LETTER

Complete Sequence of a 93.4-kb Contig from Chromosome 3 of *Trypanosoma cruzi* Containing a Strand-Switch Region

Björn Andersson,[1,3] Lena Åslund,[1] Martti Tammi,[1] Anh-Nhi Tran,[1] Jörg D. Hoheisel,[2] and Ulf Pettersson[1]

[1]Department of Genetics and Pathology, Biomedical Center, S-751 23 Uppsala, Sweden; [2]Deutsches Krebsforschungszentrum, D69120, Heidelberg, Germany

We have initiated large-scale sequencing of the third smallest chromosome of the CL Brener strain of *Trypanosoma cruzi* and we report here the complete sequence of a contig consisting of three cosmids. This contig covers 93.4 kb and has been found to contain 20–30 novel genes and several repeat elements, including a novel chromosome 3-specific 400-bp repeat sequence. The intergenic sequences were found to be rich in di- and trinucleotide repeats of varying lengths and also contained several known *T. cruzi* repeat elements. The sequence contains 29 open reading frames (ORFs) longer than 700 bp, the longest being 5157 bp, and a large number of shorter ORFs. Of the long ORFs, seven show homology to known genes in parasites and other organisms, whereas four ORFs were confirmed by sequencing of cDNA clones. Two shorter ORFs were confirmed by a database homology and a cDNA clone, respectively, and one RNA gene was identified. The identified genes include two copies of the gene for alanine-aminotransferase as well as genes for glucose-6-phosphate isomerase, protein kinases and phosphatases, and an ATP synthase subunit. An interesting feature of the sequence was that the genes appear to be organized in two long clusters containing multiple genes on the same strand. The two clusters are transcribed in opposite directions and they are separated by an ~20-kb long, relatively GC-rich sequence, that contains two large repetitive elements as well as a pseudogene for cruzipain and a gene for U2snRNA. It is likely that this strand switch region contains one or more regulatory and promoter regions. The reported sequence provides the first insight into the genome organization of *T. cruzi* and shows the potential of this approach for rapid identification of novel genes.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AF052831–AF052833.]

Parasitic disease is a large health problem in developing countries. In South and Central America, 16–18 million people are affected with Chagas' disease, which is often severely debilitating and has a 10% mortality rate [according to the World Health Organization (WHO)]. Chagas' disease is caused by the protozoan parasite *Trypanosoma cruzi,* which is transmitted by insects of the Reduvidae family or by blood transfusions. The drugs that are used currently for therapy against the disease are only partially effective and have severe side effects. A parasite genome initiative was launched by the WHO in 1994 to use genome-scale analysis techniques to rapidly gain further insight into the biology of these organisms and thereby make it possible to find effective therapies against parasitic disease. The *T. cruzi* genome network includes laboratories from South America and several laboratories in North America and Europe.

T. cruzi, together with other kinetoplastids, are interesting for several reasons. They are interesting evolutionarily, since they diverged early in eukaryote evolution. In addition, there are several genetic features that are unique to the kinetoplastid group. *T. cruzi* genes are intronless, but individual genes are *trans*-spliced from polycistronic mRNAs to a splice leader RNA (DeLange et al. 1984; McCarthy-Burke et al. 1989; Donelson and Zeng 1990). Many housekeeping genes are present in large tandemly repeated clusters containing two to >100 copies. The duplications may be involved in the regulation of expression of these genes. The *T. cruzi* genome shows high plasticity and the sizes of the chromo-

[3]Corresponding author.
E-MAIL bjorn.andersson@medgen.uu.se; FAX 46 18 526849.

Table 1. Results of Gene Finding

Search methods	Number of positives
Total number of ORFs larger than 300 bp (starting with ATG)	151
Number of ORFs larger than 700 bp (starting with ATG)	29
GRAIL (human)	38 hits
Database homology	8 ORFs + 1 RNA gene
GRAIL + database homology	8 ORFs[a]
ORFs confirmed by single-pass cDNA sequencing	4[b]

[a]All GenBank-positive ORFs were identified by GRAIL.
[b]One of these ORFs is located in a VIPER repetitive element.

somes vary extensively between strains, and the size of the two homologous chromosomes can also differ greatly within each strain (Cano et al. 1995; Henriksson et al. 1995, 1996). The gene order seems to be conserved despite this variation and the cause of the size variation is unknown. The kinetoplast, the large organelle present in trypanosome cells also shows unique features, the most well known being extensive editing of mRNA molecules. The genome size of *T. cruzi* varies between strains (Dvorak et al. 1982; Thompson and Dvorak 1989). The haploid genome of the CL Brener strain is ~50 Mb.

