ABSTRACT

Optimized construction of low-redundancy cDNA mini-libraries using low-stringency RT-PCR is described. cDNAs are generated using arbitrary consensus-degenerate hybrid oligonucleotide primers and nanogram amounts of Schistosoma mansoni mRNA. A number of conditions such as temperature and salt concentration are combined to create balanced, low-stringency conditions that permit a normalized amplification of a diversified, random set of sequences. On average, 350 different sequences are obtained per mini-library, which represents a significantly higher diversity of messages per library when compared to previously published conditions (i.e., 20–40 sequences/mini-library). The optimized high-throughput approach described here is likely to help in the discovery of expressed genes in any complex organism.

INTRODUCTION

As the sequence of the human genome is being completed (8,14), it is becoming apparent that the identification of expressed genes in a complex genome is a major challenge. Indeed, the identification of genes can only be made with confidence by mapping expressed sequence tag (EST) sequences and full-length mRNA sequences onto the genomic sequence (8,14). Exclusive in silico prediction of exons and introns on a genomic sequence that exhibits large coding segments and a large number of repetitive elements is subject to enormous error rates. This brings renewed importance to the acquisition of large sets of EST sequences obtained from mRNA extracted from different tissues and at different stages of the cell life cycle to cover the complete transcriptome of an organism. In fact, despite the large set of more than 3.7 million EST sequences currently available from human tissues, it is apparent that only a fraction of the human gene complement has its full sequence determined. In this respect, a new low-stringency RT-PCR EST (ORESTES) sequencing strategy was shown to produce a complex pattern (15) and was subsequently used to produce EST libraries (3,4). However, the primers that are used do not include amino acid codon variability, and, typically, a restricted collection of sequences has been obtained per mini-library: about 20 different sequences for each mini-library with Schistosoma mansoni mRNA (4) or about 40 different sequences per mini-library with human mRNA (3).

In the present work, we have chosen to optimize low-stringency RT-PCR using partially degenerate oligonucleotide primers to increase the diversity of the messages obtained, thus facilitating EST library construction and high-throughput sequencing. Special care was taken to ensure that the advantages of the low-stringency RT-PCR technique were preserved, namely the amplification of the central portion of messages (3), the amplification of less abundant messages (i.e., normalization of libraries) (3), and the ability to use nanogram amounts of mRNA that permits the characterization of messages present in small numbers of cells (15).

S. mansoni is a digenetic blood fluke that causes schistosomiasis in humans and is a major cause of morbidity in the world. S. mansoni has a very large (270 Mbases) and complex genome (11) in which several families of repeats have been identified and characterized (5,12), thus imposing considerable difficulties in terms of the amount of work and financial support required for full genome sequencing and gene identification. Therefore, we decided to generate almost 3000 S. mansoni ESTs in the present work of optimizing modified ORESTES as an attempt to contribute further toward finding genes in this complex organism.
MATERIALS AND METHODS

