS, Kelley JC, Klimek KM, Kelley JC, Liu LI, Marmaros SM, Merrick JM, Moreno-Palanques RF, McDonald LA, Nguyen DT, Pellegrino SM, Phillips CA, Ryder SE, Scott JL, Saudek DM, Shirley R, Small KV, Spriggs TA, Utterback TR, Weidman JF, Li Y, Barthlow R, Bednarik DP, Cao L, Cepeda MA, Coleman TA, Collins EJ, Dimke D, Feng P, Ferrie E, Fischer C, Hastings GA, He WW, Hu JS, Huddleston KA, Greene JM, Gruber J, Hudson P, Kim A, Kozak DL, Kunsch C, Ji H, Li H, Meissner PS, Olsen H, Raymond L, Wei YF, Wing J, Xu C, Yu GL, Ruben SM, Dillon PJ, Fannon MR, Rosen CA, Haseltine WA, Fields C, Fraser CM, Venter CJ 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature 377* (Suppl). 3-174.

Adra CN, Leonard D, Wirth LJ, Cerione RA, Lim B 1993. Identification of a novel protein with GDP dissociation inhibitor activity for ras-like proteins CDC4Hs and rac I. *Gens Chromosom Cancer 8*: 253-261.

Calakos N, Scheller RH 1994. Vesicle-associated membrane protein and synaptophysin are associated on the synaptic vesicle. *J Biol Chem 269*: 24534-24537.

Franco GR, Adams MD, Soares MB, Simpson AJG, Venter JC, Pena SDJ 1995a. Identification of new *Schistosoma mansoni* genes by EST strategy using a directional cDNA library. *Gene 152*: 141-147.

Franco GR, Simpson AJG, Pena SDJ 1995b. Sequencing and identification of expressed *Schistosoma mansoni* genes by random selection of cDNA clones from a direction library. *Mem Inst Oswaldo Cruz 90*: 215-216.

Kolman C, Soll D 1993. SPL1-1, a *Saccharomyces cerevisiae* mutation affecting tRNA splicing. *J Bacteriol 175*: 1433-1442.

McMahon HT, Bolshakov VY, Janz R, Hammer RE, Siegelbaum SA, Sudhof TC 1996. Synaptophysin, a major synaptic vesicle protein, is not essential for neurotransmitter release. *Proc Natl Acad Sci USA 93*: 4760-4764.

Smithers S, Terry RJ 1965. The infection of laboratory hosts with cercarial of *S. mansoni* and the recovery of adult worms. *Parasitology 55*: 695-700.