Original Research Article,

Anti-cytoskeleton immunoscreening of *Trypanosoma brucei* expression library reveals novel immunogenic conserved putative proteins

Abstract

Background

The overall shape of the trypanosome is defined by an internal cytoskeleton consisting of a network of microtubules that are cross linked both to each other and the inner face of the plasma membrane. However, the total compliment and identity of the trypanosome cytoskeleton proteins are not yet fully determined despite the fact that some of them may be good targets for diagnostics, drugs and/or vaccines discovery.

Methods

In this study, rabbit anti-*Trypanosoma brucei* detergent insoluble cytoskeleton sera were produced *in vivo* and used to probe a *T. brucei* expression library. The picked plaques were made clonal by a series of library screening followed by PCR amplification, cDNA sequencing and identification of the proteins coded by these sequences using BLAST.

Results

The previously well-known cytoskeleton proteins (paraflagella rod protein and histone H2B), putative cytoskeleton proteins (Dynein light chain and nucleoporin), conserved hypothetical protein (Tb10.61.2430) and novel cytoskeleton protein coding cDNA sequences (not in the sequenced and published *T. brucei* genome) were identified in this study.

Conclusion

This approach is therefore, useable in the search for novel proteins whose utility in the design and development of diagnostics, drugs and/or vaccines can further be studied.

Key words: Antibodies, cDNA sequencing, cytoskeleton, expression library, immune sera, immunoscreening, PCR, proteins, *Trypanosoma brucei*.

Introduction

Trypanosoma brucei is a protozoan parasite causing human African trypanosomiasis (HAT) prevalent in 36 sub-Saharan Africa countries with 2804 reported cases in 2015 [1]. HAT is transmitted through a bite of tsetse fly of genus Glossina affecting mostly the poor populations living in rural communities in Sub Saharan Africa and is fatal if left untreated. There is currently no commercial vaccine against African trypanosomiasis and the current drugs have been in use for a long time with some of them showing adverse side effects, melarsoprol killing about 5% of the patients and parasites developing resistance [2, 3].

Trypanosome cytoskeletal proteins have been targeted for a possible drug or vaccine candidate. The interest in cytoskeletal proteins is due to the vital roles they play in this hemoflagellate parasite. Such roles include motility and intracellular trafficking of proteins and organelles, which are vital for the survival and pathogenesis of *T. brucei* [4, 5]. The most prominent features in *T. brucei* cytoskeleton are the subpellicular corset of microtubules and the flagellum. Besides microtubules and the flagellum, various other trypanosome proteins such as microtubule-associated repetitive proteins (MARPs) and cytoskeletal-associated proteins (CAP) have been associated with the cytoskeleton due to the vital roles they play in maintaining its integrity [6]. However, studies on these cytoskeletal proteins have not yet yielded any commercial vaccine candidate. Purification of individual trypanosome-cytoskeleton proteins is difficult because they are prone to aggregation and getting lost in the pellet upon centrifugation. Studies done decades ago demonstrated that cytoskeleton proteins can be identified using monoclonal antibodies, however, this is expensive and time consuming [7]

In this study therefore, we demonstrated the potential of immunoscreening of cDNA (Complementary DNA) expression library for the identification of novel trypanosome cytoskeletal proteins that may be vaccine or drug targets by raising poly-specific immune serum against *T.brucei* detergent-insoluble cytoskeleton and using it to screen a λ ZAP *T. brucei* expression library. This approach circumvents the need for monoclonal antibodies and is cheaper compared to the use of monoclonal antibodies.

2.0 Methods

2.1 Anti-detergent insoluble cytoskeleton poly-specific immune sera screening of a *λ ZAP T. brucei* expression library

Poly-specific immune sera raised against whole *T.brucei* detergent-insoluble cytoskeleton was used to screen a λ ZAP *T. brucei* expression library.Competent XL1-Blue cells were transformed with 2 µl of the phage library. The cells were left to grow on sterile agar plates at 42ºC until plaques appeared after 4 hours. A nitrocellulose membrane soaked in 10mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) was overlaid onto the plates and incubated further for 4 hours at 37ºC. The expressed proteins were impregnated onto the membrane and probed with the anti-cytoskeleton immune sera (1:1000). The positive colonies were identified by incubating the membrane with anti-rabbit IgG (1:10000) horseradish peroxidase (HRP) and later developed in a developing solution (60mg of 4-Chloro-1-napthol dissolved in 20ml ice cold methanol, just before use; mixed with 100ml of 10mM Tris-Cl (PH 7.5), 150mM NaCl containing 60µl of 30% H_2O_2). Positive plaques that showed up as dark purple rings on the membrane were stabbed out with a wide bore pipette from the agar plates and put into 500 µl SM buffer in 1.5 ml tube. Each positive colony was re-screened until all the plaques on the plate were being recognized by the immune sera. The clonal plaques were picked and re-suspend in 200µl SM buffer for PCR (polymerase chain reaction) analysis.

2.3 PCR amplification of the clonal plaques and cDNA sequencing

Commercially available T7 and T3 (New England Biolabs) were used as forward and reverse PCR primers respectively. Amplification reactions included (100µl final volume) 5µl template, 2µl 10mM dNTPs (Deoxyribonucleotide triphosphates), distilled water, 1µl of each primer and 1.0U DNA (Deoxyribonucleic acid) polymerase. PCR (Polymerase Chain Reaction) was performed as 35 cycles with denaturing at 95°C for 2 min, annealing at 52°C for 1 minute, extension at 72°C for 2 minutes and finished with a final cycle of 72°C for 7 minutes using the Qiagen PCR kit. The amplicons were thereafter analyzed using agarose gel electrophoresis. Clonal plaques were identified and confirmed after those picked from the same plate using immune sera gave single and same size bands after running the PCR amplicons on a 2% agarose gel. Positive clonal amplicons (>200 bp) were then purified using the Qiagen PCR purification

kit and sent for commercial sequencing to MWG Biotech, Germany and Segoli sequencing facility at ILRI, Nairobi Kenya for comparison and reliability purposes.

