READERS' OPINION

Trypanosomatid EST: a neglected information resource regarding flagellated protozoa?

Adeilton Brandão

Laboratório de Epidemiologia Molecular de Doenças Infecciosas,
Instituto Oswaldo Cruz-Fiocruz, Av. Brasil 4365, 21045-900
Rio de Janeiro, RJ, Brasil

Expressed Sequence Tag (EST) sequence analysis rapidly gained widespread use and application for the discovery of gene transcripts in a variety of organisms (Adams et al. 1991). As of April 11, 2008, 51,391,051 ESTs entries were recorded in the dbEST NCBI (National Center of Biotechnology Information). Most of these (61%) are generated from laboratory models and economically important organisms (21 species; see Table I for the 10 most frequently sequenced organisms). ESTs have proven invaluable tools for accelerating the discovery of transcribed sequences in the genome and generating markers for genome mapping. Several bioinformatic methods for the analysis of ESTs at both small and large scales have been developed in the last 10 years. A thorough revision of these methods as well as a comprehensive analysis regarding the breadth of their applications for EST processing, quality sequence analysis and functional discovery has been published by Nagaraj et al. (2007). Accompanying this road map analysis is a guide through the world of EST bioinformatics tools available at http://biolinfo.org/EST/. This guide contains an extensive list of software and methods.

For the purpose of this opinion article, I am focusing on the *Trypanosoma cruzi* EST collection. This analysis may, in general terms, be applied to other trypanosomatids. To illustrate this opinion with examples, I have searched the *T. cruzi* EST dataset as described below.

The full collection of *T. cruzi* ESTs in the dbEST-GenBank - NCBI was downloaded and screened for the presence of the spliced leader fragment at the 5' end. After redundancy elimination, this EST set was aligned with ClustalX (Thompson et al. 1997) in order to identify genes with evidence of alternative trans-splicing. Each EST set was compared to genomic sequences available either in GenBank (www.ncbi.nlm.nih.gov) or GeneDB (www.GeneDB.org) to confirm the presence of poly-pyrimidine-rich regions and trans-splicing sites (AG dinucleotide). This procedure allowed the discarding of any artifacts from the cDNA libraries primed with spliced leader fragments. The presence of a poly-pyrimidine-rich region (at least 10 bases long) before the trans-splicing site in the DNA genomic segment being analyzed is an indication that the transcribed mRNA has been correctly primed. According to this criterion, I selected 88 sets of ESTs for further analyses. These analyses included BLAST similarity searches to label genes, inspection of 5' untranslated region (UTR) for additional poly-pyrimidine-rich regions and validation of trans-splicing sites. The EST figures for other parasites and model organisms were obtained directly from the dbEST summary page at http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.

ESTs from parasitic organisms - Parasitic organisms account for only 2.5% of the total EST entries in the dbEST and they represent, approximately, 60 human and animal parasitic species (Table II). The most frequently sequenced parasite is *Schistosoma mansoni.* Among the protozoan parasites (28 species, 4.1% of parasite ESTs and 1% of total EST entries), the apicomplexa organisms have the largest number of ESTs. Specialized databases for *Plasmodium* spp (www.plasmodb.org) and *Toxoplasma gondii* (www.toxodb.org) provides resources that complement the information coming from these organisms' respective genome projects. This effort to continuously update the EST collection has also been important in correcting gene predictions and annotations of the *Plasmodium* genome. *Plasmodium* was also the first parasite to have an EST sequencing project (Chakrabart et al. 1994).

