Generation and analysis of expressed sequence tags from *Trypanosoma cruzi* trypomastigote and amastigote cDNA libraries ⋆

Fernán Agüero [a,*], Karim Ben Abdellah [b], Valeria Tekiel [a], Daniel O. Sánchez [a] and Antonio González [b]

[a] *Instituto de Investigaciones Biotecnológicas, Universidad Nacional de General San Martín - CONICET, Buenos Aires, Argentina*

[b] *Instituto de Parasitología y Biomedicina, CSIC, Granada, Spain*

Abstract

We have generated 2,771 expressed sequence tags (ESTs) from two cDNA libraries of *Trypanosoma cruzi* CL-Brener. The libraries were constructed from trypomastigote and amastigotes, using a spliced leader primer to synthesize the cDNA second strand, thus selecting for full-length cDNAs. Since the libraries were not normalized nor pre-screened, we compared the representation of transcripts between the two using a statistical test and identify a subset of transcripts that show apparent differential representation. A non-redundant set of 1,619 reconstructed transcripts was generated by sequence clustering. This dataset was used to perform similarity searches against protein and nucleotide databases. Based on these searches, 339 sequences could be assigned a putative identity. 1,116 sequences in the non-redundant clustered dataset (68.8%) are new expression tags, not represented in the *T. cruzi* epimastigote ESTs that are in the public databases. Additional information is provided online at http://genoma.unsam.edu.ar/projects/tram. To the best of our knowledge these are the first ESTs reported for the life cycle stages of *T. cruzi* that occur in the vertebrate host.

Key words: *Trypanosoma cruzi*, EST, trypomastigote, amastigote, cDNA library

Abbreviations: EST: expressed sequence tag; SL: spliced leader; CDS: coding sequence

Expressed sequence tags (ESTs) are single pass reads obtained from randomly selected cDNA clones. As such, they provide a highly cost-effective method to dis-

⋆ Note: nucleotide sequence data reported in this paper are available in the GenBank, EMBL and DDBJ databases under the accession numbers CF887912-CF890703.

* To whom correspondence should be addressed.
Email address: fernan@iib.unsam.edu.ar (Fernán Agüero).

cover new genes, to obtain data on gene expression and regulation, and to construct genome maps. The concept of using cDNAs as a route to expedited gene discovery was first demonstrated in the early 1980s [1]. For trypanosomatid protozoa, the production of ESTs has proven to be an efficient method for rapid gene discovery in the absence of a complete genome [2–7].

Trypanosoma cruzi is the agent of the American trypanosomiasis (Chagas disease), for which there is no definitive chemotherapeutic treatment. The parasite has a complex life cycle, with four main stages occurring in two hosts. In the insect host *T. cruzi* is found in the form of epimastigotes and metacyclic trypomastigotes. In the vertebrate host, it is found in the form of bloodstream trypomastigotes and intracellular amastigotes.

Now that three trypanosomatid genome projects are near completion (*Leishmania major*, *T. brucei* and *T. cruzi*), EST data will begin to be exploited for other uses. One such use is the correct annotation of genes within the genome. Despite the fact that trypanosomatid genomes contain mostly intronless genes, gene finding methods still fall short of producing correct predictions for every gene: genes might be missed entirely by gene finding programs, or their annotated start positions might not correspond to those of the expressed product. EST data is also useful for correctly annotating untranslated regions (UTRs) in genomic sequences. This is particularly important in trypanosomatid protozoa, given the widespread use of post-transcriptional mechanisms to regulate gene expression [8]: knowledge of target regions in UTRs can guide the search for cognate RNA binding proteins. Also, the discovery of RNA processing events is dependent on the comparison of EST or cDNA sequences against genomic data. In trypanosomatids both *cis*-splicing and alternative *trans*-splicing have been described [9,10]. Thus, a representative collection of ESTs represents a valuable resource for interested researchers.

