RESEARCH NOTE

The Schistosome Genome Project: RNA Arbitrarily Primed- PCR Allows the Accelerated Generation of Expressed Sequence Tags

Emmanuel Dias Neto+, Richard Harrop **, Rodrigo Corrêa-Oliveira, Sérgio DJ Pena *, R Alan Wilson **, Andrew JG Simpson***

Centro de Pesquisas René Rachou - FIOCRUZ, Av. Augusto de Lima 1715, 30190-002 Belo Horizonte, MG Brasil *Departamento de Bioquímica e Imunologia, ICB-UFMG, Caixa Postal 486, 30161-970 Belo Horizonte, MG, Brasil **Department of Biology, University of York, Heslington, York, Y015DD, UK ***Laboratório de Genética de Câncer, Instituto Ludwig de Pesquisas Contra o Câncer, R. Prof. Antônio Prudente 109/ 4º andar, 01509-010 São Paulo, SP, Brasil

Key words: *Schistosoma mansoni* - RNA arbitrarily primed - PCR - Schistosome Genome Project - expressed sequence tags - carboxypeptidase - reverse transcriptase

In 1992, a collaborative research project was established between Fundação Oswaldo Cruz (Brazil), the Federal University of Minas Gerais (Brazil) and The Institute for Genomic Research (USA) which aimed to generate Expressed Sequenced Tags (ESTs) from the parasitic trematode *Schistosoma mansoni.* Initially, a directional cDNA library was used and ESTs produced by sequencing one, or both ends of the randomly selected clones. Such an approach generated 154 sequences of different genes which had not been described previously in *S. mansoni* (GR Franco et al. 1995a *Gene 152*: 141-147, b *Mem Inst Oswaldo Cruz 90*: 215-216). This represented a greater than twofold increase in the number of sequences available in the databases before the project was initiated. The adoption of such a strategy proved to be extremely valuable in the analysis of abundant gene transcripts as well as in the discovery of new genes. However, there are some problems inherent to this method, such as the sequencing of vectors without insert and more importantly, the high level of redundancy. For example, of the 429 clones sequenced, 46 (10.7%) represented vectors without insert, and of 202 identified ESTs, homology with only 77 different genes was found, indicating a redundancy of 62% (Franco et al. 1995a *loc. cit.*). Whilst the former problem can be overcome, there is no simple way to solve the latter. These factors reduce the number of useful sequences obtained and thus increase the costs and the time required to tag all of the genes expressed by the parasite. In addition, to tag cDNAs derived from either rare, tissue or stage-specific mRNAs requires complex pre-processing of libraries, such as normalization, subtraction or differential hybridization (C Hoog 1991 *Nucl Acids Res 19*: 6123-6127).

With the aim of increasing the overall efficiency of EST generation, we have applied a technique of RNA arbitrarily primed PCR (RAP-PCR) to this process. RAP-PCR and its close relative "Differential Display", were first described by J Welsh et al. (1992 *Nucl Acids Res 20*: 7213-7218), and P Liang and AB Pardee (1992 *Science 257*: 967-971), respectively. Both techniques enable the identification of differentially expressed genes. We have adapted RAP-PCR to generate normalized plasmid mini-libraries which were subsequently sequenced.

Parasite material was pelleted, washed with DEPC treated water and stored at -70°C until required. mRNA was then extracted using the Micro-Fast Track™ mRNA isolation kit (Invitrogen, San Diego, CA) according to the manufacturer's instructions. The mRNA was eluted with 200 µl of a solution of 10mM Tris-HCl pH 7.5 in DEPC treated water and stored in aliquots of 10 µl at -70°C until required for cDNA synthesis. Genomic DNA was extracted from cercariae and adult worms as described by THDA Vidigal et al. (1994 *Exp Parasitol 79*: 187-194). Prior to cDNA synthesis, an aliquot of 10 µl of mRNA (1 to 5 ng) was heated to 65°C for 10 min and subsequently mixed with 2 nmol of dNTPs, 100 units of M-MLV reverse transcriptase (Promega Co.), 25 pmol of a randomly choosen primer between those available in our laboratories, in a buffer consisting of 25 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$ and 10 mM DTT in a final reaction volume of 20 µl. The mixture was incubated at 37°C for 30 min and the resulting cDNA stored at -70°C until required. The cDNA (usually 1.0 µl) was mixed with 200 µM of dNTPs, 6.4 pmol of the same or a different primer,

