Identification of Transcribed Sequences (ESTs) in the *Trypanosoma cruzi* Genome Project

Adeílton Brandão, Turan Urmenyi*, Edson Rondinelli*, Antonio Gonzalez, Antonio B de Miranda, Wim Degrave/+**

Departamento de Bioquímica e Biologia Molecular, Instituto Oswaldo Cruz, Av. Brasil 4365, 21045-900 Rio de Janeiro, RJ, Brasil *Instituto de Biofísica Carlos Chagas Filho, ICB, Universidade Federal de Rio de Janeiro, Ilha do Fundão, Rio de Janeiro, RJ, Brasil **Instituto de Parasitologia e Biomedicina, Calle Ventanilla 11, 18001 Granada, Spain

Random single pass sequencing of cDNA fragments, also known as generation of Expressed Sequence Tags (ESTs), has been highly successful in the study of the gene content of higher organisms, and forms an integral part of most genome projects, with the objective to identify new genes and targets for disease control and prevention and to generate mapping probes. In the Trypanosoma cruzi *genome project, EST sequencing has also been a starting point, and here we report data on the first 797 sequences obtained, partly from a CL Brener epimastigote non-normalized library, partly on a normalized library. Only around 30% of the sequences obtained showed similarity with Genbank and dbEST databases, half of which with sequences already reported for* T. cruzi.

Key words: expressed sequence tags sequencing - normalized cDNA library - *Trypanosoma cruzi* - clone CL Brener - Parasite Genome Projects

Generation of expressed sequence tags (EST) through partial sequencing of randomly selected cDNA clones is a process widely used to accelerate gene identification in genome projects. This approach for gene identification has been responsible for the rapid growth of public sequence databases and now represents more than half of the sequence records in Genbank (Boguski 1995). Identification of the gene content of an organism through cDNA analysis has many advantages, in particular the speed, easiness and large scale of the approach. Especially in the human genome initiative, EST sequencing has yielded data on hundreds of thousands of gene fragments, from a wide variety of tissues. The majority of the genome projects on more complex organisms devote at least some part of the effort to EST sequencing, such as is the case with parasite genome projects. An additional advantage to this technically straightforward approach is that potentially interesting new drug targets, vaccine candidates and new antigens can be identified in a short time frame and with limited resources. Moreover, tissue specific or life stage specific gene expression can be analyzed. In addition, ESTs can be used to tag loci on the chromosomal map or on genomic libraries and greatly enhance physical mapping. However, a number of drawbacks should be noted, especially when low quality or non-normalized libraries are used. A few highly expressed messengers are present in high copy number in the library, while intermediately expressed messengers can make up to 50% of the library. Hence, rare messengers, believed to be the most promissing ones in a parasite gene identification project, are hidden like a needle in the hay (Adams et al. 1995, Bonaldo et al. 1996). Through random clone picking and sequencing, one reaches rapidly up to 40% of redundancy with obvious low cost effectiveness (Cooke et al. 1996a). Various approaches have been tried to minimize this problem, including elimination of abundant genes identified through hybridization of clones with specific probes (Cooke et al. 1996b), normalization of the cDNA library (Bonaldo et al. 1996), generation of mini libraries through differential display or arbitrarily primed RT-PCR (Dias Neto et al. 1996, 1997).

Yet another problem turns gene discovery efforts through EST sequencing a difficult task. In general, a fair amount of sequence should be obtained from the protein coding region, to allow for a better probability of gene identification, since such identification is entirely based upon computer

Financial support: UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases; CYTED-Subprogram of Biotechnology (Spain), CNPq, Centro Brasileiro Argentino de Biotecnologia (CBAB/CABBIO).

+Corresponding author. Fax: +55-21-590.3495. E-mail: wdegrave@gene.dbbm.fiocruz.br

Received 20 August 1997
Accepted 10 September 1997

similarity analysis using BLASTN, BLASTX or FASTA against the Genbank, EMBL and dbEST databases. However, identification of similarity with a known protein sequence does not necessarily imply an identical or even similar function (Bairoch 1996). Zhang et al. (1997) reported the identification of EST's using only 9 nt sequences for database screening, however, such screening was done within the human genome project, where a very large number of (EST) sequences are known.