Here we report the sequencing and analysis of ESTs derived from directionally cloned, spliced-leader cDNA libraries constructed from trypomastigotes and amastigotes of *Trypanosoma cruzi*.

Two cDNA libraries were constructed starting from polyA(+) RNA obtained from trypomastigotes (TcTR) and amastigotes (TcAM) of *T. cruzi* CL-Brener. Parasites were obtained by infection of semi-confluent monolayers of simian LLC-MK2 cells. Briefly, culture supernatants containing parasites released by cell lysis were collected at 96, 120 and 144 hrs post-infection, and the parasites harvested by centrifugation at 2000 $\times$ g. Pellets were then incubated for 2 hrs at 37°C in a CO_2 atmosphere to allow the freely swimming trypomastigotes to appear in the supernatant. This fraction was gently removed and trypomastigotes were collected by centrifugation at 2,400 $\times$ g and used in the construction of the TcTR library. The pellet containing remaining trypomastigotes and amastigotes was resuspended in LIT media supplemented with 10% fetal bovine serum, and incubated at 37°C for 36 hrs. Under these axenic conditions, trypomastigotes differentiate into amastigote-like

forms [11]. The amastigote and amastigote-like cells obtained after a final wash to remove cellular debris were used for the construction of the TcAM library.

PolyA(+) RNA was prepared from 10^9 cells using the QuickPrep^{TM}Micro mRNA Purification Kit (Amersham Pharmacia Biotech). Synthesis of the cDNA first strand was done with Superscript II reverse transcriptase (Gibco BRL) and an oligo-dT-Not primer (5'-CCTGCGGCCGCT(18)-3'). Synthesis of the cDNA second strand was performed using the Klenow fragment of DNA polymerase with a spliced leader (SL) primer (5'-GATACAGTTTCTGTA-3'). After methylation with EcoRI methylase, phosphorylated EcoRI linkers (5'-ACGGAATTCGT-3') were ligated to the cDNA. The resulting cDNA mixture was then digested with NotI and EcoRI restriction enzymes, subjected to size fractionation on SizeSep^{TM}400 Spun Columns (Pharmacia), and cloned into the dephosphorylated NotI/EcoRI sites of pBluescript KS+ (TcTR) or pBluescript SK+ (TcAM) (Stratagene).

Templates for sequencing were prepared as described [5] from randomly selected clones, and 5' sequenced with T3 (TcTR) or T7 (TcAM) primers (New England Biolabs) using an ABI 377 automated DNA sequencer (Applied Biosystems Inc, Foster City, CA). Chromatogram trace files were basecalled with phred [12]. Sequences were then trimmed to remove low quality bases and vector from the ends.

We generated a total of 2,771 EST sequences from the two libraries (see Table 1). Because the synthesis of the second strand of the cDNA was primed with a SL primer, we expected the cloned cDNAs to contain complete 5' ends. Moreover, we were able to discriminate between cDNAs containing complete 5' ends and cDNAs where non-specific annealing of the SL primer occurred by analyzing the sequence of the first 23-25 bases of the trimmed ESTs. In the former case we would expect the ESTs to contain the complete sequence of the SL primer followed by the last 8 bases of the SL (CTATATTG), which were deliberately omitted from the primer. Allowing for 3 mismatches in the 23 bases analyzed, we were able to find a complete match in 1,578 ESTs. Thus, at least 56.9% of the sequences could be considered to contain a complete 5' end based on this criterion.