We have undertaken a project to determine the entire DNA sequence of the third smallest chromosome of the CL Brener reference strain of *T. cruzi* within the *T. cruzi* genome network. We present here the sequence of three overlapping cosmid inserts covering 93.4 kb of sequence from chromosome 3 (GenBank accession nos. AF052831, AF052832, and AF052833). The sequence reveals a potential regulatory region where transcription of long polycistronic messenger RNAs may be initiated. Also, several novel expressed sequences and repeat-rich regions have been identified.

RESULTS

The cosmids were selected from a sublibrary enriched for cosmids containing inserts from chromosome 3 and its homolog (Frohme et al. 1998). The first cosmid (1o17) was chosen as a starting cosmid for a region of the chromosome and the following two cosmids (1b21 and 1m17) were selected on the basis of confirmed overlaps to the already sequenced cosmids. Their localization to the chromosome of interest was confirmed by hybridization to pulsed-field gel-separated *T. cruzi* chromosomes, in which all three cosmids hybridized only to chromosome 3 and its longer homolog (data not shown).

The cosmids were sequenced by a shotgun strategy using M13 subclones. A total of ~1500 sequence reads using the universal primer, 600 reverse reads, and 46 reads using 26 specific primers were required for complete double-stranded coverage of all three cosmid inserts. The length of the inserts was 39 kb for 1o17, 41.3 kb for 1b21, and 35 kb for 1m17. An additional 300 sequence reads were discarded because of poor quality or *Escherichia coli* contamination. The sequence was edited to reach an accuracy of ~99.99%. The overlap between cosmids 1o17 and 1m17 was ~15 kb, whereas the overlap between 1m17 and 1b21 was ~7 kb. The total length of the resulting contig was 93.4 kb. The cosmids were 100% identical in the overlapping regions, confirming that they are from the same region.

The results of gene finding are summarized in Tables 1 and 2, and a map of the contig is shown in Figure 1. A total of 151 open reading frames (ORFs) exceeding 300 bp in length starting with an ATG start codon were identified and 29 of these were longer than 700 bp. To isolate the ORFs most likely to represent real genes, all ORF sequences were used to search GenBank at the DNA and protein levels, which resulted in seven positives, representing a wide variety of genes. In addition, a gene for U2snRNA was identified from a database search using non-ORF regions. In an attempt to identify further genes the computer program GRAIL for human

Table 2. Database Homologies Outside ORFs

Number of sites/position	Homology
1 (33967–33393)	U2 snRNA gene
1 (26500–28807)	VIPER element
2 (28808–28849, 33287–33385)	SIRE element
1 (36984–38744)	TRS 1.6 reverse transcriptase-like repeat element
1 (92642–92795)	partial TRS 1.6 repeat element
3 (8034–7714, 81544–81982, 90531–90963)	novel 400-bp repeat element