This research was funded by grants from WHO/OMS, PAPES-FIOCRUZ and CNPq.
+Corresponding author. Fax: 55-31-295.3115, e-mail: emmanuel@gene.dbbm.fiocruz.br.
Received 18 January 1996
Accepted 8 May 1996

0.5 units of Taq DNA polymerase (Cenbiot, RS, Brazil) in a buffer consisting of 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris.HCl pH8.3, in a final volume of 10 µl. The reaction was covered with mineral oil, and the samples were submitted to one cycle through a denaturation step of 1 min at 95°C, annealing at 37°C for 2 min and primer extension for 2 min at 72°C, followed by 34 cycles of amplification as follows: 95°C for 45 sec, 55°C annealing for 1 min, and 72°C for 90 sec. Genomic DNA used as a control, was submitted to the same conditions. The reaction products (3µl) were resolved by electrophoresis on 6% polyacrylamide gels and visualized after silver staining (CJ Sanguinetti et al. 1994 *Biotech 17*: 915-918), instead of using radioisotopes and sequencing gels, commonly employed by other authors. The profiles are reproducible even when different amounts of cDNA were used, and the cDNA amplification profile is totally different from the profile derived from genomic DNA amplification. The pooled bands derived from different primer combinations, were concentrated and ligated into the pUC 18 plasmid, using the SureClone™ Kit (Pharmacia Co.). The ligated products were then used to transform competent DH5α bacteria according to J Sambrook et al. (1989 *Molecular cloning: A laboratory manual* 2nd ed). Successfully transformed cells were selected on LB medium plates containing X-gal, IPTG and 150 µg/ml of ampicilin. The insert size of each recombinant plasmid was determined by transferring a small quantity of individual white bacterial colonies to a 0.5ml eppendorf tube using a sterile toothpick. The DNA insert was amplified using 5 pmol primers flanking the cloning site (pUCF and pUCR) in the presence of 0.5 units of Taq DNA polymerase, 200 µM of dNTPs, and the same buffer described above. The reaction was covered with a drop of mineral oil and submitted to 25 cycles through the following temperature profile: denaturation at 95°C for 1 min, annealing at 55°C for 1 min and extension at 72°C for 1 min. The final extension step was 5 min. To select clones containing inserts of different sizes, PCR products derived from the different bacterial colonies were electrophoresed in 4% polycrylamide gels which were subsequently silver-stained. Only those clones containing inserts of differing sizes were selected for further study. The selected clones were grown overnight in 5 ml of LB containing 150µg/ml ampicilin at 37°C on a shaking platform (250rpm), and the supercoiled plasmids were prepared from the colonies using a modified protocol of the Wizard minipreps™ (Promega Co). Cycle sequencing reactions were performed according to the manufactors instructions (Autocycle™ sequencing kit - Pharmacia, Upsala, Sweeden). The sequencing was undertaken using the automatic sequencers available in the Laboratory of Molecular Biology at the Centro de Pesquisas René Rachou (ALF - Pharmacia), or in the Cancer Research Unit, of the Department of Biology of the York University (ABI). The obtained sequences were initially edited in order to subtract both vector and primer sequences. The edited sequences were then submitted for analysis to the servers available at the National Center for Biotechnology Information, National Institutes of Health, USA and European Molecular Biology Laboratories, Heidelberg, Germany. Searches for homologies at the amino acid and/or nucleotide level used the algorithms of BLAST (SF Astchul et al. 1990 *J Mol Biol 215*: 403-410) and FASTA (WR Pearson & DJ Lipman 1988 *Proc Natl Acad Sci USA 185*: 2444-2448). To consider a gene as putatively identified, we used cutoff values of scores higher than 100 and a P value less than 0.05 or a P value less than 10^{-5}, independent of the score.

Using this RAP-EST strategy 185 ESTs were obtained. Of these, none represented vectors without insert, illustrating the value of the PCR screening prior to the selection of the clones for sequencing. Of the 185 sequences, 1.2% represented genes which had already been sequenced in *S. mansoni*, 6.9% represent genes with partial *S. mansoni* matches, 18.6% show homology to genes sequenced in other organisms and 70% had no sequence homology to anything present in GenBank. The overall level of redundancy using this method was only 2%. All the ESTs produced are available in the public databases. We were able to highlight 2 ESTs which showed homology to non-schistosome genes, a carboxypeptidase and a reverse transcriptase. Neither gene has been described before in *S. mansoni*.