EST's were first generated from human cDNA libraries (Adams et al. 1991, Boguski, 1995), but nowadays other organisms, especially parasites such as *Schistosoma mansoni*, *Trypanosoma b. rodhesiense*, *Leishmania major* and *Plasmodium falciparum* are being subjected to gene identification by EST sequencing (Chakrabarti et al. 1994, el-Sayed et al. 1995, Franco et al. 1995). Thus, as a part of the collaborative effort to sequence the whole genome of *T. cruzi* clone CL Brener, we initiated the EST sequencing of a *T. cruzi* epimastigote cDNA library on an ABI 373A sequencer. The single pass sequencing is being performed either by Taq dyeprimer cycle sequencing or T7 dyedeoxi termination on double stranded plasmid DNA. Here we present the analysis of the first 797 sequences so far obtained. An additional 95 ESTs were obtained in Granada, Spain, by A Gonzalez.

MATERIALS AND METHODS

cDNA library construction - Non-normalized and normalized libraries were constructed in the laboratory of Dr Rondinelli and the methods used in their preparation will be published elsewhere.

Plasmid DNA preparation - Clones were randomly selected from LB agar plates and grown overnight in 4 ml LB/ampicilin at 37°C until saturation. Plasmid DNA was extracted by the alkaline lysis method using the Flexiprep kit from Pharmacia. DNA from 90 clones of the non-normalized library were double digested with the enzymes Eco RI and Hind III (New England Biolabs) to evaluate the mean insert size.

Sequencing - The double stranded plasmids were sequenced by Taq dyeprimer cycle sequencing or Sequenase dyedeoxi terminator with primers T7 or M13 reverse in one direction (5'-end) accordingly to protocols supplied by the manufacturer (ABI-Perkin-Elmer) and reactions were loaded on an automated sequencer ABI 373A. At the end of each sequencing run, chromatograms were visually inspected and sequences with poor readings were discarded while useful ones were edited, *i.e.*, vector sequences were removed and the end of the high quality sequence readings was determined. Sequences were then converted do GCG format, and transferred by ftp to our SGI Challenge-L 4xR4400 250MHz server for FASTA similarity searches. Original chromatograms without modifications were also saved.

ESTs are being deposited in dbEST. The following accession numbers have already been allocated: AA399704-AA399708; AA426656-AA426705; AA433291-AA433388; AA441733-AA441781; AA525699-AA525749; AA532113-AA532212; AA532063-AA532112.

RESULTS AND DISCUSSION

EST sequencing on the epimastigote non-normalized library - Ninety clones from the epimastigote non-normalized library were randomly picked and their mean insert size was determined, by restriction digestion, to be around 1 kb, with a minimum of 0,35 Kb and maximum of 3,4 Kb. In total, 267 clones were successfully sequenced using this library, 33% of which resulted in a positive match after FASTA analysis with Genbank/dbEST. Of these, however, 27% and 23%, respectively, represented ribosomal RNA and mitochondrial sequences. The remainder matched with histones, tubulines, heat shock proteins, sialidases and elongation factors (all around 5% of the matches), and ribosomal proteins (13%), as well as 12% with a variety of other genes. In summary, 67% of the sequences did not have similarity in the databases, and of the ones that did, 59% were with genes, previously described for *T. cruzi* (19.4% of the total). Thus, only 14% of the clones yielded a match with the database, and were not previously described in *T. cruzi*. The fairly high number of ribosomal RNA sequences in the library probably reflects the fact that the mRNA preparation was purified on oligo-dT cellulose in a single pass, or may indicate that internal priming on poly-A rich stretches in rRNA has occurred.

EST sequencing on the epimastigote normalized library - It was expected that the normalized library, from which abundant expressed gene sequences were removed, would yield better results in relation to the objectives of gene discovery through random sequencing. A total of 530 clones were sequenced thus far, using the same strategy as in the case of the non-normalized library. Only 28% of the sequences showed homology with the nucleotide sequence databases, and of these, 48% were already described in *T. cruzi*. Ribosomal RNA, mitochondrial sequences and tubulines are present in less than 1% of the total number of clones, however histone genes account for nearly 5%.

The low number of sequences with homology to sequences in the databases, after FASTA analysis, is striking but should not come as a surprise since this phenomenon was observed in all eukaryote genome projects and to a lesser extent in bac-

teria. We expect to be able to identify a few more gene sequences after ORF analysis and protein sequence comparison. One could argue that the majority of the sequences obtained would consist of 5'-untranslated sequences, however, the average read length was 350 bp and we expect the average distance from the mini-exon to the ATG start codon to be shorter. Moreover, Andersson (personal communication) found roughly the same value when sequencing ESTs from the 3'-end, and also after ORF analysis on genomic sequences derived from a cosmid. This finding reminds us that most of the biochemical and regulatory pathways in this parasite are still unrecognized and if we take into account its genome size estimation of 0.8-1.0 x 10^8 bp (Cano et al. 1995) much more than 6.000 genes, as demonstrated to be present in *Saccharomyces cerevisiae* (Mewes et al. 1997), are needed to direct the parasite life cycle.