Because individual ESTs are typically unedited and automatically processed, they contain low-quality regions and sequencing errors. These facts, together with the high redundancy of EST collections, led us to use the ESTs in the form of a clustered dataset. By clustering ESTs, redundancy is reduced, sequencing errors are usually compensated in the creation of majority-rule consensus sequences, and longer sequences can be obtained, through the joining of overlapping ESTs. Therefore, we next generated a non-redundant set of sequences from the 2,771 ESTs. This was done by sequence clustering, using STACKPack^{TM}[13]. Before clustering, sequences were masked against a local database of *T. cruzi* repetitive elements and retrotransposons. The resulting non-redundant EST set contained 1,619 sequences, made up of 437 consensus sequences derived from the clusters and 1,182 single-

Description	Number	Percent
Total ESTs generated	2,771	100
TcTR library	1,502	54.2
TcAM library	1,269	45.8
Non-redundant dataset	1,619	100
ESTs aligned to *T. cruzi* shotgun reads	1,490	92.0
ESTs similar to other *T. cruzi* ESTs	505	31.1
Matches *vs* protein databases	439	27.1
Detail of matches *vs* protein databases	439	100
Ribosomal proteins	107	24.5
Hypothetical proteins / Unknown function	100	22.7
(i) Trypanosomatid proteins	162	36.8
(ii) Proteins from other organisms	70	15.9

Table 1
BLAST matches to protein and nucleotide databases. All 1,619 ESTs in the non-redundant dataset were compared to sequences in public databases using BLASTN (nucleotide) or BLASTX (protein) [15]. Matches were considered significant if the alignments had BLAST E values $\leq 10^{-5}$ (BLASTX) or $\leq 10^{-40}$ (BLASTN). In the case of alignments against genomic shotgun reads, a number of ESTs that were below this cutoff were considered to be significant based on the inspection of the alignments (see main text). ESTs with significant similarity to genes coding for non-ribosomal proteins were divided in two groups based on the presence or lack of a descriptive annotation. Next, we further divided annotated proteins in groups based on (i) their similarity or (ii) lack of similarity to genes encoding proteins from other trypanosomatids. A table listing the matches found is provided online at http://genoma.unsam.edu.ar/projects/tram.

tons (ESTs with no similarity against any other EST). A file containing all 1,619 sequences is available online [14].

The non-redundant dataset was then compared against different databases using BLAST [15]. Protein databases were searched to identify the protein encoded by each sequence. Nucleotide databases were searched to compare this dataset to another *T. cruzi* EST dataset, and to map the cDNAs to *T. cruzi* genomic sequences. Table 1 shows a summary of the results.

A comparison of our non-redundant dataset against the 9,922 *T. cruzi* epimastigote ESTs present in GenBank yielded 505 (31.1%) matches (see Table 1). Thus, more than two thirds of our clustered sequences represent new ESTs for *T. cruzi*.

To map the ESTs against the draft *T. cruzi* genome, we compared all the sequences in our non-redundant set against unassembled shotgun reads obtained from the sequencing consortium (1,118,787 reads representing a 17.9X coverage (715 Mb) of the genome as of August 2003) using BLASTN. Significant matches were those showing a BLAST Expect value $\leq 10^{-40}$. Based on this cutoff we were able to align 1,230 sequences (75.9%) against the draft genome. However, a closer examination of the 389 sequences that gave no significant hits, showed that many of them were failing to give significant BLAST hits due to the filtering of low complexity sequences (the default BLAST behavior). As an example, we were unable to map cluster cl434 to genomic reads without turning off the low complexity filtering option of BLAST (see http://genoma.unsam.edu.ar/projects/tram/cl434.html). A new search against the genomic reads with the new parameters allowed us to increase the number of mapped sequences to 1318 (81.4%). For the remaining sequences, a careful analysis of the BLASTN alignments (with and without low complexity filtering) revealed that a significant number of them (172) were meaningful BLASTN alignments but with BLAST E values that were higher than the cutoff chosen. In all of these cases this was due to the length of the EST being too short to give a higher score or a more significant E value. Thus, as expected given the high genomic coverage attained we have been able to align over 90% (1,490) of our ESTs against the draft *T. cruzi* genome. Another 113 short ESTs that produced weak alignments and six sequences that were presumed to be a contamination, based on their matches to *Mycoplasma* and to the human genome (data not shown) were not considered further. Nine sequences, however, could not be aligned to any genomic read, and do not appear to be derived from a contamination based on the comparisons performed (cl21, cl74, cl409, tcam-49, tctr-655, tctr-718, tctr-805, tctr-1030, and tctr-1242). Two of them (tcam-49 and tctr-1030, accession numbers **CF887960** and **CF890227**) contain the complete SL sequence at the 5' end and thus are clearly from *T. cruzi*. If these sequences do represent remaining gaps in the genome, they could be used as markers to identify and close these gaps, thus helping the task of finishing and production of an assembled genome.