Sequences were deposited into GenBank® (accession nos. BG930137–BG932910). S. mansoni live adult worms were collected from freshly perfused hamsters, extensively washed in ice-cold PBS, and immediately frozen in dry ice. Total RNA was isolated from approximately 500 mg tissue by resuspending in TRIzol® (Invitrogen, Carlsbad, CA, USA) and processing according to the manufacturer’s instructions. The mRNA was isolated with an Oligotex mRNA isolation kit (Qiagen, Valencia, CA, USA) and resuspended in 40 μL buffer. Alternatively, 30 mg tissue were used for extraction with the microMACS mRNA isolation kit (Miltenyi Biotec) and eluted with 120 μL RNase-free water. A total of 2–5 μg mRNA was obtained in both cases. Subsequently, mRNA samples were treated with RQ1 RNase-free DNase (1 U/10 μg) for 15 min at 37°C, DNase was inactivated at 65°C for 10 min. Reverse transcription was carried out for 1 h at 42°C using 200 U SUPERSCRIPT II™ reverse transcriptase (Invitrogen) in 20 μL reaction medium containing 250 ng mRNA and 20 pmol of the desired primer. Primer APYAF1 has the following sequence: 5'-CGGAATTCCTGTTTAAATATGGTATTGTGAN-GAYKSNGGNTC-3' and was designed according to the principles indicated by Rose et al. (10). It also had been designed previously in our laboratory to match one of the conserved domains of the family of apyrases or nucleoside triphosphate diphosphohydrolases (13). The primer has a 512-fold degeneracy, with degenerate bases at positions 3, 6, and 12 and partially degenerate bases at positions 7, 8, and 9 starting from the 3'-end. In addition, it has an EcoRI site at the 5'-end to facilitate the cloning of the resulting PCR products. The oligonucleotide primer, Hib8, has a degeneracy of 512 and the following sequence: 5'-CGGAATTCCTGTTTAAATATGGTATTGTGAN-GAYKSNGGNAAT-3'. The transcriptase was inactivated by incubation at 70°C for 15 min. The sample was treated for 20 min at 37°C with 2 U RNase H, and 0.5-μL aliquots of the transcriptase reaction mixture were subjected to PCR under the conditions described below, using 50 μL PCR medium, 2.5 U AmpliTaq Gold® DNA polymerase (Applied Biosystems, Foster City, CA, USA), and 100 pmol of the same primer used in the reverse transcription step. Under standard low-stringency conditions (3,4), the reaction medium contained undiluted (1×) PCR buffer, and PCR was carried out as follows: an initial step of 94°C for 8 min, an annealing step at 37°C for 2 min, and an extension step at 72°C for 2 min, followed by 34 cycles each at 95°C for 45 s, 55°C for 1 min, and 72°C for 90 s. Under our modified low-stringency conditions, the reaction medium contained either undiluted (1×) buffer or PCR buffer that was diluted to 0.7x or 0.4x, as indicated. In the latter two cases, MgCl2 was added for a final concentration of 2.5 mM. Touchdown PCR was performed as follows: an initial step at 94°C for 8 min, followed by 30 cycles each of denaturation at 94°C for 1 min, annealing at variable temperatures for 1 min, and extension at 72°C for 1 min. In the first cycle, the annealing temperature was set to 60°C, and for each of the 29 subsequent cycles the annealing temperature was decreased by 0.5°C (i.e., it varied from 60°C to 45.5°C at 0.5°C decrements along the 30 cycles). This was followed by 10 cycles each at 94°C for 1 min, 45°C for 1 min, and 72°C for 1 min. A GeneAmp® PCR System 9700 (Applied Biosystems) was used in the maximum ramp speed mode unless specifically stated otherwise. When a slow temperature ramp was used, the speed of the temperature increase in the annealing to extension step was set to 15% of its maximum speed. Human cDNA was generated with primer APYAF1 and the touchdown PCR strategy using mRNA from breast tumor tissue that was kindly provided by Dr. A. Simpson from the Human Cancer Genome Project (HCGP, http://www.ludwig.org.br/ORESTES/). A total of 138 clones from the resulting modified ORESTES library were sequenced and analyzed by BLAST against the UniGene public databank. Controls for the absence of contaminating DNA were done by performing PCR with an aliquot of mRNA that had not been subjected to treatment with reverse transcriptase. This is a particularly important and critical control for the method because the low-stringency and random arbitrary priming conditions used in the PCR are likely to amplify any contaminating DNA that is present in the mRNA preparations.

An aliquot of the reaction was analyzed by SDS-PAGE and stained with silver, and the remaining reaction was digested with EcoRI. Digested cDNA was size-selected on agarose gels, cloned, and sequenced by the dideoxy-terminator method using BigDye™ Terminator chemistry and an ABI Prism® 377 automated sequencer (both from Applied Biosystems).

RESULTS AND DISCUSSION

Effect of Touchdown PCR on Diversity of Amplified Messages

We decided to start with the standard low-stringency RT-PCR protocol conditions for ORESTES described by Dias-Neto et al. (3,4) and to introduce a degenerate primer as the first attempt to increase the diversity of amplified cDNA. The degenerate primer APYAF1 was arbitrarily chosen as the first candidate.