2.3 Basic Local Alignment Search Tool (BLAST) sequence analysis

Sequences of the clones were obtained by long reads from one or both ends performed by a commercial provider (MWG Biotech, Germany). Prior to homology search, each sequence was screened for vector contamination using the National Center for Biotechnology Information (NCBI)-Vec screening program to ensure that each sequence contained only pure trypanosome-derived sequences. Sequence segments that belonged to the λ ZAP expression vector such as adapters, linkers and primers were thereafter deleted. Contigs were built and corrected, where possible, by matching to the *T.brucei* genome project database at http://www.ebi.ac.uk. The resulting sequences were used to query the non-redundant protein and DNA databases at NCBI, NIH, Bethesda, USA and EMBL (Hinxton) using blastn (BLAST on nucleotide query vs. nucleotide database), blastx (BLAST on translated nucleotide query vs. protein database) and tblastx (BLAST on translated nucleotide query vs. translated database) and omniBLAST algorithms.

The database queried contained version 4 genome dataset of *T.bruc*ei s927 available at ftp:/ftp.sanger.ac.uk/pub/databases/T.brucei_sequences/T.brucei_genome_v4.

Blastn which compares a nucleotide sequence query against a nucleotide database and blastx which compares the six-frame conceptual translation products of both strands of the query nucleotide sequences against a protein database were used to identify homologues in GeneDB and NCBI data bases. On the other hand tblastx which compares the six-frame translations of a nucleotide query sequence against the six-frame translations of nucleotide sequences in databases was used for identifying novel genes in error prone nucleotide query sequences. The omniBLAST search done on a set of nucleotide databases (blastn and tblastx or tblastn) available in GeneDB was used to characterize a number of sequences that couldn't give any significant BLAST hit using the above conventional BLAST algorithms. Furthermore, sequence alignment

for DNA or proteins was first performed using ClustalW and then ClustalX and T-coffee to ensure that reliable results were obtained. In order to identify the transmembrane domains and thus prediction of the probable location and function of the encountered putative trypanosome cytoskeletal proteins, the*T.brucei* genome was analyzed using the Hidden Markov Modeling (TMHMM) v2.0 available from http://www.cbs.dtu.dk/services/TMHMM/

3. 0 Results and discussion

3.1.0 Immunoscreening of a *T. brucei* bloodstream form expression library

In order to identify the antigens represented in the raised anti-cytoskeletal sera, the polyclonal-polyspecific sera were used to screen a λ ZAP *T. brucei* bloodstream form expression library. Anti-*T.brucei* detergent insoluble cytoskeleton antibodies generated in rabbits detected many positive clones during the first round of immune-screening, all of which scored negative when screened with control pre-immune rat serum or the secondary detecting antibody. The positive clones were recovered, purified by dilution and re-screened twice until homogeneously positive. Only 103 clones that were positively detected by the anti-cytoskeleton screening were further PCR amplified and sequenced. Bioinformatics analysis of the cDNA sequences against web-based databases (GenBank and GeneDB) identified 31 of the 103 inserts (Table 1) as trypanosome genes. The other remaining 72 cDNA sequences did not find any hit with the various BLAST algorithms used, however, OmniBLAST registered hits with either trypanosome contigs or ESTs (expressed **sequence** tag).

3.1.1 Well identified and studied cytoskeleton proteins

The paraflagella rod protein C (PFR-C) and histone H2B (H2B) were the only two well-known and studied cytoskeleton proteins identified in this study. Three cDNA sequences matched PFR-C with 97% identities on the histone shock protein (HSP); aligning from nucleotide (nt) 1-345 on the query with nt 453-797 on the subject, e-value of 6.2e-71 (Figure 1) on the established

trypanosome cytoskeleton protein, PFR-C [8]. This protein is located on *T. brucei* chromosome 3, has a permanent systematic name, is experimentally characterised and has orthologues in *T. cruzi* and *L. major* [9]. It is functionally a structural component of the microtubule-based flagellum involved in cell motility and calmodulin binding [10, 11, 12]. It is a known component of the trypanosome detergent-insoluble cytoskeleton [13] encoded by a cluster of similar tandamely arranged genes and unknown number of minor proteins [14, 15] and present in all life cycles except in the amastigote [13, 14]. On the other hand, sequences from 4 other isolated clones matched H2B with 100% identities from nt 66-404 on the query against nt 1-339 of the subject on HSP, e-value of 5.1e-72 (Figure 2). Information obtained from the database shows that H2B plays a role in chromosome organization and biogenesis, functions in DNA binding and belongs to the nucleosome core. It was earlier described as a specific identifiable remnant of the nucleus retained in its original cellular position in studies involving cytoskeleton preparations [7, 12, 16] thus confirmed in this study as a definitive component of the trypanosome detergent insoluble cytoskeleton.

3.2.2 Putative cytoskeleton proteins

Dynein light chain (DLC) and nucleoporin protein (NP) were the two putative cytoskeleton proteins identified in this study.Fifteen cDNA sequences obtained from 15 independent clones hit the DLC with 99% identities and an e-value of 2.9e-124. The alignment occurred from nt 38-608 on the query to nt 1-571 on the subject in a HSP (Figure 3). This protein has a systematic name Tb10.389.0060 in the GeneDB and GenBank accession number XM.822729 [8, 17]. It is located on chromosome 10. In order to confirm the result of the BLAST hit for all the fifteen isolates, clustalW multiple alignments was done. The results obtained (alignment not shown) indicated that the sequences of these isolates were highly identical thus in agreement with the uniform BLAST results. This protein belongs to the cytoplasmic dynein complex catalyzing