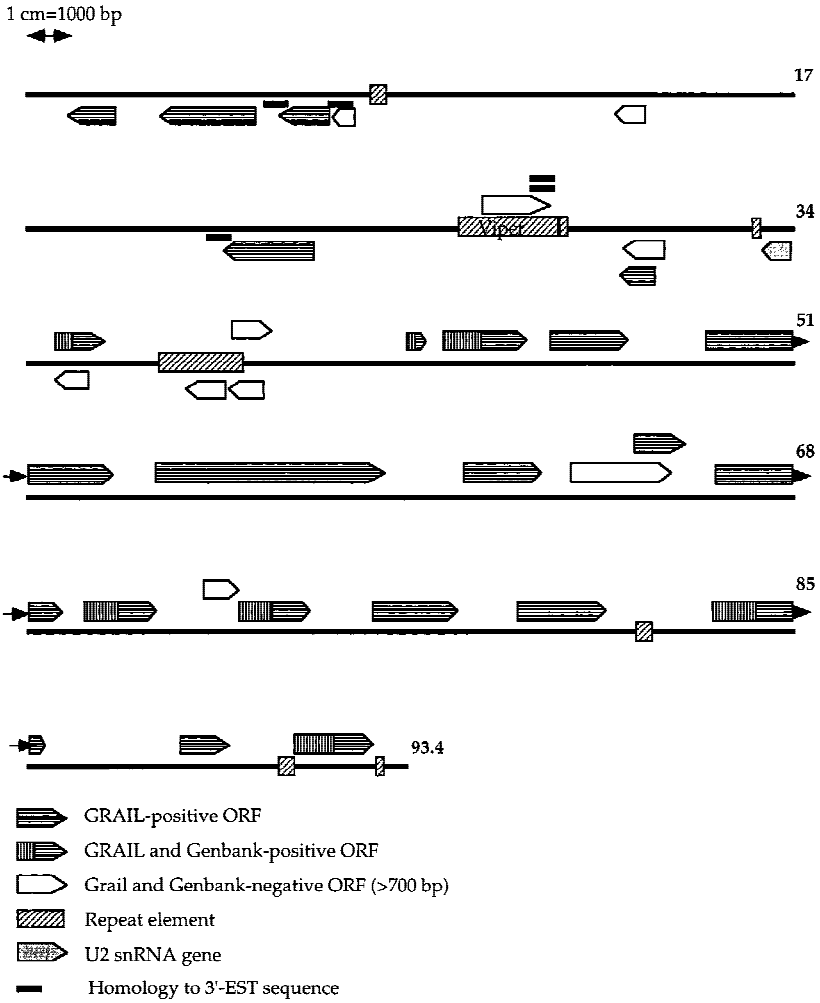

Figure 1 Schematic illustration of the entire contig. Gene candidates and repetitive elements are indicated by arrows and boxes. Each line represents 17 kb of genomic sequence. All ORFs longer than 700 bp have been included, as well as two shorter ORFs confirmed by a database homology and a cDNA sequence, respectively.

DNA was used to identify protein-coding ORFs. A total of 38 ORFs >300 bp in length were identified as protein coding by GRAIL, including the seven ORFs that were confirmed by database searches. For four other ORFs, cDNA clones were identified by hybridization of 1o17 cosmid DNA to filters containing an array of spotted cDNA clones from a normalized *T. cruzi* epimastigote library. Single-pass sequences from the 3′ end of the cDNA inserts were found to match the ORF sequences, thus making it likely that these ORFs are expressed. A total of five cDNA sequences were found to match cosmid sequences in four locations. Two of these ORFs were not identified by GRAIL, including the long ORF in the VIPER repetitive element, to which two cDNA sequences matched.

From approximately base 34,000 of the contig the GRAIL-positive ORFs are present on only the forward strand, whereas the ORFs before base 34,000 are present only on the reverse strand in the contig as shown in Figure 1. This is the case for all ORFs confirmed by database searching or by the presence of corresponding cDNA clones, including one that was GRAIL-negative. All ORFs >700 bp conform to this pattern, except for those that are located within two large repeat elements. This pattern indicates that the genes before base 34,000 are transcribed together on the reverse strand, whereas those after are transcribed together on the forward strand. The region between possibly could contain a bidirectional promoter region or two separate promoter regions, one for each transcriptional unit. The region between the first confirmed genes in each region is ~20 kb in length (Fig. 2). From approximately base 23,000 to base 43,000, there is a relatively high GC content, 54.8%, which can be compared to 48.5% GC for the sequence of the entire contig. Two large repetitive elements were found in this region, an element with homology to the *Trypanosoma brucei* TRS1.6 element and a *T. cruzi* VIPER element that contains a large ORF. In the middle of this region there is a truncated copy of the cruzipain gene, which is present as a tandem cluster in a different area of chromosome 3. The cruzipain gene is likely to be a pseudogene however, since the ORF shows only 290 amino acids compared to the 467 of the previously characterized locus. This gene may therefore not be expressed. Upstream of this locus, at a distance of 845 bp, a U2snRNA gene is located in the opposite direction from the cruzipain pseudogene (Figs. 1 and 2; Table 2). The upstream region of the U2snRNA gene is GC-rich, but shows no discernible similarity to the upstream region of the U2snRNA gene of *T. brucei* (Fantoni et al. 1994). It is unclear whether this represents an actual promoter sequence. This entire region provides several possiblities as to the location of potential promoters and regulatory regions, but no such elements have been identified as yet.