The trypanosomatid EST collection - Trypanosomatids, which include data from nine species, account for less than 0.1% of the total ESTs entries and 3.0% of total parasitic ESTs (Table III). *T. cruzi* is the most frequently sequenced trypanosomatid, ranking in the 18th position in the whole dataset. This raises the question of why trypanosomatids are not covered by extensive EST sequencing. Several reasons related to both technical nature and scientific goals are implicated in the answer to this question. First, the number of laboratories performing EST sequencing in trypanosomatid and parasites in general is small compared to the number involved with model or economically important organisms. EST sequencing has been surpassed by more powerful techniques for gene transcription analysis, such as microarrays, SAGE and its variants. EST sequencing is limited due to its dependence on sequencing facilities, requirement for the construction of good cDNA libraries (total and normalized) and typical biasing of clones obtained towards more frequently transcribed genes. In addition, the resources needed to completely characterize libraries mean that every EST sequencing project is executed on terms similar to those of full genome projects. Single laboratories are not prepared to undertake massive EST sequencing as they would for a typical SAGE or microarray experiment. The costs of these latter techniques are within reach of single laboratories. Thus, EST sequencing-generated information regarding the trans-cription activities of genes is less rewarding today than it was 10 years ago. In contrast to the direct sequencing of cDNAs, however, the quantitative and arraying methods do not

Financial support: Instituto Oswaldo Cruz-Fiocruz
Corresponding author: abrandao@fiocruz.br
Received 27 April 2008
Accepted 31 July 2008

provide sufficient sequence information to address features like polycystronic transcription and trans-splicing. Though an EST database encompassing all of the organisms represented in the current version of GenBank (Lee et al. 2005) has been constructed, some specific features of trypanosomatid may not be evident in this kind of general analysis. The main assumption for this analysis is based on a kind of gene organization, including gene interruptions and cis-splicing, that is not present in trypanosomatid. The overwhelming majority of trypanosomatid genes do not present introns and mRNA processing occurs via trans-splicing. Thus, refined analysis driven towards trypanosomatid features should be performed in order to extract relevant information regarding gene transcription and organization. To this end, I have taken *T. cruzi* as an example of what transcriptionally relevant information we can obtain from the present collection of ESTs. I propose that more monoxenous and heteroxenous trypanosomatid ESTs should be produced as tools to understand the evolution of transcription in these protozoa.

TABLE I

The first 10 organisms ranked by amount of EST entries in dbEST

Rank	Organism	Amount of EST
1	*Homo sapiens* (human)	8,136,592
2	*Mus musculus* and *M. domesticus* (mouse)	4,850,258
3	*Arabidopsis thaliana* (thale cress)	1,526,133
4	*Bos taurus* (cattle)	1,517,053
5	*Sus scrofa* (pig)	1,476,357
6	*Zea mays* (maize)	1,462,612
7	*Danio rerio* (zebrafish)	1,379,829
8	*Xenopus tropicalis* (western clawed frog)	1,271,375
9	*Oryza sativa* (rice)	1,220,261
10	*Triticum aestivum* (wheat)	1,051,465

source: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html; dbEST release 041108, April 11, 2008.

What information can be extracted from a trypanosomatid EST collection? The T. cruzi example - T. cruzi, the flagellated protozoan that causes Chagas disease in humans, exhibits gene organization and molecular processes (e.g., intronless genes and polycistronic transcription) similar to those of prokaryotes cells (Hausler et al. 1997). The immature polycistronic transcript is processed via a trans-splicing mechanism, which adds a small, spliced, 39 nucleotide RNA leader to the 5' end of the newly generated monocistronic messengers (Campbell et al. 2003). Variant mRNAs can be found in trypanosomes due to different transcription initiation sites that generate distinct 5' UTRs lengths (alternative trans-splicing) (Nepomuceno-Silva et al. 2001). Conversely, alternative polyadenylation is also a mechanism for the generation of transcriptional diversity (different 3' UTR lengths) from the same gene (Kubo et al. 2006). Ten years ago, the first *T. cruzi* ESTs were publicly available (Brandão et al. 1997). The pattern arising from that work showed a picture that would be confirmed at the end of the genome sequencing project for this protozoan. Most of the transcripts coded for genes either of unknown function or specific to this parasite. The EST sequencing projects that have followed this first initiative, including a larger numbers of clones, did not diverge from this first project. However, cDNAs from other developmental stages (e.g., trypomastigote and amastigote) were included in sequence database (Verdun et al. 1998, Aguero et al. 2004, Cerqueira et al. 2005). These numbers show that a great deal of research effort should be focused on deciphering the codes behind the *T. cruzi* genome.