movement along a cytoplasmic microtubule and functions as a microtubule motor (www.GeneDB.org). The dynein light chain [Tb11.02.3390] gene [18] is completely different from the one isolated in this study. The latter is located on chromosome 10 whereas the former is on Chromosome 11. A ClustalW alignment of the DLC with the well characterized [Tb11.02.3390] (Alignment not shown) showed that they are much diverged both at the nucleotide and amino acid level suggesting that it's a novel paralog. The nucleotide sequence similarity between them is only 3% but they are predicted to perform related functions. They are able to perform similar function by virtue of similar three-dimensional domains they constitute upon folding. Furthermore, motif and domain analysis of the cDNA of the DLC classified it among the Tctex-1 family members; a feature characteristic of all dynein chains. Meanwhile a cDNA insert isolated five times during different rounds of screening hit NP with 99% identities on the query from nt 4-552 against the subject's nt 4114-4662, e-value 3.5e-120 (Figure 4). This protein has a GeneDB systematic id [Tb11.03.0140] [14]. It was also in the GenBank with accession number [XM 823098.1], [10, 18] and has orthologues in *T.cruzi* and *L.major*. It's a major component of the nuclear-pore complex in eukaryotic cells and plays a vital role in the transport of cytoplasmic components especially mRNA and proteins across the nuclear envelope [19]. It was also encountered in *T. brucei* bloodstream plasma membrane fraction by fractionation techniques [12]. This suggests that it is a definitive protein.

3.1.3 Conserved hypothetical protein Tb10.61.2430 (CHP)

Two independently isolated cDNA inserts matched the conserved hypothetical protein (CHP) upon BLAST search in the GeneDB. It also matched the same hypothetical protein as the top hit in the NCBI GenBank data base on the HSP (Figure 5). This protein has no signal peptide suggesting that it may not be a membrane associated protein. It has no transmembrane domain

and no GPI (Glycosylphosphatidylinositol) anchors implying that it's not bound to the non-cytoplasmic surface of the membrane by an anchor. It was difficult to assess CHP in details as no functional insights about it could be found in the queried databases. However, by performing reciprocal BLAST analysis, it was possible to gauge its relative uniqueness based on whether homologues could be identified in any other organisms. Its orthologues were identified in the closely related trypanosomatids, *Leishmania major* and *T. cruzi.* The assigned conserved hypothetical protein status thus arises from its close relatedness to the sequences in *T. cruzi* and *Leishmania* genomes.CHP was also encountered in *T. brucei* fractions enriched with plasma membrane and cytoskeleton proteins hitherto defined only as a conserved hypothetical protein [12], thus confirming it as a definitive *T. brucei* protein.

3.1.4 *Trypanosoma brucei* chromosome 3 clone RPCI93-48O8 (RPC)

Sequences of 2 isolates hit clone RPC with 99% identities (HSP) from nt 1-241 on the query against nt 14012-14252 of the subject and an e-value of 3e-147 (Figure 6) containing a sequence of *T. brucei* located on chromosome 3. RPC was also generated during the partial sequencing project of the trypanosome genome and directly submitted to The Institute for Genomic Research, 9712, Medical Center Dr, Rockville, MD 20850, USA, [20, 21]. It was given Accession number [AC091330.18] but was never annotated or assigned to any gene category, contig or EST. A translated amino acid sequence of the above clone contained two transmembrane domains, a possible indication of membrane localization but one is unable to conclude whether it is a *bona fide* plasma membrane protein intimately associated with the subtending cytoskeleton or an integral membrane protein. However, the gene product of RPC is real and may as well be a cytoskeleton protein associated with the plasma membrane. In a search for more information on this sequence (sequence that matched the above chromosome clone),

primers were designed basing on the obtained cDNA clone. PCR amplification using *Trypanosoma brucei* genomic DNA was successful and after sequencing the purified amplicons, the resulting sequence was identical to the original clone upon alignment. This is an indication that the above cDNA insert exists in the trypanosome and was not a foreign contaminant of the cDNA library. However, further studies need to be done on it so as to determine its location, function(s) and expression levels among others.

3.1.5 Uncharacterisable isolate.

A total of seventy two cDNA sequences from independent clones obtained during the same and various screening sessions were identical upon multiple sequence alignment therefore, it was the same clone being encountered so many times despite the different recognition criteria employed in picking them from the screened library. Apparently, after a series of homology searches against NCBI non-redundant or trypanosome specific GeneDB databases, T.b.gambiense DAL972 chromosome 9, complete sequence was obtained as the most significant hit with a GenBank accession number FN554972.1 (Table 2). However, an OmniBLAST search against the trypanosome GeneDB matched them with *Trypanosoma brucei* contigs; an indication that the queried sequences were sections of contiguous sequences of DNA created by assembling overlapping sequenced fragments of the trypanosome chromosome. Also, omniBLAST search hit ESTs, further ruling out the possibility of the sequences being contaminants. Since ESTs are neither annotated nor fully sequenced, there are no corresponding protein translations in the protein databases, a fact which puts this sequence into a novel trypanosome conserved putative protein category that had not been submitted to any curated data bases. The sequences which may exist as single or multiple copies may be denoted as specific orphan sequences since they cannot be linked to any other sequences submitted in the databases and cannot be assigned any

function. One interesting feature is that all these sequences had a polyA tail at their 3' end, a characteristic of mature and complete mRNAs encoding proteins. Since polyadenylation is a post transcription modification in trypanosomes and other eukaryotes, the presence of the polyA tail on all these sequences shows that the sections sequenced were C-termini of the expressed cDNA sequences. However, this study wasn't able to obtain the complete ORFs thus further studies are needed to deduce the size and substantiate on the identity of these orphan sequences.