The *S. mansoni* carboxypeptidase homologue is a clone of 303 bp (accession number - L46953) which has similarity of 52.6% in an overlap of 57 aminoacids and 63.8% at the nucleotide level with a human carboxypeptidase precursor. The aminoacid alignment (Fig.) shows the conserved regions which enabled the identification of the gene, and the active sites are indicated with asterisks (D Hendriks et al. 1993 *Biol Chem Hoppe-Seyler 374*: 843-849). The enzyme is responsible for the hydrolysis of the carboxyl-terminal peptide bond in the polypeptide chain. Such enzymes are involved in peptide hormone maturation. The physiological function of some carboxypeptidases, however, appears to be to protect the organism from the action of potent peptide hormones that may escape from the tissues or be released into the cir-

```
                *                   *
SM)DNHFNDGITNGARWYSLNGGMQDXNYLHTNSFXITLELGCXKFPNASXLPRYWNESKMSAE
HM)  NFPNGVTNGYSWYPLQGGMQDYNYIWAQCFEITLELSCCKYPREEKLPSFWNNNKASPD
HH)  DYFPDGITNGASWYSLSKGMQDFNYLHTNCFEITLELSCDKFPPEEELQREWPDDKLFQK
AH)DSSFKDGITNGGAWYSVPGGMQDFNYLSSNCFEITLELSCDKFPNEDTLKTYWEQNRNSPD
HE)DSSFVDGTTNGGAWYSVPGGMQDFNYLSSNCFEITVELSCEKFPPEETLKTYWEDNKNSPD
```

Aminoacid residues alignment of the carboxypeptidase derived from: SM - *Schistosoma mansoni* clone SMRAP040 (L46953); HM: human carboxypeptidase M precursor (P=1.6e-24); HN: human carboxypeptidase N precursor (P=1.7e-23); AH: *Lophius americanus* carboxipeptidase H (P=4.2e-23); HE: human carboxypeptidase E (P=8.2e-22). The aminoacids residues conserved in all carboxypeptidases are shown in bold, and those present in at least another carboxypeptidase are underlined. ***** indicates components of the active site (Hendriks et al. 1993 *Biol Chem Hoppe-Seyler 374:* 843-849).

culation (Hendriks et al. 1993 *loc. cit.*). The role of the carboxypeptidase in *S. mansoni* has yet to be established.

The *S. mansoni* reverse transcriptase homologue is a clone of 216 bp (accession number - L46976). The aminoacid alignment shows a similarity of approximately 66% over a 60 amino acid overlap with various reverse transcriptases. When compared with the reverse transcriptases of three other organisms (*Caenorhabditis elegans, Anopheles gambiae* and *Gallus gallus*), 63.6% of the 60 aminoacids are conserved, in average, in the same position of this same region of the molecule. If this enzyme is still active it could be functioning as a telomerase, or could be an indication of retroviral infection. The presence of a reverse transcriptase in *S. mansoni* was suggested before (M Tanaka et al. 1989 *Parasitol 79*: 187-194), and Franco et al. (1995a *loc. cit.*), have sequenced a highly expressed EST that showed homology with a retrovirus-related pol polyprotein (GenBank accession code - T18625), confirming the presence of viral nucleic acids incorporated into the parasite genome. A comparison made using the FASTA program showed no significant similarities between our clone (L46976) and that reported in Franco et al. (1995a *loc. cit.*).

The Genome Project represents a valuable resource for researchers. The composition of a cDNA library reflects the abundance of mRNA sub-populations. Thus, if there is no pre-screening of these libraries, there will be a strong bias due to the abundance of some messages. On the other hand, when the RAP-PCR technique is used, the sub-populations screened are strongly biased by the primer sequence, with the populations being selected, that have a maximum of homology between primer and template. To change the bias, all that is required is to change the primer combination to produce the cDNA or to perform the PCR. The redundancy of these minilibraries was found to be low, and the number of new genes in schistosome, was higher than described before, enabling a much faster development of the project.

The sequences reported here are deposited in the GenBank, under the accession numbers L46914 to L47098. The clones are available upon request to the first author

Acknowledgments: to Diana Noronha Nunes, Mike Anderson, Nilton Barnabé Rodrigues and Ricardo Pereira de Moura for their valuable technical assistance. To Neusa Araújo and Cecília P de Souza for providing the parasites used in this study.