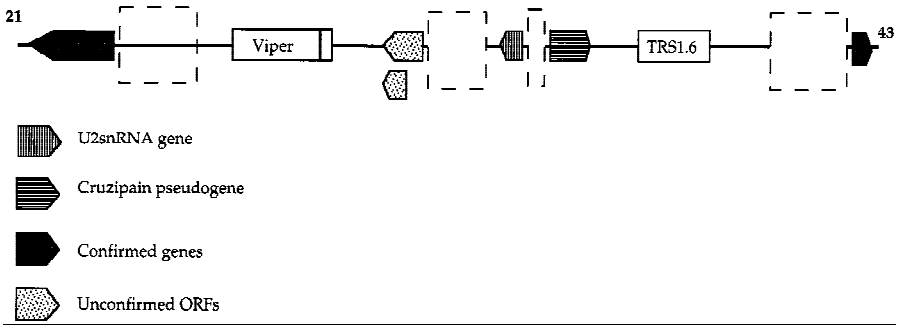

Figure 2 Schematic illustration of the beginning of the two gene rich regions and the sequence between them. Repetitive elements and putative genes are shown as arrows and boxes. The dashed boxes indicate possible locations of regulatory and promoter sequences.

Nucleotide and protein database comparisons using BLAST yielded several significant similarities, which are listed in Tables 2 and 3. The homologies included: the U2snRNA gene mentioned above, which shows almost 100% homology to the database entry, and therefore appears to be a functional gene; two ORFs adjacent to each other showed close to 100% homology to *T. cruzi* alanine aminotransferase as mentioned above; and an ORF showed homology to the cysteine proteinase cruzipain. This truncated locus shows close to 100% homology to the previously published cruzipain sequence in the 5′-untranslated part of the gene and in the 5′ end of the translated region. The homology decreases gradually after 180 amino acids and is completely lost after the premature stop codon. As yet no message from this gene has been identified and the possible function, if any, of this gene is therefore unknown. The ORF sequences in the gene-rich regions were generally more GC rich than the surrounding sequences and often reached GC levels of >60%.

Four ORFs showed homology to known genes in other organisms. These homologies were to tyrosine phosphatases and serine/threonine kinases, an ATP synthase subunit and glucose-6-isomerase from many different organisms, including *Saccharomyces cerevisiae* and *Homo sapiens* as well as several prokaryotes. Blast and BEAUTY searches of the remaining ORFs showed no GenBank homologies to the nucleotide or protein databases, indicating that these represent novel genes. Work is in progress to confirm more genes by identification of the corresponding messenger RNAs by RT–PCR from *T. cruzi* RNA or by further cDNA library screening.

Five sequences showed blast homology to known *T. cruzi* repetitive elements, including the VIPER and TRS1.6 elements mentioned above, two SIRE elements, and one other short interspersed repeat that is also present in the 3′-untranslated region of the gene for *T. cruzi* tyrosine aminotransferase. A 3- to 400-bp sequence was identified that occurs in three different places in the contig. The copies are very well conserved with only a few base pairs differing between each copy. The location of the copies are at approximately positions 7500, 82,000, and 91,000. The orientation of the two latter copies is the same, whereas the first copy is in the opposite orientation, which is consistent with the direction of the ORFs present in the same regions. These sequences are not located within ORFs and they show no homology to known sequences in GenBank. It has been found by hybridization of this sequence to separated chromosomes, that these sequences are only present on chromosome 3 and its homolog. The function of this sequence is unknown. Further studies of this repeat element are in progress.

DISCUSSION

The sequencing of these cosmids provides the first *T. cruzi* genomic sequence longer than a few kilobase pairs. Genome sequencing has been initiated recently in other parasites, mainly *Plasmodium falciparum*, in which large amounts of unfinished sequence have been submitted to the public databases, and in chromosome 1 of *Leishmania major*, in which unfinished and finished data from several cosmids have been submitted. The genomic sequence of *P. falciparum* is very AT rich and quite different from that of *T. cruzi*, whereas preliminary data from *L. major* shows more similarities to *T. cruzi*.