Figure 1A, lane LS, shows that RT-PCR performed under the standard low-stringency RT-PCR conditions (3,4) results in a very limited amplification with an almost undetectable amount of cDNA and even more limited amounts of products greater than or equal to 500 bases. In contrast, Figure 1A, lane LS+TD, shows the same reaction, except that touchdown PCR was introduced during the annealing temperature steps. PCR began with an annealing temperature at 60°C. At each of the subsequent 30 cycles, the annealing temperature was reduced by 0.5°C, down to 45°C. Figure 1A, lane LS+TD, shows that products between 250 and 600 bases were obtained. However, certain products were amplified preferentially, resulting in the appearance of defined bands in the gel.

Dilution of Salts

Different reaction conditions for touchdown PCR were further tested to
change the pattern of amplification and possibly to obtain a more diverse set of products. The experiments illustrated in Figure 1B show how diluting the buffer solution used during touchdown PCR affects the pattern of amplification. Except for MgCl₂, whose concentration was kept constant, all salts and reagents present in the PCR buffer were diluted by the extent indicated in the figure. Figure 1B, lane 0.7X, shows that a dilution of buffer to 0.7× produced a smear of larger products (600–700 bases) when compared to the 1× buffer. The further dilution of the buffer to 0.4× yielded no amplification.

It has been reported (7) that a dilution of the PCR buffer to 0.8× increases the efficiency of the amplification of larger PCR products in multiplex PCR. A similar dilution of the buffer to 0.7× is also effective in increasing the size of PCR products in low-stringency touchdown PCR with degenerate primers (Figure 1B, lane 0.7X).

Slow Temperature Ramp in PCR

Figure 1C shows an experiment similar to that of Figure 1B, with the introduction of a slow ramp temperature increase between the annealing and extension steps in the reaction (temperature ramp at 15% of the maximum). It has recently been described (6) that the use of a slow temperature ramp step increases the efficiency of PCR amplification with arbitrary primers. A considerable increase in the complexity of the amplification profile can be seen (Figure 1C), even when using undiluted PCR buffer, the amplification profile being much more diverse than that obtained under similar conditions with no temperature ramp (compare with Figure 1, B and C). The effect of the slow temperature ramp was very pronounced when used in combination with the dilution of buffer to either 0.7× or 0.4× (Figure 1C). Note that a combination of buffer dilution to 0.4× and a slow temperature ramp gave the best cDNA amplification profile with the most diverse product sizes, some of them 2–3 kb.

Table 1. Classification of Sequences from Libraries F24I and F33 by Similarity to GenBank Sequences

<table>
<thead>
<tr>
<th></th>
<th>F24I library</th>
<th>F33 library</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of total sequences</td>
<td>401</td>
<td>642</td>
</tr>
<tr>
<td>No. of unique sequences</td>
<td>219</td>
<td>415</td>
</tr>
<tr>
<td>Nonredundancy</td>
<td>55%</td>
<td>65%</td>
</tr>
<tr>
<td>Partially or totally known sequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. mansoni ESTs</td>
<td>71.4%</td>
<td>49.8%</td>
</tr>
<tr>
<td>S. mansoni full-length sequences</td>
<td>32.9%</td>
<td>33.3%</td>
</tr>
<tr>
<td>(rRNA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New sequences</td>
<td>28.6%</td>
<td>50.2%</td>
</tr>
<tr>
<td>similar to S. japonicum ESTs</td>
<td>1.7%</td>
<td>0.3%</td>
</tr>
<tr>
<td>similar to non-Schistosoma sequences</td>
<td>10.9%</td>
<td>13.1%</td>
</tr>
<tr>
<td>no GenBank match</td>
<td>15.7%</td>
<td>36.7%</td>
</tr>
</tbody>
</table>

aNew sequences refer to our library sequences that did not exhibit similarity to any S. mansoni sequence available in GenBank. These were separated into three categories according to the best hit found against other species (threshold E value less than 10⁻¹⁰).