It may be queried why the major trypanosome cytoskeleton proteins such as tubulin were not encountered in this study. In an attempt to answer that question, the bacteriophage lambda ZAP expression library was separately probed using anti-α tubulin and anti-β tubulin monoclonal antibodies in the same manner as the anti-*T.brucei* cytoskeleton serum that isolated the above sequences. It turned out that monoclonal immunoscreening of the bacteriophage cDNA expression library could not detect tubulin, which could be due to the fact that the expression library being used was deficient of this major trypanosome cytoskeletal protein; thus the library used may not have been rich in such major gene products. However, this alone is not conclusive considering limitations of monoclonal antibodies as probes for cDNA libraries since they recognize single epitopes and often recognize an epitope that is a product of post translation modifications such as acetylation of Lys^{40} in alpha tubulin [22]. On the other hand, components from the trypanosome cytoskeleton proteome are highly glycosylated [23]. It is therefore possible that some of the potential antigens could have been carbohydrate epitopes that cannot be expressed using the expression cDNA library which may be deficient of the trypanosome-specific requirements for glycosylation and GPI anchoring [24].

One obvious bias involved with immunoscreening of the trypanosome expression library is that the picks are likely to be highly immunogenic trypanosome cytoskeleton proteins thus components that are poorly immunogenic are likely to be missed out [25]. However, this method proved particularly useful for the molecular identification of minor components such as DLC, NP, H2B, PFR-C proteins and uncharacterizable isolates in a complex cytoskeletal structure.

Conclusion: This approach is simple and less expensive compared to other techniques such as mass spectrometry and 2D-gel electrophoresis for the identification of antigens with differential immunogenicities that can be used as targets for diagnostics, drug and/or vaccine discovery against African trypanosomiasis in developing countries like Uganda. We therefore, recommend the screening of *T. brucei* expression libraries with sets of sera produced at different times to identify a repertoire of immunogenic proteins for the control and management of African trypanosomiasis in humans and livestock.

Ethics approval and consent to participate

"Not applicable"

Consent for publication

"Not applicable"

Availability of data and material

Datasets generated during this study has been submitted in a publicly available repository, GenBank Submission ID is 2130078

REFERENCES

[1]WHO. (2017). Human African trypanosomiasis: epidemiological situation. http://www.who.int/trypanosomiasis_african/country/en/

[2] Bacchi, C. J. (2009). Chemotherapy of Human African Trypanosomiasis. Interdisciplinary Perspectives on Infectious Diseases. Volume 2009, Article ID 195040, 5 pages. http://dx.doi.org/10.1155/2009/195040

[3] Buscher *et al.,* (2017). Human African trypanosomiasis. The Lancet, Volume 390, Issue 10110, 2397 - 2409

[4] Vickerman K. (1985). Developmental cycles and biology of pathogenic trypanosomes.*Brit. Med. Bull.* 41:105-114.

[5] Baines and Gull (2008). WCB is a C2 domain protein defining the plasma membrane sub-pellicular microtubule corset of kinetoplastid parasites. *Protist*. 159(1):115-25.

[6] Vedrenne C., Giroud C., Robinson D.R., Besteiro S., Bosc C., Bringuad F., and Baltz T. (2002). Two related subpellicular cytoskeleton associated proteins in *Trypanosoma brucei* stabilize microtubules. *Molecular biology of the cell*. 13: 1058-1070.

[7] Woods A., Sherwin T., Sasse R., MacRae T.H., Baines A.J., and Gull K. (1989). Definition of individual components within the cytoskeleton *Trypanosoma brucei*by a library of monoclonal antibodies.*Journal of cell science*. 93: 491-500.

[8] Berriman*et al*., (2005). The Genome of the African Trypanosome *Trypanosoma brucei*.*Science.*309: 416-422.

[9] Bastin P., Sherwin T., and Gull K. (1998). Paraflagella rod is vital for Trypanosome motility. *Nature.*391, 548.

[10] Hill K .L. (2003).Biology and mechanism of trypanosome cell motility. *Eukaryot cell minireviews.* 2 (2): 200-208.

[11] Ridgley, E., P. Webster, C. Patton, and L. Ruben. (2000). Calmodulin-binding properties of the paraflagellar rod complex from *Trypanosoma brucei*. Mol. *Biochem. Parasitol*.109:195–201.

[12] Bridges D.J., Pitt A.R., Hanrahan O., Brennan K., Voorheis H.P., Herzyk P., Koning H.P., and Burchmore R.J.S. (2008). Characterisation of the plasma membrane subproteome of bloodstream form Trypanosoma brucei. *Proteomics* .8: 83-99.

[13] Gull K. (1999).The cytoskeleton of Trypanosomatidparasites.*Annu. Rev. Microbiol.*53: 629-655.

[14] Bastin P., Matthews K. R., and Gull K. (1996). The paraflagella rod of kinetoplastida: solved and unsolved questions. *Parasitol.Today.*12: 302-307.

[15] Maga J.A. and LeBowitz J.H. (1999). Unraveling the kinetoplastidparaflagella rod.*Trends cell Biol.* 9: 409-413.

[16] Garcia-Salcedo J.A., Gijon P., and Pays E. (1999).Regulated transcription of histone H2B genes of *Trypanosoma brucei. Eur J. Biochem.* 264: 717-723.

[17] El-Sayed*et al*., (2005). Comparative Genomics of Trypanosomatid Parasitic Protozoa.*science*309

[18] Baron D.M., Kabututu Z.P., and Hill L.K. (2007). Stuck in reverse: Loss of LC1 in *Trypanosoma brucei* disrupts outer dynein arms and leads to flagella beat and backward movement.*Journal of cell science*. 120: 1513-1520.

[19] Zhang X., et al. (1992). Localization of a Human nucleoporin 155 gene (Nup 155) to the p513 region and cloning of its cDNA.Genomics. 57: 144-151.

[20]Luck D.J.L. (1984).Genetic and Biochemical of dissection of the eukaryoticflagellum. *J.Cell. Biol.* 98: 789-794.

[21] El-sayed N.M A., Hedge P., Quackenbush J.S., Melville S.E., and Donelson J.E. (2000). The African trypanosome genome.*International journal for parasitology.* 30: 329-345.