T. cruzi is expected to have 5000–10,000 genes, and estimates of the size of the haploid genome of the CL Brener strain varies between 45 and 50 Mb. A rough estimate of the gene density would thus be one gene per 4500 bp and that the current sequence should contain ~20 genes. There are between 25 and 30 strong gene candidates in this sequence based on GRAIL searches and the length of ORFs. A total of 11 genes outside repetitive elements have been confirmed by database searching and identification of cDNA clones. The total number of genes therefore appears to be slightly higher than 20. The variation

Table 3. Possible Genes

Location	Length (bp)	Confirmation	Identity
53756–58912	5157	—	—
49164–52877	3714	—	—
66347–68791	2445	—	—
63168–65339	2178	—	—
83353–85467	2115	database homology	protein phosphatase
4970–2859	2112	—	—
23287–21299	1989	3′-cDNA sequence	—
75662–77563	1902	—	—
79058–80860	1803	—	—
43293–45047	1755	database homology	G-6-P isomerase
90872–92554	1683	database homology	protein kinase
60672–62312	1641	—	—
45694–47331	1638	—	—
69244–70737	1494	database homology	alanine aminotransferase
27365–28642	1278	database homology/ 3′-cDNA sequence	VIPER element
72919–74148	1230	database homology	alanine aminotransferase
1920–829	1092	—	—
6651–5599	1053	3′-cDNA sequence	—
88432–89436	1005	—	—
64663–65649	987	—	—
34713–35681	969	database homology	truncated cruzipain
31270–30308	963	—	—
38430–37552	879	—[a]	—[a]
38587–39459	873	—[a]	—[a]
71994–72782	789	—	—
13865–13101	765	—	—
30998–30258	741	—	—
35453–34731	723	—[b]	—[b]
39308–38592	717	—[a]	—[a]
7208–6720	489	3′-cDNA sequence	—
42461–42862	402	database homology	ATP synthase subunit

Possible genes are ORFs longer than 700 bp or otherwise confirmed.
[a]Overlaps TRS1.6 repeat element.
[b]Overlaps truncated cruzipain gene on the other strand.

in gene density in the *T. cruzi* genome is currently unknown, and it remains to be seen how representative this region is of the genome.

It has been shown that *T. cruzi* has polycistronic pre-mRNAs and it is therefore expected that genes would often be positioned close to each other on the same strand. This pattern can be seen in the region presented here. The contig appears to contain two separate gene regions separated by a region with fewer gene candidates. One of the gene-rich regions covers >50 kb and contains at least 17 candidate genes. It is unknown how much further this region continues outside the contig. The two gene-rich regions are on opposite strands. The region between the two gene rich areas is GC rich and contains a VIPER and a TRS1.6 repetitive element as well as a probable pseudogene and an RNA gene, U2snRNA. No promoter sequences have been identified previously in *T. cruzi,* other than for the spliced leader RNA (Nunes et al. 1997) and for rRNA genes (Janz and Clayton 1994), but it is likely that this region contains at least one such regulatory element. The precise location of the putative promoters is not known, but there are several possibilities. There are two confirmed gene sequences between the major gene-rich regions, but one is likely to be a pseudogene, whereas the other is the gene for U2snRNA. In *T. brucei* it has been shown that *U2snRNA* is transcribed by RNA polymerase III (Fantoni et al. 1994). Since the protein-coding genes are expected to be transcribed by RNA polymerase II, it is likely that the *U2snRNA* gene has its own regulation and is not a promoter for a whole cistron. It is more likely that there are two separate regulatory regions, one at the start of each gene cluster. The polII promoters that have been characterized previously in trypanosomes are in the upstream sequences of the actin gene (Ben Amar et al. 1991) and the hsp70 locus (Lee 1996) of *T. brucei.* Also, some data indicates that the spliced leader RNA gene (Nunes et al. 1997), where the promoter has been characterized in *T. cruzi* may be transcribed by polymerase II, but this is not entirely clear. Ths promoter shows a simi-

lar organization to other processor RNA genes in different organisms. The *T. brucei* actin promoter is located 4 kb upstream of the actin genes and appears to initiate transcription for a whole cistron, but it shows no similarities to known promoters and no clear features. It is therefore not surprising that no promoter sequences have been found in the present sequence by comparison to known sequences. This makes the upstream sequences of both gene regions good candidates for containing regulatory and promoter sequences. The *T. brucei hsp70* promoter shows a different organization in that it is associated closely with the gene and that it is also present between each tandem copy of the gene. It is possible that the intergenic sequences of the cruzipain locus contains similar promoter regions. Transfection with a clone containing several copies of the cruzipain gene leads to overexpression of the gene (Tomas and Kelly 1997). Since the 5′-end and upstream sequence of the truncated cruzipain gene in the contig sequence is well conserved there is a possibility that this locus could be involved in transcriptional regulation.