In this first survey about the gene content of *T. cruzi*, several new genes could be identified. The list from both libraries comprises: ribosomal proteins L1, L5, L7, L11, L13a, L15, L18, L17, L26, L27a, L28, L34, L35a, L36, S2, S3, S4, S6, S12, S17, S21, S25, UMS (universal minicircle sequence) binding protein, p18 protein (RNA binding protein), TAT binding protein, Jun-binding protein, protein kinase C, human MEK kinase, casein kinase, serine/threonine protein kinase, phosphoglicerate kinase, CDC-2 related protein kinase, phosphomannmutase, ATP synthase, alkylhgydroperoxide reductase, dihydrorotate synthetase, glyceraldehyde 3-phosphate dehydrogenase, pyridine nucleotide linked dehydrogenase, putrescine N-methyltransferase, peptidyl-prolyl isomerase, RAS related protein, cyclophilin, 26S proteasome ATPase subunit, HMG-CoA reductase, bovine antioxidant protein. A full list of the matches, as well as FASTA/BLAST output and the actual sequences can be found on our www pages at www.dbbm.fiocruz.br.

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, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JÁ, Brandon RC, Chiu MW, Clayton RA, Cline RT, Cotton MD, Earle-Hughes J, Fine LD, FitzGerald LM, FitzHugh WM, Fritchman JL, Georghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle PS, Kelley JM, Klimek KM, Kelley JC, Liu Li, Marmaros SM, Merrick JM, Palanques RFM, 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 A, 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 JC 1995. Initial assesssment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature 377*: 03-174.

Bairoch A, 1996. Go hunting in sequence databases but watch out for the traps. *Trends in Genetics 12:* 425-427.

Boguski MS 1995. The turning point in genome research. *Trends in Biochem Sci 20*: 295-296.

Bonaldo MF, Lennon G, Soares MB 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res 6:* 791-806.

Cano MI, Gruber A, Vazquez M, Cortés A, Levin MJ, González A, Degrave W, Rondinelli E, Zingales B, Ramirez JL, Alonso C, Requena JM, Silveira JF 1995. Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* Genome Project. *Mol Biochem Parasitol 71*: 273-278.

Chakrabarti D, Reddy GR, Dame JB, Almira EC, Laipis PJ, Ferl RJ, Yang TP, Rowe TC, Schuster SM 1994. Analysis of expressed sequence tags from *Plasmodium falciparum*. *Mol Biochem Parasitol 66*: 97-104.

Cooke R, Mache R, Hofte H 1996a. EST and genomic sequencing projects. In GD Foster & D Twell (eds). *Plant gene isolation: Principles and practice.* John Wiley & Sons Ltd, London, UK.

Cooke R, Raynal M, Laudié M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G, Li YF, Mache R, Krivitsky M, Gy IJJ, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clément B, Phillips G, Hervé C, Bardet C, Termousaygue D, Lescure B, Lacomme C, Roby D, Jourjon MF, Chabrier P, Charpenteau JL, Desprez T, Amselem J, Chiapello H, Hofte H 1996b. Further progress towards a catalogue of all *Arabdopsis* genes: analysis of a set of 5000 non-redundant ESTs. *Plant J 9:* 101-124.

Dias Neto E, Harrop R, Correa-Oliveira R, Pena SDJ, Wilson RA, Simpson AJG 1996. The Schistosome genome project: RNA arbitraly primed PCR allows the accelerated generation of expressed sequence tags. *Mem Inst Oswaldo Cruz 91:* 655-657.

Dias Neto E, Harrop R, Oliveira RC, Wilson RA, Pena SDJ, Simpson AJG 1997. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the eficient generation of ESTs from nanogram quantities of mRNA. *Gene 186:* 135-142.

el-Sayed NM, Alarcon CM, Beck JC, Sheffield VC, Donelson JE 1995. cDNA expressed sequence tags of *Trypanosoma brucei rhodesiense* provide new insights into the biology of the parasite. *Mol Biochem*

Parasitol 73 : 75-90.

Franco GR, Adams MD, Soares MB, Simpson AJ, Venter JC, Pena SD 1995. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. *Gene 15*: 141-147.

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A 1997. Overview of the yeast genome. *Nature 29:*387(6632 Suppl): 7-65.

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW 1997. Gene expression profiles in normal and cancer cells. *Science 276:* 1268-1272.