Based on the comparison against NCBI's GenPept database (containing a non-redundant collection of: GenBank CDS translations, and protein sequences from PDB, SwissProt, PIR and PRFA databases) we were able to assign a putative identity to 439 of the 1,619 sequences (20.9%). ESTs matching another 100 proteins that were annotated either as hypothetical or that lacked an informative description of their function were not counted (see Table 1). Also based on this comparison we were able to make an estimation of the number of genes represented by the 1,619 clusters. Analysis of the 439 positive BLASTX searches revealed 95 sequences that matched to 43 different genes (the remaining 345 sequences represent 345 different genes). Thus, in the subset corresponding to the 439 sequences with significant BLASTX matches, the redundancy in the clustering is 12% (439 - 95 + 43 / 439). This redundancy can be explained from a number of ESTs that do not overlap with their cognate clusters, and from the separation of paralogs into different clusters. Extrapolation of the observed redundancy to the whole dataset would mean that the

1,619 sequences in the clustered dataset actually represent 1,424 individual genes, although this has not been analyzed further.

A detailed analysis of the genes identified is outside the scope of this work, and will certainly be done by interested researchers in the field. A more detailed table, as well as the complete output of all BLAST searches is provided in the supplementary data that is available online [14]. Nonetheless, it is worth to mention some ESTs that encode proteins that have not been previously described in *T. cruzi*, and that could represent ideal targets for the design of parasite-specific drugs. EST tctr-1000 (GenBank accession number **CF890197**) has similarity to fungal sterol 24-C-methyltransferases (E.C. 2.1.1.41), which are also present in plants, but not in vertebrates; while cluster cl80 is similar to a triacylglycerol lipase-like protein that is so far present only in plants, fungi and bacteria. Other interesting findings include homologues of proteins involved in the initiation of transcription by RNA polymerase II. It is well known that trypanosomatid RNA polymerase II is able to initiate transcription in the absence of a specific promoter [8]. It is thus reasonable to expect that differences either in the polymerase itself or in additional factors are responsible for this feature. In this work we have identified amastc-675 (accession number **CF888638**), and amastc-784 (accession number **CF888747**), which showed significant similarity to the the SPT4 protein from yeast – a small zinc-finger containing protein [16], and the mammalian basic transcription factor BTF3.

Because the cDNA libraries were not normalized, and were prepared under similar conditions we decided to compare the number of ESTs obtained for a particular gene for the two libraries. To do this we used the re-constructed transcripts obtained in the clustering step, and for each cluster, we observed the number of contributed ESTs from each library (see Table 2). In most cases, the difference in abundance of ESTs for a particular cluster will arise through a combination of both (i) true differential expression (library heterogeneity) and (ii) sampling variability. To identify those cases that represent true library heterogeneity, we used a statistical analysis [17]. For each cluster, the likelihood of seeing the observed data is calculated, considering (i) or (ii) in turn. The two likelihoods are compared by subtracting the logs of the likelihoods, thus generating a log likelihood ratio (denoted R, see Table 2). Next, we performed the same analysis on 1,000 randomized datasets, observing the mean number of clusters that attained a particular value of R. A mean number of 13.35 clusters from the randomized dataset (false positives) had values of R of at least 2, 5.3 clusters had values of at least 3, and 0.52 clusters had values of at least 4. Thus, by selecting a threshold of $R > 4$, we expect to have a mean number of false positives of 0.52. In our EST dataset, 7 clusters had values of $R > 4$. This corresponds to a true positive rate of 82.5%.