Population bias of T. cruzi ESTs - Despite its relatively simple gene organization and mRNA processing, *T. cruzi* exhibits functional and population diversity that can be observed by the numerous vertebrate hosts it infects as well as the different clinical presentations of Chagas disease (Devera et al. 2003). Two major phylogenetic lineages or at least three populations groups are recognized by molecular markers (Fernandes et al. 2001, Brandão et al. 2006): *T. cruzi* I, *T. cruzi* II and *T. cruzi* III zymodemes. The *T. cruzi* EST collection was not

TABLE II

The first 10 parasites ranked by amount of EST entries in dbEST

Rank[a]	Organism	Amount of EST
1 (41)	*Schistosoma mansoni*	205,892
2 (64)	*Toxoplasma gondii*	129,421
3 (70)	*Schistosoma japonicum*	97,543
4 (88)	*Ancylostoma caninum*	80,551
5 (109)	*Ascaris suum*	56,492
6 (122)	*Plasmodium berghei* strain ANKA	47,085
7 (156)	*Plasmodium falciparum*	35,442
8 (160)	*Eimeria tenella*	34,998
9 (201)	*Strongyloides ratti*	27,366
10 (213)	*Brugia malayi*	26,215

a: number in parenthesis refers to organism rank in dbEST organisms list; source: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html; dbEST release 041108, April 11, 2008.

TABLE III

Trypanosomatids ranked by the amount of EST entries in dbEST

Rank[a]	Organism	Amount of EST
1 (324)	*Trypanosoma cruzi*	14,023
2 (399)	*Leishmania donovani chagasi*	9,839
3 (553)	*Trypanosoma brucei rhodesiense*	4,821
4 (685)	*Leishmania major*	2,191
5 (717)	*Trypanosoma carassii*	1,921
6 (899)	*Phytomonas serpens*	697
7 (1,074)	*Trypanosoma brucei brucei*	196
8 (1,110)	*Leishmania amazonensis*	147
9 (1,317)	*Leishmania infantum*	21
10 (1,485)	*Leishmania braziliensis*	2

a: number in parenthesis refers to their rank in dbEST organisms list; source: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html; dbEST release 041108, April 11, 2008.

exclusively obtained from a single strain because ESTs from the CL-Brener, Dm28c, Y, Tulahuen and Tehuantepec strains at different developmental stages have been deposited. However, they do not account for all of the diversity of *T. cruzi* populations. Because these populations can behave distinctly, with regard to both biological and clinical parameters, these biased cDNAs do not provide a full picture of their functional heterogeneity. The total number of entries in the dbEST-NCBI is almost twice as high as the estimated number of genes in *T. cruzi*. However, redundancies and multiple fragments for the same gene bias the EST collection towards more frequently expressed genes. To avoid a flawed *T. cruzi* EST analysis, transcripts from representative strains of the major phylogenetic groups should be present in the database.

Alternative trans-splicing - One of the post-transcription modifications in trypanosomatids involves trans-splicing site changes (Manning-Cela et al. 2002, Helm et al. 2008). This can be observed for several genes in *T. cruzi;* from approximately 14,000 ESTs, 195 entries were identified as possessing at least two trans-splicing sites and generating mRNAs with distinct 5' UTRs. The majority of these mRNAs correspond to genes coding for *ribosomal proteins (rp)*. In more general terms, two types of mRNAs can be deduced from these *T. cruzi* EST sets (supplementary data): (i) mRNAs that differ in length (number of bases) and composition of the 5' UTR. For example, EST CF889765 in *rp L7* contains a 5' UTR with 12 bases, whereas EST CF888327 contains a 5' UTR with 35 bases. This difference suggests that there are at least two trans-splicing sites along the 5' UTR. Other examples include *Histone H3* (ESTs CF890229, CF890408 and CF890611) and *rp L34* (ESTs CF889802, CF890219). The ESTs for these last two genes came from the trypomastigote stage of *T. cruzi* CL-Brener (Aguero et al. 2004); (ii) mRNAs with similar lengths (number of bases) but different nucleotide composition of the 5' UTR. For example, *rp L27* ESTs CF888060 and CF888883 show different composition of their 5' UTRs. They are not differentially trans-spliced. The flanking regions for each copy in the genome are different, resulting in the production of two mRNAs that code for the same protein but differ in their 5' UTRs.