[22] Schneider A, Plessmann U, and Weber K (1997).Subpellicular and flagella microtubules of Trypanosoma brucei are extensively glutamylated. Journal of Cell Science, 110, 4 431-437

[23] Nolan D.P., Geuskens M. & Pays E. (1999). N-linked glycans containinglinearpoly-N-acetyllactosamine as sorting signals in endocytosis in Trypanosoma brucei. *Current Biology*9,1169-1172

[24] Ferguson M.A.J. (1997) In: Trypanosomiasis and Leishmaniasis Biology and Control, eds G. Hide, J.C. Mottram& G.H. Coombs, P.H. Holmes, p. 65, CAB International, Oxon, UK

[25] Birkett, C.R., Parma, A.E, Gerke-Bonet, R., Woodward, R., and Gull, K.: (1992). Isolation of cDNA clones encoding proteins of complex structures:analysis of the *Trypanosoma brucei* cytoskeleton. *Gene*, 110: 65-70

FIGURE LEGENDS

Figure 1: BLAST alignment of one deduced sequences (query) with the HSP sequence in the GeneDB. The top hit in the GeneDB was a paraflagella rod C protein (Tb927.3.4300 |PFR1|PFRC, 73 kDa) of *Trypanosoma brucei* located on chromosome 3, Accession No: [XM 8389929.1] with score = 1655 (254.4 bits), E-value = 6.2e-71, Identities = 339/346 (97%), Positives = 339/346 (97%), Strand = Plus / Plus

Figure 2: **BLAST alignment of one deduced sequences (query) with the HSP sequence in the GeneDB.** The top hit in the GeneDB was a histone H2B putative protein **[**Tb10.406.0350)] in *Trypanosoma brucei* located on chromosome 10 with score = 1695 (260.4 bits), E-value = 5.1e-72, Identities = 339/339 (100%), Positives = 339/339 (100%), Strand = Plus / Plus

Figure 3: BLAST alignment of one of the deduced sequences (query) with the HSP Sequence in the GeneDB. The top hit in the GeneDB was, dynein light chain [Tb10.389.0060], a putative protein located on *Trypanosoma brucei chromosome* 10 with score = 2846 (433.1 bits), Expect = 2.9e-124, Identities = 570/571 (99%), Positives = 570/571 (99%), Strand=Plus/ Plus

Figure 4: BLAST alignment of one of the deduced sequences (query) with the high scoring points (subject) in the GeneDB. The top hit in the GeneDB was a putative nucleoporin protein of *Trypanosoma brucei* located on chromosome 11 [systematic id Tb11.03.0140] with Score = 2736 (416.6 bits), E-value = 3.5e-120, Identities = 548/549 (99%), Positives = 548/549 (99%), Strand = Plus / Plus

Figure 5: BLAST alignment of one of the deduced sequences (query) with the high scoring points (subject) in the GeneDB. The top hit in the GeneDB was a hypothetical protein [Tb10.61.2430], conserved in *Trypanosoma brucei* and located on chromosome 10 with score = 660 (105.1 bits), E-value = 9.8e-25, Identities = 132/132 (100%), Positives = 132/132 (100%),, Strand = Plus / Plus. The match with the subject occurred with very significantstatistical parameters and a perfect alignment all through the entire length of query.

Figure 6: **BLAST alignment of one of the deduced sequences (query) with the HSP sequence in the Genebank database.** The top hit in the GenBank database was clone RPCI93-48O8, accession number [gb|AC091330.18], a *Trypanosoma brucei* clone located on chromosome 3 with score = 529 bits (286), E-value = 3e-147, Identities = 290/292 (99%), Gaps = 0/292 (0%) Strand=Plus/Plus.

Tables

Table 1: Identified genes from the *T. brucei* bloodstream form expression library

Name of gene	**Systematic id**	**No of inserts**
Paraflagella rod protein C (PFR-C)	Tb927.3.4300	3
Histone H2B (H2B)	Tb10.406.0350	4
Dynein light Chain (DLC)	Tb10.389.0060	15
Putative nucleoporin (NP)	Tb11.03.0140	5
Conserved hypothetical protein (CHP)	Tb10.61.2430	2
clone RPCI93-48O8 of chromosome 3 (RPC)	Not available	2
T.b.gambiense DAL972 chromosome 9, complete sequence	Not available	72
Total		**103**

Table 2: Identity parameters from the BLAST analysis of the uncharacterizable sequences against NCBI and GeneDB databases. High p-values were obtained using all the 3 BLAST algorithms (blastn, tblastx, blastx) suggesting poor match with the subject sequences

Name	BLASTN	p-value	TBLASTX	p-value	BLASTX	p-value
A4	Tb09.160.3560	065	Tb927.4.2180	0.64	Tb09.160.1430	0.74
A5	Tb09.160.0090	0.92	Tb08.27p2.660	0.38	Tb09.160.5510	0.43
B1T7	Tb09.160.0090	0.46	Tb11.01.8000	0.43	Tb09.160.5510	0.19
B2T7	Tb09.160.0090	0.21	Tb927.2.4720	0.43	Tb927.1.3890	0.15
D3.T7	Tb09.160.0090	0.46	Tb09.V4.0031	0.54	Tb11.02.3840	0.54
KA	Tb09.160.0090	0.98	Tb927.7.2430	0.89	Tb09.160.5510	0.31
KB	Tb09.160.0090	0.98	Tb927.2.3160	0.89	Tb09.160.5510	0.21
KC	Tb09.160.0090.	0.98	Tb927.2.3160	0.89	Tb09.160.5510	0.81
KD	Tb09.160.0090.	0.98	Tb927.7.2430	0.88	Tb09.160.5510	o.43