As shown in Table 2, three of the seven clusters fulfilling this criteria correspond to ESTs encoding ribosomal proteins. It is known that global changes in the expression of ribosomal proteins can occur as a response to stress [18]. However, changes in the expression of individual ribosomal proteins could be explained if the pro-

Cluster	TcAM ESTs	TcTR ESTs	Similarity-based annotation	R factor
cl136	17	0	ribosomal protein S12	13.16
cl34	8	0	clathrin assembly protein AP19	6.19
cl237	7	0	–	5.42
cl244	7	0	–	5.42
cl89	6	0	TcSMUG mucin (L)	4.65
cl70	9	1	ribosomal protein L5	4.34
cl425	0	7	ribosomal protein SA (P40)	4.33

Table 2
Clusters showing apparent differential representation between the two libraries. The number of ESTs originated from each library was recorded for each cluster. Then, a statistic, denoted R, was calculated that measures the extent to which the differences in gene expression correspond to heterogeneities in the libraries as opposed to random sampling variability [17]. Based on the R values obtained from 1,000 randomized datasets, a threshold of $R > 4$ was selected to attain a number of false positives < 1. Clusters showing higher differences between the two libraries are at the top.

tein has an additional extraribosomal function [19]. Two of the clusters shown in Table 2 encode ribosomal proteins with known extraribosomal functions. Cluster cl136 encodes the *T. cruzi* homologue of the ribosomal protein S12. Apart from its function in the ribosome, the S12 ribosomal protein – a RNA chaperone – is one of several proteins that enhance phage T4 intron splicing *in vitro* [20]. Cluster cl425 encodes the *T. cruzi* homologue of the ribosomal protein SA, that is also found located at the cell surface, where it functions as a high-affinity laminin receptor [21]. The other clusters in Table 2 include a *T. cruzi* mucin of the SMUG gene family [22], and an homologue of the clathrin assembly protein AP-19. The L type SMUG mucins have been shown to be expressed in all life-cycle stages of *T. cruzi*, but the higher mRNA levels have been observed in the replicating epimastigotes and amastigotes. Our finding that cluster cl89 encoding this mucin shows a significantly higher representation in the TcAM library, thus agrees with the published data [22]. The clathrin assembly protein AP-19 is a small (σ) subunit of the adaptor protein (AP) complexes involved in the formation of intracellular transport vesicles. Although there is very little information regarding the role of the σ subunits, a putative function of this protein might be in the guiding of AP complexes to the appropriate membrane [23]. The observation that six of the seven clusters shown in Table 2 correspond to cases of apparent differential expression in the TcAM library – in spite of having sequenced a similar number of clones from both libraries – might be reflecting the more active state of the replicating amastigotes, as compared to the non-replicating trypomastigotes.

To the best of our knowledge, this report represents the first description of ESTs from the life cycle stages of *T. cruzi* that occur within the vertebrate host.

Acknowledgements

This work received financial support from the UNDP / World Bank / WHO Special Programme for Research and Training in Tropical Diseases (TDR), Agencia Nacional de Promoción Científica y Tecnológica (Argentina), and Fundación Antorchas (Argentina). Preliminary genomic data obtained from http://www.tigr.org/, was provided by the TIGR-SBRI-KI Sequencing Consortium supported by NIH grants AI45038, AI45061 & AI45039. We thank Antonio Lario, Rodrigo Pavón, Fernanda Peri and Diego Rey Serantes for technical assistance. FA is a postdoctoral fellow, and DOS is a member of the Research Career of the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. The work of DOS was partially supported by the Carrillo-Oñativia Fellowship, Ministerio de Salud (Argentina).

References

[1] Putney SD, Herlihy WC, Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature 1983;302:718–721.

[2] El-Sayed NMA, Alarcon CM, Beck JC, Sheffield VC, Donelson JE. cDNA expressed sequence tags of *Trypanosoma brucei rhodesiense* provide new insights into the biology of the parasite. Mol Biochem Parasitol 1995;73:75–90.