In addition to those coding for *rp*s, other genes also present variations in their 5' UTRs. For example, ESTs CF888803 (from strain CL-Brener) and CB923724 (from strain Tulahuen), which both derive from a gene encoding a hypothetical protein, present similar coding regions but have 5' UTRs that differ by 113 nucleotides in length. Pairwise comparison of these mRNAs with their corresponding genomic sequences reveals that CL-Brener exhibits two loci for this gene: one copy gives rise to mRNAs corresponding to both ESTs (CF888803 and CB923724), whereas the other copy matches only EST CB923724. The latter copy lacks the additional trans-splicing site that yields the longer mRNA (CF888803).

The transcription of some genes yielding variant mRNAs may use acceptors splicing sites different from the AG dinucleotide that follows the poly-pyrimidine-rich region. Through comparison with genomic sequences, we observed the use of five non-canonical acceptor sites by at least one copy of some genes. The GG dinucleotide was used as surrogate acceptor site for two genes encoding a hypothetical protein (ESTs CF889477 and CF890296). The *mucin TcSMUGS* (ESTs AI035079, AI612603, CF890337 and CF890370) and *rp S6* (ESTs CF889016, CF890501) transcripts present TG and CG dinucleotides, respectively, as alternative acceptor sites.

Analysis of both ESTs and the genomic organization of the genes show that variant mRNA production is more frequent in duplicated or multicopy genes. However, single copy genes that have long 5' UTRs also exhibit this phenomenon. For example, ESTs CF888585 and CF888100 originated from the same single copy gene encoding a hypothetical protein that is transcribed into two mRNAs with distinct 5' UTR lengths (82 and 200 bases).

Below, I provide a list of information that can be retrieved for further experimental validation:

- Trans-splicing site and additional/multiple trans-splicing site in 5' UTR
- Exact length and start of 5' UTR
- Composition of both UTRs
- Sequence context for additional trans-splicing site
- Use of non canonical dinucleotides for the trans-splicing site
- Exact length and end of 3' UTR
- Poly-A site and additional/multiple poly-A sites
- Sequence context for poly-A sites
- Sequence context for trans-splicing signal (poly-pyrimidine rich regions) after comparison to genomic sequence

Two of these points – the definition of polyadenylation sites and the transcription starting points – are extremely important for the comprehension of key trypanosomatid mRNA processing events. Trypanosomatids diverge from the rest of the eukaryotes at this point, because they do not have a consensus nucleotide signal that indicates probable polyadenylation sites. In addition, the coupling of trans-splicing and poly-A addition in just one molecular event represents a differential mechanism of mRNA processing. Therefore, knowing what parts of intergenic sequences and flanking transcriptional starting points are present in pre-mRNA is a necessary step in understanding these phenomena. Experimental evidence suggests that UTRs influence gene transcription and expression as well as mRNA half life and degradation. This means that EST sequences still provide a source of information for expanding these experimental findings to the trypanosomatid universe. The *T. cruzi* genome sequencing effort has made clear that the exact definition of a UTR is dependent on experimental information and it should be checked by full cDNAs or ESTs primed at either the 5' or 3' end (Brandão 2006).