Regulatory and promoter sequences in repetitive elements may be involved in regulation of *T. cruzi* genes (Requena et al. 1996). The two large repeat elements present in this region both contain large ORFs, which may be transcribed. EST sequences have been found that correspond to both types of elements and we identified cDNA clones corresponding to the ORF in the VIPER element. Thus, the repeat elements, together with several other sequences in this region are good candidates for containing promoters and other regulatory elements, and the region described in this paper is a good candidate for further studies of gene expression in *T. cruzi*.

The plasticity of the *T. cruzi* genome is high and it is therefore of great interest to study the presence of repeated sequences in the intergenic regions and the relative positions of genes for future comparative studies. In the sequence presented here, a large region that appears to contain relatively few genes is present in the region between 10 and 20 kb in the sequence. This area contains regions of 2–3 kb each, which are extremely rich in di- and trinucleotide repeats. The extent of variation between strains and between homologous chromosomes in this and other regions is unknown currently. The sequenced cosmids all hybridize equally strongly to chromosome 3, which is ~620 kb in length and to the homologous chromosome which is ~1 Mb. The cosmids could thus come from either chromosome. The hybridization data coupled with the fact that the cosmids showed no differences in the overlapping regions, a total of ~22 kb, indicates either that all three cosmids come from the same chromosome, or that this region is close to identical in chromosome 3 and its homolog.

An additional source of variation is the number of copies present of certain genes. As mentioned above, housekeeping genes in *T. cruzi* are often present in multiple copies in tandem. The number of copies in a tandem cluster has been found to vary from 2 to >50 in the clusters studied (Campetella et al. 1992; B. Andersson, unpubl.). In the current region one gene, *alanine aminotransferase*, is present in two tandem copies. It has been hypothesized that the number of copies correlates with the expression levels of the gene. It is unknown whether the gene for alanine aminotransferase is present in other loci in the *T. cruzi* genome and also if the two copies in this locus are both functional. One possibility is that one of the copies is a pseudogene. The fact that the ORF of one of the copies is 264 bp shorter than in the other speaks in favor of this hypothesis.

A different pattern was observed for a 300-bp sequence that was detected in three different locations in the contig and not in a cluster. It is possible that this is a novel repeat element, or that these are pseudogenes. The reason for its presence on only this chromosome currently is not known. The nature of this element and its distribution along the chromosome and in the genome is under further investigation.

A relatively small number of expressed sequences have been characterized previously from *T. cruzi*. Currently, efforts to produce ESTs are under way, but only a small number are available as yet. Because the kinetoplastids diverged early in evolution from other organisms, most genes may be quite different from their equivalents in other organisms. It is therefore expected that a large fraction of genes can not be identified by homology searches. Protein-coding sequences in *T. cruzi* can, however, often be identified easily using other methods. *T. cruzi* genes have no introns and protein-coding genes can therefore be identified as long ORFs. In addition, the genes tend to have a higher GC content compared to surrounding sequences. Surprisingly, we found that GRAIL for human genes could be used to identify possible genes in *T. cruzi* by its ability to identify ORFs and by codon usage, even though codon usage in humans is different from that in *T. cruzi* for a few amino acids. GRAIL was able to detect all but one of the confirmed genes and also several of the other long ORFs that are likely to be expressed. Thus, GRAIL and other programs that de-

tect codon bias will be of use for gene prediction in *T. cruzi.*

This work only represents ~0.2% of the *T. cruzi* genome, but it has already yielded some information about the organization of genes on chromosome 3 and also several novel genes that may be used in the future as drug targets and in projects to better understand the biology of *T. cruzi.* The sequencing of the remainder of chromosome 3 is ongoing and the continued characterization of novel genes and other genetic features in *T. cruzi* within the genome project will generate many insights into parasite biology and possibilities for identification of drug targets to combat Chagas' disease.

METHODS

Selection of Target Cosmids

The three cosmids were selected from a sublibrary of a whole-genome *T. cruzi* cosmid library (Hanke et al. 1996). The sublibrary was isolated by hybridization to chromosome 3, DNA-purified using pulsed-field gel electrophoresis, performed as described by Henriksson et al. (1995, 1996). The cosmids were selected by showing that they hybridize specifically to chromosome 3 and its homolog on Southern-blotted, pulsed-field gel electrophoresis separated chromosomes, and by using restriction patterns from multiple cosmids in the region and end probes from each cosmid to show overlaps. Cosmid overlaps were confirmed by using end probes from sequenced cosmids and by comparison of restriction patterns.