Figure 1. Optimization of low-stringency reverse transcription and PCR. S. mansoni cDNA was analyzed on silver-stained 6% polyacrylamide gels as shown here. The scale at left is in the number of bases. (A) Primer APYAF1 was used under low-stringency (LS) conditions with a touchdown (TD) program (as described in the text) and compared with standard low-stringency conditions of the ORESTES protocol (4). (B) Different buffer salt dilutions (0.7X and 0.4X) were tested using LS+TD conditions, as in panel A. Control refers to PCR in the absence of the reverse transcription step. (C) A slow temperature ramp between the annealing and extension steps was introduced, using conditions as in panel B. (D) Optimized low-stringency modified touchdown (MTD) PCR conditions were tested for the Hib8 primer (as described in the text) and compared with amplification produced using standard LS conditions of the ORESTES protocol (4).
Optimized Conditions of Dilution, Touchdown, and Slow Temperature Ramp

Figure 1D, lane LS+MTD, shows that a result similar to that of Figure 1C was obtained, with a complex amplification profile and a broad range of product sizes. For comparison, Figure 1D, lane LS, shows the amplification obtained with the same primer under standard low-stringency RT-PCR conditions described previously (3,4). Note that in Figure 1D, lane LS, a less efficient amplification is shown, when compared to Figure 1D, lane LS+MTD, with smaller PCR products and an undesired predominant band at 500 bases.

Nonredundancy of Libraries

To characterize the message content of the products obtained in the experiments shown in Figure 1C (lane 0.4X) and Figure 1D (lane LS+MTD), cDNAs were cloned, sequenced, and compared to the sequences available in the public GenBank database using the BLAST algorithm to look for similarities. Sequences obtained for each library were grouped into distinct classes, according to the BLAST results (Table 1). In each library, approximately 29% of the sequences did not match any S. mansoni sequences available in GenBank and are likely new S. mansoni gene fragments.

Library F24I exhibited 55% nonredundant sequences after 401 clones were sequenced (Table 1), and the single most redundant gene (i.e., 86 sequences out of 401, 21%) was 28S ribosomal RNA, a contaminant that has a sequence stretch that matches exactly the 3’-end sequence of the degenerate oligonucleotide primer used in this experiment. Library H8-1 was obtained with a different primer (Hib8), and a total of 306 sequences were obtained and analyzed, presenting a similar 55% nonredundancy (data not shown). The rRNA contaminants present in the mRNA preparations (Table 1, 25% of the sequences) were decreased in additional experiments by the use of magnetic bead separation of mRNA. Library F33 was constructed under exactly the same conditions as F24I, except for the magnetic bead preparation of the mRNA. A total of 642 sequences was obtained showing 65% nonredundancy (Table 1), with a significant decrease in the level of rRNA contamination (only 10% of the sequences matching rRNA). In fact, the magnetic bead mRNA purification method described here is used routinely in the collaborative S. mansoni EST project that we recently started (http://bioinfo.iq.usp.br/schisto/), and preliminary results show that, indeed, rRNA is in the range of 6%–8% of total acquired sequences.

An analysis of the data after eliminating the rRNA showed that 290 sequences remained in library F24I with 71.3% nonredundancy and 576 sequences for library F33 with 71.4% nonredundancy. Figure 2 shows the percentage of nonredundant sequences in the libraries for each set of 25 new sequences collected. As shown in Figure 2, it was possible to accumulate 200–300 sequences with a nonredundancy of about 75%. On average, 357 nonredundant sequences were obtained per library with an average nonredundancy level of 71.3% (for three libraries that were extensively sequenced to exhaustion). This amount of nonredundant sequences per library is considerably higher than the 15–20 nonredundant sequences obtained for each S. mansoni mini-library in the original work of Dias-Neto et al. (4). It is also higher than the average 40–50 sequences per library obtained in the HCGP (3). The number of cloning and library picking steps involved in the construction of mini-libraries is a significant rate-limiting and labor-intensive activity. The modifications introduced here have improved one of the most desirable features of the low-stringency RT-PCR amplification strategy, namely, the ability to acquire sequence data from a large number of different genes each time the method is applied to a single sample of mRNA, thus reducing the number of mini-libraries required.