Trypanosomatid and next-generation sequencing technology: who will pay for it? - For almost 30 years, DNA sequencing has relied on the enzymatic method of dideoxinucleotide chain termination developed by Sanger et al. (1977). The introduction of innovations, such as a laser based fluorescence detection, capillary electrophoresis and automated base calling, has not altered the core of the methodology. Sanger sequencing was the tool that shaped genomic analysis into the new face of biology. Due to the economic and entrepreneurial opportunities provided up by this new biology an entire industry has been generated; furthermore, the sequencing knowledge base has moved from academia to private organizations. Some of the giants in the Biotech and pharmaceutical segments, eager to serve as new standards in genomics areas, steadily captured newcomers and advertised innovations in DNA sequencing technology. While the scientific community uses and adapts these technologies to its needs, a battle among the owners and vendors of these technologies is currently underway to define which technology will win the race and become the standard technique. In this new field of sequencing-by-synthesis, three main competitors and techniques are struggling for researchers' attention and resources. These include Roche's GS FLX Genome Analyzer (454 Life Science pyrosequencing, http://www.roche-applied-science.com), Illumina's Solexa 1G sequencer (http://www.illumina.com) and Applied Biosystem's SOLiD system (Supported Oligonucleotide Ligation and Detection) (http://solid.appliedbiosystems.com). These technologies bring technical innovations in the way DNA sequence fragments are generated. For example, 454 sequencing (Roche GS-FLX) overcome the inefficiency and bias of PCR amplification by using lipid droplets to enclose unique template DNA molecules (emulsion PCR) along with the highly parallel processing capacity of pyrosequencing (Ronaghi et al. 1996, Margulies et al. 2005). The Illumina system uses a Sanger-like, four-color sequencing system and "innovates" with solid phase bridge PCR amplification and the reversible chain terminator (Solexa sequencing). SOLiD technology from Applied Biosystems may be viewed as an intermediate between pyrosequencing and the solexa and it innovates with sequencing by ligation based-chemistry. A deep coverage of these technologies, along with its achievements, potential applications and drawbacks, has been reviewed in Schuster (2008) and Holt and Jones (2008). For trypanosomatids, the technological capacity to deal with massive genomes and transcriptomes offers an immense window of opportunity; this is especially true since one of the most significant concerns regarding trypanosomatid is their diversity of environments, hosts and genotypes. Approaches similar to metagenomic sequencing analysis using pyrosequencing (Edwards et al. 2006) could be envisaged for high throughput diversity analysis in trypanosomatid. Additionally, searching for drug-induced alterations is a good starting point in genome resequencing projects involving *T. cruzi* strains and isolates. The re-sequencing of experimentally modified genomes is an immediate application for these technologies and this has been demonstrated by the full genome methylation analysis of *Arabidopsis thaliana* with solexa sequencing (Cokus et al. 2008).

A new world of trypanosomatid transcription processing can emerge via these technologies. Since parasitic organisms are responsible for neglected diseases in lower income countries, however, the question remaining is: who will pay for it?

CONCLUSION

Taking *T. cruzi* ESTs as an example of a potential source of information useful for understanding similar processes in other trypanosomatid, I present here some points to support the widening of EST sequencing in trypanosomatids: 1: Trypanosomatids can use two mechanisms to generate transcript diversity. They may have either (i) distinct splicing acceptor sites (usually an AG dinucleotide located downstream of a tract of poly-pyrimidines) or (ii) distinct polyadenylation sites. By using both mechanisms, trypanosomatids generate mRNAs corresponding to the same gene with different UTR lengths and/or compositions. Since no removal of introns is necessary in the mRNA processing of *T. cruzi*, alternative cis-splicing should be uncommon. Until now, few experimental works have demonstrated alterations in protein composition due to alternative cis-splicing (Mair et al. 2000). 2: Additional trans-splicing in *T. cruzi* occur more frequently in *rp* genes. Nevertheless, the existence of this phenomenon in other genes leads to the assumption that this is a common feature of transcriptional activity in the *T. cruzi* epimastigote and other developmental stages. For example, ESTs from the genes coding for *Histone H3* and *rp L34* were obtained from *T. cruzi* CL-Brener trypomastigotes RNA (Aguero et al. 2004). These ESTs exhibit alternative trans-splicing at this stage. The final result of this process is that two or more mRNAs with different 5' UTRs code for one specific protein. The alternative trans-splicing may impact mRNA translation because many genes in trypanosomatid are regulated via other mRNA interactions at the post-transcriptional level (Teixeira et al. 1995, Hausler & Clayton 1996, D'Orso & Frasch 2001, D'Orso et al. 2003). Altered 5' or 3' UTRs imply the existence of mRNAs with different properties that may change parameters such as half-life, secondary structures, regulatory motifs and protein binding sites. 3: The comparison of some ESTs to the genomic sequence of *T. cruzi* (CL-Brener clone) shows that some genes use non-canonical trans-splicing sites other than the classical AG. This use of non-canonical trans-splicing sites is another possibility for generating mRNA variation in *T.* cruzi and possibly other trypanosomatid.