KE	Tb09.160.0090.	0.98	Tb.11.01.5770	0.85	Tb09.160.5510	0.64
KF	Tb09.160.0090	0.98	Tb927.7.2430	0.89	Tb09.160.5510	0.15
KP	None		Tb10.70.3310	0.35	Tb927.1.4200	0.32
KT	Tb.927.7.5090	0.999	Tb09.211.0600	0.16	Tb.927.1.2300	0.54
KG	Tb09.160.0090	0.98	Tb927.7.2430	0.90	Tb09.160.5510	0.15
KH	Tb09.160.0090	0.98	Tb.11.01.5770	0.85	Tb09.160.5510	0.12
KI	Tb09.160.0090	0.98	Tb927.7.2430	0.90	Tb09.160.5510	0.15
KJ	Tb09.160.0090	0.98	Tb927.2.3160	0.89	Tb09.160.5510	0.14
KK	Tb09.160.0090	0.98	Tb927.7.2430	0.90	Tb09.160.5510	0.15
KL	None		Tb927.1.2280	0.99	Tb927.3.3210	0.32
KM	Tb09.160.0090	0.90	Tb927.7.2430	0.90	Tb927-09-v4	6.8e-09
KN	Tb09.160.0090	0.98	Tb11.01.7540	0.79	Tb09.160.5510	0.12
KO	Tb09.160.0090	0.98	Tb927.7.2430	0.89	Tb09.160.5510	0.12
KQ	Tb09.160.0090	0.98	Tb927.1.2280	0.84	Tb09.160.5510	0.15
KR	Tb09.160.0090	0.98	Tb927.7.2430	0.89	Tb09.160.5510	0.11
KS	Tb09.160.0090	0.98	Tb927.6.1790	0.76	Tb09.160.5510	0.15
KT	Tb927.7.5090	0.999	Tb09.211.0600	0.16	Tb927.1.2300	0.012
KU	Tb09.160.0090	0.98	Tb927.1.1460	0.999	Tb09.160.0090	0.70
EW002	Tb09.160.3560	0.89	Tb11.02.0050	0.89	Tb10.05.0180	0.28
EW017	Tb09.160.0090	0.98	Tb11.02.0050	0.89	Tb10.05.0180	0.28
EW015	Tb09.160.0090	0.98	Tb10.6k15.1470	0.90	Tb11.01.1180	0.47
EW014	Tb09.160.3560	0.998	Tb927.1.3610	0.98	Tb09.160.3560	0.998
EW013	Tb09.160.0090	0.98	Tb11.02.0050	0.89	Tb10.05.0180	0.28
EW012	Tb09.160.0090	0.98	Tb08.27p2.660	0.98	Tb10.05.0180	0.33
EW010	Tb09.160.0090	0.98	Tb927.1.1460	0.997	Tb09.160.0090	0.70
EW009	Tb09.160.3560	0.997	Tb.927.5.350	0.0045	Tb11.01.1180	0.12
EW008	Tb09.160.0090	0.98	Tb927.2.3160	0.89	Tb09.160.5510	0.15
EW004	Tb09.160.3560	0.997	Tb09.160.5540	0.56	Tb11.01.1180	0.12
EW003	Tb09.160.0090	0.98	Tb10.6k15.1470	0.67	Tb11.01.1180	0.15

T.brucei expression library immunoscreening manuscript Figures

FIGURES

```
Query:     1 GGCTGCAGGCACAGTTGGCGCACGTCCC-ACACAGACATTGAAGCAAGTGGAGGATATCA 59
             || ||||  | |||||||||||| ||||| ||||||||||||||||||||||||||||||
Sbjct:   453 GGATGCAA-CGCAGTTGGCGCAGGTCCCCACACAGACATTGAAGCAAGTGGAGGATATCA 511

Query:    60 TGAACGTAACGCAAATCCAGAATGCGCTTGCCTCAACTGACGACCAGATCAAGACGCAGT 119
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   512 TGAACGTAACGCAAATCCAGAATGCGCTTGCCTCAACTGACGACCAGATCAAGACGCAGT 571

Query:   120 TGGCGCAGCTTGAAAAAACGAACGAGATCCAGAACGTTGCGATGCATGATGGTGAGATGC 179
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   572 TGGCGCAGCTTGAAAAAACGAACGAGATCCAGAACGTTGCGATGCATGATGGTGAGATGC 631

Query:   180 AGGTCGCCGAGGAGCAAATGTGGACGAAGGTACAGCTTCAGGAGCGCTTGATCGATCTGA 239
             |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct:   632 AGGTCGCCGAGGAGCAAATGTGGACGAAGGTACAGCTTCAGGAGCGCTTAATCGATCTGA 691
```

```
Query:   240 TTCAGGACAAATTCCGCTTGATCAGCAAATGTGAGGAGGAGAACCAGGCCTTCAGCAAAA 299
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   692 TTCAGGACAAATTCCGCTTGATCAGCAAATGTGAGGAGGAGAACCAGGCCTTCAGCAAAA 751

Query:   300 TCCATGAGGTGCAGAAACAGGCGAATCAGGAAACGAGTCAGATGAA 345
             |||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   752 TCCATGAGGTGCAGAAACAGGCGAATCAGGAAACGAGTCAGATGAA 797
```

