Sequencing and Identification of Expressed *Schistosoma mansoni* Genes by Random Selection of cDNA Clones from a Directional Library

Glória R Franco, Andrew JG Simpson*, Sérgio DJ Pena

Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas - UFMG, Avenida Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brasil *Laboratório de Biologia Molecular, Centro de Pesquisas "René Rachou"-FIOCRUZ, Avenida Augusto de Lima 1715, 30190-002 Belo Horizonte, MG, Brasil

We have initiated a gene discovery program in Schistosoma mansoni *based on the technique of Expressed Sequence Tags (ESTs), i.e. partial sequences of cDNAs obtained from single passes in automatic DNA sequencers. ESTs can be used to identify genes on the basis of their homology with sequences from other species deposited in DNA or protein databases. Transcripts with sequences without matches in the databases may represent novel parasite-specific genes. This approach has shown to be very efficient and in less than two years a broad range of novel genes has already been ascertained, more than doubling the number of known* S. mansoni *genes.*

Key words: *Schistosoma mansoni* - automated DNA sequencing - EST - gene identification

Although a decade has passed since the first *Schistosoma mansoni* gene was cloned, relatively little information is currently available concerning its genome and its gene complement. Indeed, GenBank contains less than 100 different full-length cDNA or genome sequences from schistosomes (GenBank 1993). In the context of the global effort to systematically map and sequence the human genome (Olson 1993), a program that embraces the study of a number of model organisms such as yeast, *Drosophila* and *Caenorhabditis elegans*, we have thus embarked on a program to characterize the schistosome genome and its component genes in depth. Our view is that since current efforts at understanding the biology of the organism, drug design and vaccine production are being undertaken with meager molecular knowledge, a gene discovery program may contribute importantly to these areas.

The collaboration involves the Federal University of Minas Gerais, the Fundação Oswaldo Cruz (Brazil) and The Institute for Genomic Research (U.S.A.) and is focused on defining the expressed gene content of *S. mansoni* expressed by producing Expressed Sequence Tags (ESTs). ESTs are single pass sequences of at least 150 bp derived from one or both ends of cDNAs which we obtain using automated DNA sequencers. ESTs provide an "identity tag of the gene" that can be used for the identification of the respective gene based on its homology with database sequences derived from the same or other organisms in the DNA data bases (Adams et al. 1991). The cDNAs from which the ESTs are derived then become important tools for selecting and identifying full length copies of the gene within DNA libraries and for mapping the position of the gene within the organisms genome.

An adult worm size-selected cDNA library was constructed and cloned into BA vector (a phagemid derived from pEMBL)[•], essentially as described (Adams et al. 1993). The ESTs were obtained by partial sequencing from one or both ends of randomly selected cDNAs using M13 universal, M13-40 or M13 reverse primers, in automated DNA sequencers (ALF Pharmacia or 373A Applied Biosystem). Homology searches were done at National Center for Biotechnology Information site using the program Basic Local Alignment Search Tool (Altschul et al. 1990) for nucleotide sequences (BLASTN) or amino acid sequences (BLASTX). The genes were identified based on the high score alignments between the EST and the database sequences. The cDNAs were classified as "identified genes" when they showed matches with schistosome sequences or with those of other organism in the databases, or "non-identified genes" when they showed only partial homology with non-schistosome genes or had no significant database match.

We obtained a total of 536 ESTs from 390 cDNA clones. From these clones, 21.1% corresponded to undesirable sequences such as rRNA, mitochondrial sequences and plasmids without inserts. The remaining were grouped as follows: (1) 36.4% with no database match which may correspond to untranslated regions of the cDNAs or genes expressed only in *S. mansoni*; (2) 5.1% with a partial match with non-schistosome sequences and whose alignments consisted either of short re-

gions probably corresponding to either conserved motif or domains or regions enriched with a specific amino acid; (3) 15.9% homologous to previously sequenced schistosome genes; and (4) 20.5% showing homology with genes from other organisms. The first two groups consisted of "non-identified genes" and the two last groups of "identified genes". The ESTs with no database match were compared one with other in order to see if they correspond to the same or different genes. The corresponding cDNAs were divided into classes of distinct genes, some of them with high frequency in the library and others representing medium abundance or low abundance transcripts. The ESTs collected to date represent a broad range of transcripts including enzymes and known antigens as well as cytoplasmic, structural, regulatory, membrane, nuclear and secretory proteins.

Using the EST strategy we have partially sequenced 168 different genes from the cDNA library. Fourteen of them represent genes previously characterized in *S. mansoni* and 154 correspond to new genes of this organism that have never sequenced before. From these, 42 were identified based on matches with non-schistosome entries in the databases, 20 were not precisely identified although they showed partial homology with sequences of different organisms and 92 were not identified, because they did not match any entry in the databases.

This EST project was initiated in 1992 and has already more than doubled the number of defined *S. mansoni* genes. It represents the first initiative of a systematic study of the parasite's genome and expressed gene complement. In addition, to increasing the *S. mansoni* EST data base we are currently embarking on the full length sequencing of selected genes and the construction of a genetic map using the cDNAs as probes.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science 252:* 1651-1656.

Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC 1993. Rapid cDNA sequencing (Expressed Sequence Tags) from a directionally cloned human infant brain cDNA library. *Nature Genet 4:* 373-389.

Altschul SF, Gish W, Miller W, Myers EW, Lipman D 1990. Basic local alignment search tool. *J Molec Biol 215:* 403-410.

GenBank 1993, Release 80, December 10.Olson MV 1993. The human genome project. *Proc Natl Acad Sci USA 90:* 4338-4344.