Despite all of the potential of EST analysis, the trypanosomatid collection in the EST section of GenBank remains deceptively small in comparison to those of either model organisms (e.g., *Mus musculus, Homo sapiens* and *A. thaliana*) or other protozoan parasites. Trypanosomatid EST sequencing efforts reflect the same situation as the diseases they cause: as a tool for information, EST sequencing is neglected in research focused on these parasites. Protozoa researchers should endeavor to increase the number of EST entries from trypanosomatids in public databases.

Next-generation sequencing technology provides a broad window of opportunity to tackle the so-called diversity in trypanosomatids from a medical perspective. For example, this includes the task of identifying key targets in transcription and metabolic networks with the goal of drug development and immunotherapy. However, the costs of these technologies are prohibitively high for most laboratories. Due to this obstacle, "old generation sequencing methodologies" such as EST analysis and SAGE still play important roles in trypanosomatid research.

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.

Aguero F, Ben Abdellah K, Tekiel V, Sanchez DO, Gonzalez A 2004. Generation and analysis of expressed sequence tags from *Trypanosoma cruzi* trypomastigote and amastigote cDNA libraries. *Mol Biochem Parasitol 136*: 221-225.

Brandão A 2006. The untranslated regions of genes from *Trypanosoma cruzi*, perspectives for functional characterization of strains and isolates. *Mem Inst Oswaldo Cruz 101*: 775-777.

Brandão A, Fernandes O 2006. *Trypanosoma cruzi*: mutations in the 3' untranslated region (3' UTR) of calmodulin gene are specific for lineages *T. cruzi* I, *T. cruzi* II and the Zymodeme III isolates. *Exp Parasitol 112*: 247-252.

Brandão A, Ürmenyi TP, Rondinelli E, Gonzalez A, Miranda AB, Degrave W 1997. Identification of Transcribed Sequences (ESTs) in the *Trypanosoma cruzi* Genome Project. *Mem Inst Oswaldo Cruz 92*: 863-866.

Campbell DA, Thomas S, Sturm NR 2003. Transcription in kinetoplastid protozoa, why be normal? *Microbes Infect 5*: 1231-1240.

Cerqueira GC, DaRocha WD, Campos PC, Zouain CS, Teixeira SMR 2005. Analysis of expressed sequence tags from *Trypanosoma cruzi* amastigotes. *Mem Inst Oswaldo Cruz 100*: 385-389.

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.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature 452*: 215-219.

Devera R, Fernandes O, Coura JR 2003. Should *Trypanosoma cruzi* be called "cruzi" complex? A review of the parasite diversity and the potential of selecting population after in vitro culturing and mice infection. *Mem Inst Oswaldo Cruz 98*: 1-12.

D'Orso I, De Gaudenzi JG, Frasch AC 2003. RNA-binding proteins and mRNA turnover in trypanosomes. *Trends Parasitol 19*: 151-155.

D'Orso I, Frasch AC 2001**.** TcUBP-1, a developmentally regulated U-rich RNA-binding protein involved in selective mRNA destabilization in trypanosomes. *J Biol Chem 276*: 34801-34809.

Edwards RA, Rodrigues-Brito B, Wegle L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC, Rohwer F 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics 7*: 57.