Figure 1: BLAST alignment of one deduced sequences (query) with the HSP sequence in the GeneDB. The top hit in the GeneDB was a paraflagella rod C protein (Tb927.3.4300 |PFR1|PFRC, 73 kDa) of *Trypanosoma brucei* located on chromosome 3, Accession No: [XM 8389929.1] with score = 1655 (254.4 bits), E-value = 6.2e-71, Identities = 339/346 (97%), Positives = 339/346 (97%), Strand = Plus / Plus

```
Query:    66 ATGGCCACTCCTAAGAGCACACCGGCCAAGACGCGCAAGGAGGCGAAGAAGACGCGGCGC 125
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:     1 ATGGCCACTCCTAAGAGCACACCGGCCAAGACGCGCAAGGAGGCGAAGAAGACGCGGCGC 60

Query:   126 CAGCGCAAGCGCACATGGAACGTCTACGTCAGCCGCTCACTCCGCTCGATCAACAGCCAG 185
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    61 CAGCGCAAGCGCACATGGAACGTCTACGTCAGCCGCTCACTCCGCTCGATCAACAGCCAG 120

Query:   186 ATGTCGATGACCAGCCGGACGATGAAGATCGTGAACTCATTTGTGAACGACCTGTTTGAG 245
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   121 ATGTCGATGACCAGCCGGACGATGAAGATCGTGAACTCATTTGTGAACGACCTGTTTGAG 180

Query:   246 CGCATTGCTGCAGAGGCTGCTACGATCGTGCGTGTGAACCGGAAGCGGACCCTGGGCGCT 305
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   181 CGCATTGCTGCAGAGGCTGCTACGATCGTGCGTGTGAACCGGAAGCGGACCCTGGGCGCT 240

Query:   306 CGCGAGTTGCAGACGGCTGTGCGCCTTGTGCTGCCTGCTGACCTCGCGAAGCACGCGATG 365
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   241 CGCGAGTTGCAGACGGCTGTGCGCCTTGTGCTGCCTGCTGACCTCGCGAAGCACGCGATG 300

Query:   366 GCAGAGGGGACGAAGGCTGTGTCACACGCTTCCAGCTAA 404
             ||||||||||||||||||||||||||||||||||||||
Sbjct:   301 GCAGAGGGGACGAAGGCTGTGTCACACGCTTCCAGCTAA 339
```

Figure 2: **BLAST alignment of one deduced sequences (query) with the HSP sequence in the GeneDB.** The top hit in the GeneDB was a histone H2B putative protein **[**Tb10.406.0350)] in*Trypanosoma brucei* located on chromosome 10 with score = 1695 (260.4 bits), E-value = 5.1e-72, Identities = 339/339 (100%), Positives = 339/339 (100%), Strand = Plus / Plus

```
Query:    38 ATGTCTGATGAGACGCCTCAGGAGGGGCCTAAAAGCCCGGGGTCTGACACCGCACCCGCA 97
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:     1 ATGTCTGATGAGACGCCTCAGGAGGGGCCTAAAAGCCCGGGGTCTGACACCGCACCCGCA 60

Query:    98 CCAGTTGCCGCTGACGACTCACCCGAAACGGTGTCAGGGCAGGAGGCGGGATCTAGTTCC 157
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    61 CCAGTTGCCGCTGACGACTCACCCGAAACGGTGTCAGGGCAGGAGGCGGGATCTAGTTCC 120

Query:   158 AACGCCGAGAAGGAAGCGATGGGGTCCTCAGCGGCTCCGGCAGCTGCTACGGGACCACCA 217
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   121 AACGCCGAGAAGGAAGCGATGGGGTCCTCAGCGGCTCCGGCAGCTGCTACGGGACCACCA 180

Query:   218 GCTGGTGCACAGGAAGATCATGACAGCGAGGAGGAAGACACTGCAGCAGGAACCGGGTCG 277
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   181 GCTGGTGCACAGGAAGATCATGACAGCGAGGAGGAAGACACTGCAGCAGGAACCGGGTCG 240

Query:   278 GCGCCGGATGCTGCCGTCGACATTAACGGTGTTCCCGCTAATGATGTGAAGAACATTATC 337
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   241 GCGCCGGATGCTGCCGTCGACATTAACGGTGTTCCCGCTAATGATGTGAAGAACATTATC 300

Query:   338 CTGCAGGTCCTTAGCCCTTGCTTTGATGACGAGGGTGGCGAAGATGACGCGCAGCGCTAC 397
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   301 CTGCAGGTCCTTAGCCCTTGCTTTGATGACGAGGGTGGCGAAGATGACGCGCAGCGCTAC 360
```

```
Query:   398 GACCATATCAAAGCACACGGATGGATTCAACACATTTGCGATGGAATTATGGAAAAACTA 457
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   361 GACCATATCAAAGCACACGGATGGATTCAACACATTTGCGATGGAATTATGGAAAAACTA 420

Query:   458 CTGGCAATGCGCCGCCCCTACAAATATGTTGTTCACTGTGTAATCATGCGGAAATCACGT 517
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct:   421 CTGGCAATGCGCCGCCCCTACAAATATGTTGTTCACTGTGTAATCATGCGGAAATCAGGT 480

Query:   518 GCGGGCATTCACTTGTGTTCTAGTTGCTACTATGGCCAGGCGGATGGGTGGGTGAACCAT 577
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   481 GCGGGCATTCACTTGTGTTCTAGTTGCTACTATGGCCAGGCGGATGGGTGGGTGAACCAT 540

Query:   578 GCGCATGACCTCTCTGCGCATGTCTACGCAG 608
             ||||||||||||||||||||||||||||||
Sbjct:   541 GCGCATGACCTCTCTGCGCATGTCTACGCAG 571
```