Construction of M13 Shotgun Libraries

Cosmid DNA was purified using a Qiagen midi-prep kit and sheared by nebulization to an average size of ~2 kb. The random fragments were cloned into a modified M13 vector using the ''double adaptor'' method as described (Andersson et al. 1996b).

M13 Template Preparation

M13 clones were grown in a 96-well format and high-quality DNA templates were prepared using a 96-well glass-fiber filter protocol as described by Andersson et al. (1996a).

DNA Sequencing And Sequence Assembly

Fluorescent, automated DNA sequencing was performed using Applied Biosystems 377 DNA sequencers (Perkin Elmer) and sequencing reagents from Perkin Elmer and Amersham, using automated fluorescent methods (Smith et al. 1986). The sequence reads were assembled and the contiguous sequences edited using the program gap4 in the Staden package (Staden 1996) and the PHRED-PHRAP package (courtesy of P. Green, University of Washington, Seattle). Gap closure and finishing was carried out using a mapped-gap strategy (Edwards and Caskey 1991; Muzny et al. 1994; Richards et al. 1994) and walking using specific oligonucleotide primers.

Sequence Analysis

Expressed sequences were identified by database searches using BLAST (Altschul et al. 1990) and by the program GRAIL (Uberbacher and Mural 1991). Searches for further identification of gene function were performed using BEAUTY (Worley et al. 1995).

cDNA Hybridization

More than 16,000 cDNA clones derived from a normalized epimastigote library were spotted onto nylon membranes and screened using radioactively labeled (Megaprime, Amersham) cosmid DNA. Hybridization was carried out in 0.5 M Na-phosphate at pH 7.2, 7% SDS, 1 mM EDTA, and 100 μg/ml yeast tRNA at 65°C overnight. The filters subsequently were washed in 40 mM Na-phosphate at pH 7.2, 0.1% SDS, at 65°C, and exposed to X-ray film.

ACKNOWLEDGMENTS

This work was funded by grants from the Beijer Foundation, the Knut and Alice Wallenberg Foundation, the Swedish Medical Research Council (K98-16X-12633-01A), and by the United Nations Development Programme/World Bank/WHO Special Programme for Research and Training in Tropical Diseases. We thank Edson Rondinelli for kindly providing the cDNA library.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Andersson B., J. Lu, K.E. Edwards, D.M. Muzny, and R.A. Gibbs. 1996a. A method for 96-well M13 DNA template preparations for large scale sequencing. *BioTechniques* 20: 1022–1027.

Andersson, B., M.A. Wentland, J.Y. Ricafrente, W. Liu, and R.A. Gibbs. 1996b. A ''double adaptor'' method for improved shotgun library construction. *Anal. Biochem.* 236: 107–113.

Ben Amar, M.F., D. Jeffries, A. Pays, N. Bakalara, G. Kendall, and E. Pays. 1991. The actin gene promoter of *Trypanosoma brucei. Nucleic Acids Res.* 19: 5857–5862.

Campetella, O., J. Henriksson, L. Åslund, A.C.C. Frasch, U. Pettersson, and J.J. Cazzulo. 1992. The major cysteine proteinase (cruzipain) from *Trypanosoma cruzi* is encoded by multiple polymorphic tandemly organized genes located on

different chromosomes. *Mol. Biochem. Parasitol.* 50: 225–234.

Cano, M.I., A. Gruber, M. Vazquez, A. Cortez, M.J. Levin, A. Gonzalez, W. Degrave, E. Rondinelli, B. Zingales, J.L. Ramirez, C. Alonso, J.M. Requena, and J.F. da Silveira. 1995. Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. *Mol. Biochem. Parasitol.* 71: 273–278.

De Lange, T., T.M. Berkvens, H.J.C. Veerman, A.C.C. Frasch, J.D. Barry, and P. Borst. 1984. Comparisons of the genes coding for the common 5′ terminal sequence of messenger RNAs in three trypanosome species. *Nucleic Acids Res.* 12: 4431–4443.

Donelson, J.E. and W. Zeng. 1990. A comparison of trans-RNA splicing in trypanosomes and nematodes. *Parasitol. Today* 6: 327–334.