Fernandes O, Santos SS, Cupolillo E, Mendonça B, Derre R, Junqueira AC, Santos LC, Sturm NR, Naiff RD, Barret TV, Campbell DA, Coura JR 2001. A mini-exon multiplex polymerase chain reaction to distinguish the major groups of *Trypanosoma cruzi* and *T. rangeli* in the Brazilian Amazon. *Trans R Soc Trop Med Hyg* 95: 97-99.

Hausler T, Clayton C 1996. Post-transcriptional control of hsp70 mRNA in *Trypanosoma brucei*. *Mol Biochem Parasitol 76*: 57-71.

Hausler T, Stierhof YD, Blattner J, Clayton C 1997. Conservation of mitochondrial targeting sequence function in mitochondrial and hydrogenosomal proteins from the early-branching eukaryotes. *Crithidia, Trypanosoma* and *Trichomonas. Eur J Cell Biol 3*: 240-251.

Helm JR, Wilson ME, Donelson JE 2008. Different trans RNA splicing events in bloodstream and procyclic *Trypanosoma brucei. Mol Biochem Parasitol 159*: 134-137.

Holt RA, Jones SJM 2008. The new paradigm of flow cell sequencing. *Genome Res 18*: 839-846.

Kubo T, Wada T, Yamaguchi Y, Shimizu A, Handa H 2006. Knockdown of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3' UTRs. *Nucleic Acids Res 34*: 6264-6271.

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res 33*: D71-74.

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 2000. A new twist in trypanosome RNA metabolism, cis-splicing of pre-mRNA. *RNA 6*: 163-169.

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

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*: 376-380.

Nagaraj SH, Gasser RB, Ranganathan SA 2007. Hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief in Bioinform 8*: 6-21.

Nepomuceno-Silva JL, Yokoyama K, de Mello LD, Mendonca SM, Paixao JC, Baron R, Faye JC, Buckner FS, Van Voorhis WC, Gelb MH, Lopes UG 2001. TcRho1, a farnesylated Rho family homologue from *Trypanosoma cruzi*, cloning, trans-splicing, and prenylation studies. *J Biol Chem 276*: 29711-29718.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem 242*: 84-89.

Sanger F, Nicklen S, Coulson AR 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA 74*: 5463-5467.

Schuster CS 2008. Next-generation sequencing transforms today's biology. *Nat Methods 5*: 6-18.

Teixeira SM, Kirchhoff LV, Donelson JE 1995. Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of *Trypanosoma cruzi*. *J Biol Chem 270*: 22586-22594.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res 25*: 4876-4882.

Verdun RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC, Sanchez DO 1998. Gene discovery through expressed sequence Tag sequencing in *Trypanosoma cruzi*. *Infect Immun 66*: 5393-5398.