Figure 3: BLAST alignment of one of the deduced sequences (query) with the HSP Sequence in the GeneDB. The top hit in the GeneDB was, dynein light chain [Tb10.389.0060], a putative protein located on *Trypanosoma brucei*chromosome 10 with score = 2846 (433.1 bits), Expect = 2.9e-124, Identities = 570/571 (99%), Positives = 570/571 (99%), Strand=Plus/ Plus

```
Query:     4 AATTTCGTTTCCCTCGGGGAGTGGTTTCCTGATGCCGCGCTGAGCCTGTTGGAGGGTTTC 63
             | ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4114 AGTTTCGTTTCCCTCGGGGAGTGGTTTCCTGATGCCGCGCTGAGCCTGTTGGAGGGTTTC 4173

Query:    64 GCGATGGGTGTGGCTCCCGCCGCCACTGCACTTCATCCACATTCATCAAGTTACTGTGCA 123
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4174 GCGATGGGTGTGGCTCCCGCCGCCACTGCACTTCATCCACATTCATCAAGTTACTGTGCA 4233

Query:   124 ACGGATTACTTGACACCATTCATAATTATAGTAAGTGTCCGTGCTTTGAAGTTACAGCGC 183
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4234 ACGGATTACTTGACACCATTCATAATTATAGTAAGTGTCCGTGCTTTGAAGTTACAGCGC 4293

Query:   184 ACAGACACCTATGACGACGCTGAAACAAAAGCACTTCTGGGGTTTGCCGCAGCGCTGGAG 243
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4294 ACAGACACCTATGACGACGCTGAAACAAAAGCACTTCTGGGGTTTGCCGCAGCGCTGGAG 4353

Query:   244 TGTCTTAGTGATGCGTGGTTCTGGGCTTTACTGCCGCTCCATATGATAGTTGACGACAAG 303
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4354 TGTCTTAGTGATGCGTGGTTCTGGGCTTTACTGCCGCTCCATATGATAGTTGACGACAAG 4413

Query:   304 TGCCGGGCTGTGGCTGTTGAGCAGTGTTTGCGCCGTAACGCCCACCGGTTCCAGGGAGGG 363
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4414 TGCCGGGCTGTGGCTGTTGAGCAGTGTTTGCGCCGTAACGCCCACCGGTTCCAGGGAGGG 4473

Query:   364 GCGTGCCGCACCAACACGGACTACGTACATTTGGCAGAGCTGTTGAAAATTAATGCTCGG 423
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4474 GCGTGCCGCACCAACACGGACTACGTACATTTGGCAGAGCTGTTGAAAATTAATGCTCGG 4533

Query:   424 CTGTTAGATGTGGAGATGCTGTTAGAGAAAGTCCCGGCAGATGCACCTGTGAATGCGCCA 483
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4534 CTGTTAGATGTGGAGATGCTGTTAGAGAAAGTCCCGGCAGATGCACCTGTGAATGCGCCA 4593

Query:   484 AGCATTCGCACTCACTCGTCTCTGCAGGAAGCGCTTCACCGTTTCAGTCGGGGGTTTACA 543
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  4594 AGCATTCGCACTCACTCGTCTCTGCAGGAAGCGCTTCACCGTTTCAGTCGGGGGTTTACA 4653

Query:   544 AAGAGATGA 552
             |||||||||
Sbjct:  4654 AAGAGATGA 4662
```

Figure 4: BLAST alignment of one of the deduced sequences (query) with the high scoring points (subject) in the GeneDB. The top hit in the GeneDB was a putative nucleoporin protein of *Trypanosoma brucei* located on chromosome 11 [systematic id Tb11.03.0140] with Score = 2736 (416.6 bits), E-value = 3.5e-120, Identities = 548/549 (99%), Positives = 548/549 (99%), Strand = Plus / Plus

```
Query:     1 CTTCCTCGGCTTCCGAGTATACCCGAAGTGGAAGATTTCTCTCCTCGCAGGCGCCTTCAT 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3391 CTTCCTCGGCTTCCGAGTATACCCGAAGTGGAAGATTTCTCTCCTCGCAGGCGCCTTCAT 3450

Query:    61 CGGAAGCCCACGTCCAAATCTTTTGGGAGCCAGGAAATACGTGCCCTCGCCTCACTCGTT 120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3451 CGGAAGCCCACGTCCAAATCTTTTGGGAGCCAGGAAATACGTGCCCTCGCCTCACTCGTT 3510

Query:   121 AAGTCCAAATAG 132
             ||||||||||||
Sbjct:       3511 AAGTCCAAATAG 3522
```

Figure 5: BLAST alignment of one of the deduced sequences (query) with the high scoring points (subject) in the GeneDB. The top hit in the GeneDB was a hypothetical protein [Tb10.61.2430], conserved in *Trypanosoma brucei* and located on chromosome 10 with score = 660 (105.1 bits), E-value = 9.8e-25, Identities = 132/132 (100%), Positives = 132/132 (100%),, Strand = Plus / Plus. The match with the subject occurred with very significantstatistical parameters and a perfect alignment all through the entire length of query.

```
Query  1AATGATGTATGTTTTCCCCCTTGACGTGGTAAGATGTATATCGttccttttccttttttc  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  14012  AATGATGTATGTTTTCCCCCTTGACGTGGTAAGATGTATATCGTTCCTTTTCCTTTTTTC  14071

Query  61ctttccttcttttttctttctttttcGAGGGGAATGAGGTTATCAGGCAGCAGAGGGTTTGT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  14072  CTTTCCTTCTTTTTCTTTCTTTTCGAGGGGAATGAGGTTATCAGGCAGCAGAGGGTTTGT  14131

Query  121GGCTCAAATACTGCCTATTCGATTTGCGCTAAATATACCACGCGCTCTActtctttcctt  180
             |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  14132  GGCTCAAATAATGCCTATTCGATTTGCGCTAAATATACCACGCGCTCTACTTCTTTCCTT  14191

Query  181ttttaactgctttttttaaaaattcttttccttttcCTATTAACGTTACCGGCAATATTCCT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  14192  TTTTAACTGCTTTTTTAAAAATTCTTTTCCTTTCCTATTAACGTTACCGGCAATATTCCT  14251

Query  241    TCCCCCTTATCCACACACACACACACACACACACACACACACACACACACACAC  292
               |||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct14252  TCCCCCTTATCCACACACACACACACACACACACACACGCACACACACAC  14303
```

Figure 6: **BLAST alignment of one of the deduced sequences (query) with the HSP sequence in the Genebank database.**The top hit in the Genbank database was clone RPCI93-48O8, accession number [gb|AC091330.18], a *Trypanosoma brucei* clone located on chromosome 3 with score = 529 bits (286), E-value = 3e-147, Identities = 290/292 (99%), Gaps = 0/292 (0%) Strand=Plus/Plus.