Dvorak, J.A., T.E. Hall, M.S.J. Crane, J.C. Engel, J.P. McDaniel, and R. Uriegas. 1982. *Trypanosoma cruzi:* Flow cytometric analysis, I. Analysis of total DNA/organism by means of mithramycin-induced fluorescence. *J. Protozool.* 29: 430–437.

Edwards, A. and C.T. Caskey. 1991. Closure strategies for random DNA sequencing. *Methods: Companion Methods Enzymol.* 3: 41–47.

Fantoni, A., A.O. Dare, and C. Tschudi. 1994. RNA polymerase III-mediated transcription of the trypanosome U2 small nuclear RNA gene is controlled by both intragenic and extragenic regulatory elements. *Mol. Cell. Biol.* 14: 2021–2028.

Frohme, M., J. Hanke, L. Åslund, U. Pettersson, and J.D. Hoheisel. 1998. Selective generation of chromosomal cosmid libraries within the *Trypanosoma cruzi* genome project. *Electrophoresis* 19: 478–481.

Hanke, J., D. Sanchez, J. Henriksson, L. Åslund, C. Häussler, U. Pettersson, A.C.C. Frasch, and J. Hoheisel. 1996. Mapping the *Trypanosoma cruzi* genome: Analysis of representative cosmid libraries. *BioTechniques* 21: 686–693.

Henriksson, J., B. Porcel, M. Rydåker, A. Ruiz, J.J. Cazzulo, A.C.C. Frasch, and U. Pettersson. 1995. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi. Mol. Biochem. Parasitol.* 37: 64–73.

Henriksson, J., L. Åslund, and U. Pettersson. 1996. Karyotype variability in *Trypanosoma cruzi. Parasitol. Today* 12: 108–114.

Janz, L. and C. Clayton. 1994. The PARP and rRNA promoters of *Trypanosoma brucei* are composed of dissimilar sequence elements that are functionally interchangeable. *Mol. Cell. Biol.* 14: 5804–5811.

Lee, M.G. 1996. An RNA polymerase II promoter in the hsp70 locus of *Trypanosoma brucei. Mol. Cell. Biol.* 16: 1220–1230.

McCarthy-Burke, C., Z.A. Taylor, and G.A. Buck. 1989. Characterization of the spliced leader genes and transcripts in *Trypanosoma cruzi. Gene* 82: 177–189.

Muzny, D.M., S. Richards, Y. Shen, and R.A. Gibbs. 1994. PCR based strategies for gap closure in large scale sequencing projects. In *Automated DNA sequencing and analysis techniques* (ed. J.C. Venter), pp. 182–190. Academic Press, Orlando, FL.

Nunes, L.R., M.R.C. Carvalho, A.M. Shakarian, and G.A. Buck. 1997. The transcription promoter of the spliced leader gene from *Trypanosoma cruzi. Gene* 188: 157–168.

Requena, J.M., M.C. Lopez, and C. Alonso. 1996. Genomic repetitive DNA elements of *Trypanosoma cruzi. Parasitol. Today* 12: 279–283.

Richards, S., D.M. Muzny, A.B. Civitello, F. Lu, and R.A. Gibbs. 1994. Sequence map gaps and directed reverse sequencing for the completion of large sequencing projects. In *Automated DNA sequencing and analysis techniques* (ed. J.C. Venter), pp. 191–197. Academic Press, Orlando, FL.

Smith, L.M., J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connel, C. Heiner, S.B.H. Kent, and L.E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321: 674–679.

Staden, R. 1996. The Staden sequence analysis package. *Mol. Biotechnol.* 5: 233–241.

Thompson, C.T. and J.A. Dvorak. 1989. Quantitation of total DNA per cell in an exponentially growing population using the diphenylamine reaction and flow cytometry. *Anal. Biochem.* 177: 353–357.

Tomas, A.M. and J.M. Kelly. 1997. Stage-regulated expression of cruzipain, the major cystein protease of *Trypanosoma cruzi* is independent of the level of RNA. *Mol. Biochem. Parasitol.* 76: 91–103.

Uberbacher, E.C. and R.J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* 88: 11261–11265.

Worley, K.C., B.A. Wiese, and R.F. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* 5: 173–184.

Received March 29, 1998; accepted in revised form June 26, 1998.