Supplementary data

ESTs from *T. cruzi* genes that transcribe variant mRNAs

mRNA variant	EST accession nr	Gene
1	CF887977, CF888075	Hypothetical
1	CF888178, CF890246	Hypothetical
1	AI664647, AI562498	Hypothetical
1	CF888254, CF888217	Hypothetical
1	CF889155, AA676081	Hypothetical
1	CF888585, CF888100	Hypothetical
1	CF888349, CF889145	Hypothetical
1	AA952481, AA952688	Hypothetical
1	CF889037, CF890237	Hypothetical
1	CF889770, CF889519	Hypothetical
1	CF890645, CF890008	Hypothetical
1	AA890871, AA882713	Hypothetical
1	CF890059, AI667958	Hypothetical
1	CF888750, AA676007	Hypothetical
1	CF890193, AA908133	Hypothetical
1	CF889072, CF888688	ubiquitin/ribosomal protein S27
1	CF888156, CF889712	tyrosine aminotransferase
1	CF888478, CF888305	tryparedoxin peroxidase
1	CF889765, CF888327	ribosomal protein L7
1	CF890333, CF888745	ribosomal protein L36
1	CF889747, CF890652	ribosomal protein L35
1	CF887984, CF889802, CF890219	ribosomal protein L34
1	CF888375, CF889089	ribosomal protein L29
1	CF888003, CF890600	ribosomal protein L24
1	AI035047, CF888541	ribosomal protein L18a
1	AI035047, CF888541	ribosomal protein L18a
1	CF890512, CF889311, AA882940	ribosomal protein L17
1	CF887967, CF889407	ribosomal protein L13
1	CF888639, CF889594	ribosomal protein S27
1	AI035119, CF889213	nuclear movement protein
1	CF888449, CF888325	HSP85
1	CF888779, CF889169, CF888419, CF889595	HSP70
1	AI035084, CB923670, CB923671	histone H4
1	CF890229, CF890408, CF890611, CF888049, CF889398	Histone H3
1	CF888115, CF890014	fructose-bisphosphate aldolase
1	CB923566, CF888081	Amastin
1	CF888828, CF888202, CF889558	Retrotransposon
1	CF889016, CF890501	ribosomal protein S6
1	CF890296, AA958070	Hypothetical
1	CF889477, CF888385	Hypothetical
1	AI035079, AI612603, CF890337, CF890370	mucin TcSMUGS
1	CF889637, CF889806, CF889403	Hypothetical
1	CF889503, CF888214	28S ribosomal RNA
1	CF888803, CB923724	Hypothetical
2	CF888342, AI562549	zinc finger protein ZFP1
2	CF888682, AI668092	Hypothetical
2	AA675968, CF889876, CF890283, CF888193	Hypothetical
2	CF888182, CF889720	Hypothetical
2	CF888473, CF888351	Hypothetical
2	CF888152, CF888975	RNA-binding protein
2	CF889168, AI717799	ribosomal protein S7
2	CF889628, CF889491, CF888992	ribosomal protein S5
2	CF887982, CF888539	ribosomal protein S4
2	CF888397, CF889572	ribosomal protein S3a
2	CF889524, CF890419	ribosomal protein S30
2	CF888810, CF889357	ribosomal protein S29
2	AI667889, CF889198	ribosomal protein S25
2	AA399706, CF889943	ribosomal protein S24E

↳

mRNA variant	EST accession nr	Gene
2	CF888451, CF889605	ribosomal protein S23
2	CF888369, CF888467	ribosomal protein S21
2	CF889822, CF888358	ribosomal protein S2
2	AI668059, CF888404	ribosomal protein S17
2	CF889322, AA958211	ribosomal protein S15
2	CF890460, CF889116	ribosomal protein S14
2	AI080896, CF890477, AA441734	ribosomal protein S13
2	CF890528, CF888680	ribosomal protein S12
2	CF889722, CF889378, CF890584	ribosomal protein S11
2	AI668740, CF888018	ribosomal protein P2
2	CF888831, CF888367	ribosomal protein P0
2	CF888749, CF889063	ribosomal protein L7a
2	CF888420, CF890140	ribosomal protein L5
2	CF890636, CF888921	ribosomal protein L38
2	CF890090, CF889315	ribosomal protein L37
2	CF889575, AI717855	ribosomal protein L36
2	CF888809, CF888761	ribosomal protein L31
2	CF890411, AI612552	ribosomal protein L30
2	CF888352, CF888857	ribosomal protein L28
2	CF888860, AA556105	ribosomal protein L27A/L29
2	CF888934, CF888259	ribosomal protein L27
2	CF888060, CF888883	ribosomal protein L27
2	CF889060, CF888144	ribosomal protein L2
2	CF888944, CF888853	ribosomal protein L17
2	CF888664, CF888445	ribosomal protein L14
2	CF890583, CF888135	pre-mRNA cleavage complex II Clp1
2	CF889424, CF890492, CF888948	mucin-associated surface protein
2	CF888858, CF888962	HSP60
2	CF889868, CF888622	DNA-directed RNA polymerase II
2	CF888359, CF889617	casein kinase

1: 5’ UTR exhibiting an alternative internal trans-splicing site; 2: gene presents more than one copy at different genomic loci.