Using eight different primers and the same methodology described above, we generated 2771 modified ORESTES sequences from S. mansoni. All libraries presented similar results (data not shown).

Normalization of cDNA Libraries Obtained by the Modified Touchdown Method—Estimates Using Human cDNA

There is another important characteristic of the low-stringency RT-PCR


Figure 2. Nonredundancy of libraries. Libraries were generated under conditions shown in Figure 1, C and D, using primers APYAF1 (libraries F24I and F33) or Hib8 (library H8-1). Library clones were analyzed as described in the text. The percent of nonredundancy is shown in relation to the cumulative number of sequences obtained for library F24I (squares), F33 (triangles), and H8-1 (circles).
Research Report

method, the ability to amplify messages that occur in relatively low abundance, generating normalized cDNA libraries (3). We wanted to check for the persistence of normalization with the use of touchdown and degenerate primers. However, this characteristic was difficult to analyze in the context of the very small amount of sequence information available for *S. mansoni*. The assessment of the degree of normalization (3) was carried out for human cDNA libraries by comparing a set of sequences obtained in a given library that had been generated using low-stringency RT-PCR and nondegenerate primers against the homologous EST sequence clusters obtained from poly-dT-primed EST libraries available from GenBank. The number of redundant sequences contained in each UniGene cluster from GenBank gives an estimate of the number of times a gene was cloned and sequenced by other groups, and it is expected to be roughly proportional to the relative abundance of that message. This appraisal seems to work for human genes since there are over 3.7 million human EST sequences available in GenBank, and others have used it as a ready means of estimating human gene expression (1). In fact, we have recently provided experimental evidence that supports this approach, using semi-quantitative RT-PCR to determine the level of expression of five human genes that represent UniGene clusters that vary in size from 10 to 324 members (2). Unfortunately, there are only 14,000 *S. mansoni* EST sequences deposited in GenBank, and it would be difficult to determine if the modifications introduced here affect the ability of low-stringency RT-PCR to amplify rare *S. mansoni* messages. Therefore, we decided to sequence a library of human cDNA clones obtained by low-stringency RT-PCR, using the same degenerate primer APYAF1 that was used with *S. mansoni* and the touchdown PCR strategy described above.

In this library, 25 unique clones were similar to UniGene cluster sequences. Figure 3 shows the percentage of modified ORESTES clones with the indicated sizes that fell into UniGene clusters. For comparison, a human cDNA library generated with randomly primed PCR by Macke et al. (Macke, J., P. Smallwood, and J. Nathans. 1996. Human retina cDNA randomly primed sublibrary, GenBank accession nos. W26664–W29128) was analyzed. The results in Figure 3 show that clones obtained with our modified ORESTES conditions and degenerate primer (solid bars) belonged predominantly to UniGene clusters with few members (i.e., 79% of our modified ORESTES clones fell into UniGene clusters with up to 128 members). In contrast, only 37% of the clones from the randomly primed library (open bars) fell into UniGene clusters with up to 128 members, the majority of them (63%) belonging to clusters of larger sizes. The profile obtained in Figure 3 (solid bars) indicates that using touchdown PCR and degenerate primers designed with the general strategies described here preserves the desired normalization of libraries (i.e., relatively rare messages continued to be amplified preferentially). Note that primers that have totally degenerate bases in the last two 3′-ends.

Figure 3. Normalization capacity of modified ORESTES and of a randomly primed cDNA library. Comparison between a cDNA library produced using the modified ORESTES protocol described in the text (using human mRNA extracted from breast tumor) and a human cDNA library produced using randomly primed PCR. The percentage of modified ORESTES sequences (solid bars) or randomly primed sequences (open bars) that were similar to UniGene clusters of the size indicated on the abscissa are shown (i.e., containing the indicated number of sequence members).