[3] Levick MP, Blackwell JM, Connor V, Coulson RM, Miles A, Smith HE, Wan KL, Ajioka J. An expressed sequence tag analysis of a full-length spliced-leader cDNA library from *Leishmania major* promastigotes. Mol Biochem Parasitol 1996;76:345–348.

[4] Brandao A, Urmenyi T, Rondinelli E, González A, de Miranda AB, Degrave W. Identification of transcribed sequences (ESTs) in the *Trypanosoma cruzi* genome project. Mem Inst Oswaldo Cruz 1997;92:863–866.

[5] Verdún RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC, Sánchez DO. Gene discovery through expressed sequence tag sequencing in *Trypanosoma cruzi*. Infect Immun 1998;66:5393–5398.

[6] Porcel BM, Tran AN, Tammi M, Nyarady Z, Rydaker M, Urmenyi TP, Rondinelli E, Pettersson U, Andersson B, Aslund L. Gene survey of the pathogenic protozoan *Trypanosoma cruzi*. Genome Res 2000;10:1103–1107.

[7] Agüero F, Campo C, Cremona L, Jäger A, Di Noia JM, Overath P, Sánchez DO, Frasch AC. Gene discovery in the freshwater fish parasite *Trypanosoma carassii*: identification of trans-sialidase-like and mucin-like genes. Infect Immun 2002;70:7140–7144.

[8] Clayton CE. Life without transcriptional control? From fly to man and back again. EMBO J 2002;21:1881–1888.

[9] Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, Stern LS, Wang Z, Ullu E, Tschudi C. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. RNA 2000;6:163–169.

[10] Manning-Cela R, González A, Swindle J, Alternative splicing of LYT1 transcripts in *Trypanosoma cruzi*. Infect Immun 2002;70:4726–4728.

[11] Piras MM, Piras R, Henriquez D, Negri S. Changes in morphology and infectivity of cell culture-derived trypomastigotes of *Trypanosoma cruzi*. Mol Biochem Parasitol 1982;6:67–81.

[12] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. Genome Res 1998;8:175–185.

[13] Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W. STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Res 2001;29:234–238.

[14] Supplementary material available online.
URL `http://genoma.unsam.edu.ar/projects/tram`

[15] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

[16] Malone EA, Fassler JS, Winston F. Molecular and genetic characterization of SPT4, a gene important for transcription initiation in *Saccharomyces cerevisiae*. Mol Gen Genet 1993;237:449–459.

[17] Stekel DJ, Git Y, Falciani F. The comparison of gene expression from multiple cDNA libraries. Genome Res 2000;10:2055–2061.

[18] Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA. Remodeling of Yeast Genome Expression in Response to Environmental Changes. Mol Biol Cell 2001;12:323–337.

[19] Wool IG. Extraribosomal functions of ribosomal proteins. Trends Biochem Sci 1996;21:164–165.

[20] Coetzee T, Herschlag D, Belfort M. *Escherichia coli* proteins, including ribosomal protein S12, facilitate in vitro splicing of phage T4 introns by acting as RNA chaperones. Genes Dev 1994;8:11575–11588.

[21] Ardini E, Pesole G, Tagliabue E, Magnifico A, Castronovo V, Sobel ME, Colnaghi MI, Menard S. The 67-kDa laminin receptor originated from a ribosomal protein that acquired a dual function during evolution. Mol Biol Evol 1998;15:1017–1025.

[22] D'Orso I, Di Noia JM, Sánchez DO, Frasch AC. AU-rich Elements in the 3'-Untranslated Region of a New Mucin-type Gene Family of *Trypanosoma cruzi* Confers mRNA Instability and Modulates Translation Efficiency. J Biol Chem 2000;275:10218–10227.

[23] Page LJ, Robinson MS. Targeting signals and subunit interactions in coated vesicle adaptor complexes. J Cell Biol 1995;131:619–630.