**Position of Modified ORESTES cDNA Clones in Relation to Full-Length *S. mansoni* Messages**

A subset of sequences among the 2771 *S. mansoni* DNA sequences generated with the modified ORESTES technique was identified that matched either full-length *S. mansoni* mRNA sequences from GenBank (Figure 4A) or full-length protein orthologs from other species (Figure 4B). It can be seen that 45% of the modified ORESTES sequences covered the center of full-length messages and only 10% of the sequences included either the 5′- or 3′-ends.

The utilization of degenerate primers and modified touchdown procedures did not affect the preferential amplification of the central portion of messages, which is a property of the method that seems to depend on the combined probability of two events, the first step of annealing the primer to mRNA during reverse transcription and
a subsequent annealing of the primer to the resulting cDNA strand in the subsequent PCR (3).

CONCLUSIONS

Taken together, all the above data indicate that significant improvements were introduced to the standard low-stringency RT-PCR protocol, such as the use of hybrid consensus-degenerate oligonucleotide primers, the use of touchdown PCR, and the use of diluted PCR buffer in combination with a slow temperature ramp between annealing and extension. All the modifications described here were shown to increase the number of different target messages that are amplified in a single round of the method. The amplification of a broad range of products permits the construction of low-redundancy cDNA clone libraries, thus increasing considerably the cost-effectiveness of gene discovery by high-throughput EST sequencing.

For *S. mansoni*, there has been a concerted effort from a number of laboratories around the world, which has resulted in the acquisition of nearly 14,000 *S. mansoni* EST sequences, of which 98% are from regular 3′ or 5′ EST libraries. In this context, the ORESTES approach described here provides an opportunity to generate sequence information from the internal portion of genes that will complement the existing 3′ and 5′ EST sequence data and will help to obtain the full-length sequence of *S. mansoni* genes. High-throughput sequencing of expressed messages from either regular libraries or low-stringency RT-PCR libraries should increase the coverage of the *S. mansoni* transcriptome and help to discover new genes. Recently, a BAC library has been established from *S. mansoni* genomic DNA that provides a new resource for the physical mapping and sequencing of the genome of the parasite (9). The combination of genomic sequence and expressed gene sequence information should result in a powerful set of tools that can be used in further studies of functional relationships in this important human pathogen. Moreover, the high-throughput approach described here is likely to be an important additional tool to help in the gene discovery in any complex organism.

ACKNOWLEDGMENTS

This work was supported by a grant from the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP). J.L.R.F. and R.D. received doctorate fellowships from FAPESP. S.V.-A. received an established investigator fellowship from the Conselho Nacional do Desenvolvimento Científico e Tecnológico (CNPq).

Figure 4. Relative position of modified ORESTES sequences with respect to full-length messages. Mapping of our *S. mansoni* cDNA sequences onto full-length sequences from GenBank is shown. The percentage of our matching sequences that pass through the relative positions along the full-length messages with which they match is indicated. (A) A BLASTN was compared with the GenBank full-length *S. mansoni* mRNA dataset, and results from 61 nonredundant matching sequences are shown. The 5′-end of full-length mRNAs is represented by 0 and the 3′-end by 1.0. (B) Matching sequences were translated into the six possible frames, a BLAST search was compared with the nonredundant GenBank protein dataset from all species, and results from 137 nonredundant matching sequences are shown. The amino terminus of proteins is represented by 0 and the carboxyl-terminal end by 1.0.
REFERENCES


Received 17 December 2001; accepted 29 January 2002.

Address correspondence to:
Dr. Sergio Verjovski-Almeida
Instituto de Quimica
Universidade de Sao Paulo
Av. Prof. Lineu Prestes, 748
Sao Paulo, SP 05508-900, Brasil
e-mail: verjo@iq.usp.br