D428–D431 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj108

TcruziDB: an integrated, post-genomics community resource for *Trypanosoma cruzi*

Fernán Agüero, Wenlong Zheng[1], D. Brent Weatherly[1], Pablo Mendes[2] and Jessica C. Kissinger[1,3,*]

Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, San Martín, B 1650 KNA, Buenos Aires, Argentina, [1]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602-2606, USA, [2]Department of Computer Science, University of Georgia, GA, USA and [3]Department of Genetics, University of Georgia, Athens, GA 30602-7223, USA

Received September 15, 2005; Revised and Accepted October 17, 2005

ABSTRACT

TcruziDB (http://TcruziDB.org) is an integrated post-genomics database for the parasitic organism, *Trypanosoma cruzi*, the causative agent of Chagas' disease. TcruziDB was established in 2003 as a flat-file database with tools for mining the unannotated sequence reads and preliminary contig assemblies emerging from the Tri-Tryp genome consortium (TIGR/SBRI/Karolinska). Today, TcruziDB houses the recently published assembled genomic contigs and annotation provided by the genome consortium in a relational database supported by the Genomics Unified Schema (GUS) architecture. The combination of an annotated genome and a relational architecture has facilitated the integration of genomic data with expression data (proteomic and EST) and permitted the construction of automated analysis pipelines. TcruziDB has accepted, and will continue to accept the deposition of genomic and functional genomic datasets contributed by the research community.

INTRODUCTION

Trypanosoma cruzi is the causative agent of American Trypanosomiasis (Chagas' Disease), for which there is no definitive chemotherapeutic treatment. The parasite has a complex life cycle, with four main stages occurring in two hosts. In the insect host, *T.cruzi* is found in the form of epimastigotes and metacyclic trypomastigotes. In the vertebrate host, it is found in the form of bloodstream trypomastigotes and intracellular amastigotes. Based on a number of polymorphic markers, it has been shown that *T.cruzi* strains can be classified into two defined subgroups, Tcruzi I and II. Furthermore, strains that belong to the Tcruzi II group are more heterogeneous and can be further separated into distinct subgroups ranging from IIa to IIe.

The *T.cruzi* genome sequence was generated using a whole-genome shotgun (WGS) approach. The sequence generation, assembly and annotation were performed by researchers that are part of an international consortium, the TSK-TSC, comprised of The Institute for Genomic Research (TIGR, USA), the Seattle Biomedical Research Institute (SBRI, USA) and the Karolinska Institute (KI, Sweden) (1). The sequenced strain, CL-Brener, is a hybrid derived from an ancient hybridization event between parental Tcruzi IIb and IIc subgroups (1). The current *T.cruzi* genome assembly consists of 32 746 contigs totaling ∼89 Mb. Of these, 4008 contigs represent the majority of the coding portion of the genome, the remaining contigs represent smaller regions primarily consisting of repeat regions. Annotation for the 4008 contigs provided by the TSK-TSC (1) is represented in TcruziDB. The remaining contig sequences were not annotated, but they are available for searches within the database.

TcruziDB was established before the genome sequencing project was completed to serve as a resource for the research community (2). In 2003, TcruziDB provided access to unassembled shotgun reads and preliminary contig assemblies. Since that time, the genome sequence has been published (1) together with a whole-genome proteomic analysis covering the four main life cycle stages of the parasite (3). Also during this period, new EST sequences were deposited in GenBank, adding EST coverage to the two life-cycle stages that occur in the mammalian host (4,5).

In this paper, we report on the status of TcruziDB, and highlight new database architectural features and datasets that have been added by the genome consortium and the *T.cruzi* research community.

*To whom correspondence should be addressed. Tel: +1 706 542 6562; Fax: +1 706 542 3910; Email: jkissing@uga.edu.

DATA INVENTORY UPDATES

Version 4.0 of TcruziDB, released in July 2005, contained the recently published genome sequence and annotation of *T.cruzi* strain CL-Brener (1) as submitted by the sequencing consortium (TSK-TSC Genome Release v5.0) together with proteomic data contributed by Dr R. Tarleton (3) and EST data available in public repositories, Table 1. The annotated genome sequence data consists of 32 746 contigs, 4008 of which contain coding regions that were annotated. In the 5.0 TSK-TSC genome sequence release, 19 613 protein coding sequences and 3603 pseudogenes were predicted and all are represented within TcruziDB.

A dataset representing whole-organism proteolytic peptides was deposited in TcruziDB by members of the *T.cruzi* research community. This peptide dataset was updated in TcruziDB version 4.0 following the publication of the annotated genome (3). The peptides obtained from metacyclic trypomastigotes (CL strain) and amastigotes, trypomastigotes and epimastigotes (Brazil strain) were separated by multidimensional liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS): 139 147 high mass accuracy tandem mass spectra from this analysis were matched with a confidence of >99% to 2755 proteins in the annotated *T.cruzi* genome (3). Pre-computed datasets organized by life-cycle stage are available for viewing and download (Figure 1D).

Version 4.1 of TcruziDB, released in September 2005, represents an update that adds available EST data (Table 1) to TcruziDB: 13 968 EST sequences were obtained from the NCBI GenBank (6). In the majority of the cases individual analyses of these datasets have been published (4,5,7–10). In previous versions of TcruziDB, EST data obtained from NCBI were available for download and sequence similarity searches, but they were not integrated with available genome data. Now, ESTs are clustered into RNA transcripts (assemblies) and mapped against the genome. Also, based on the addition of ESTs derived from directionally cloned, spliced-leader based cDNA libraries (5), we were also able to map the *T.cruzi* miniexon sequence onto the EST assemblies (Table 1).

Table 1. Available EST data

Library	Stage	Strain	ESTs	Observations
TEN[a]	Epimastigote	CL-Brener	9761	Normalized
TEU[b]	Epimastigote	CL-Brener	255	Non-normalized
Tomoo	Epimastigote	Y	37	
	Epimastigote/ metacyclic trypomastigotes	Dm28c	175	Differential display
TcAma	Amastigote	Tulahuen	968	Non-normalized
TcAM	Amastigote	CL-Brener	1269	Non-normalized
TcTR	Trypomastigote	CL-Brener	1503	Non-normalized

Clustered EST statistics	
Total EST sequences	13968 (100%)
Included in assemblies (>50 nt)	13250 (94.8%)
Assemblies	7201 (100%)
Singletons	4988 (69.3%)
Clusters (2-112 ESTs)	2213 (30.7%)
Assemblies with an annotated SL sequence	1537 (21.3%)

EST data obtained from GenBank were loaded into TcruziDB in the form of separate datasets, one per cDNA library.

[a]The cDNA library was sequenced and submitted to GenBank as four different clone-sets (TENF, TENG, TENS, TENU).

[b]The cDNA library was sequenced and submitted to GenBank as two different clone-sets (TEUQ, TEUF). EST assemblies were generated with CAP 4 after contaminating vector sequences were removed.

SYSTEM DESIGN AND IMPLEMENTATION

TcruziDB was migrated from a flat-file database to a relational database structure beginning with release 3.0 of the database. The relational schema that has been employed is version 3.x of the Genomics Unified Schema, GUS (http://gusDB.org) (11) and our database management system is Oracle version10g. The Web interface to the database is produced via a Java servlet. A new home page for the database was designed to include 'one click' access to the most commonly used features of the database. In addition to the relational database, several additional applications are provided such as BLAST and a variety of custom PERL scripts to facilitate data mining and text searches.

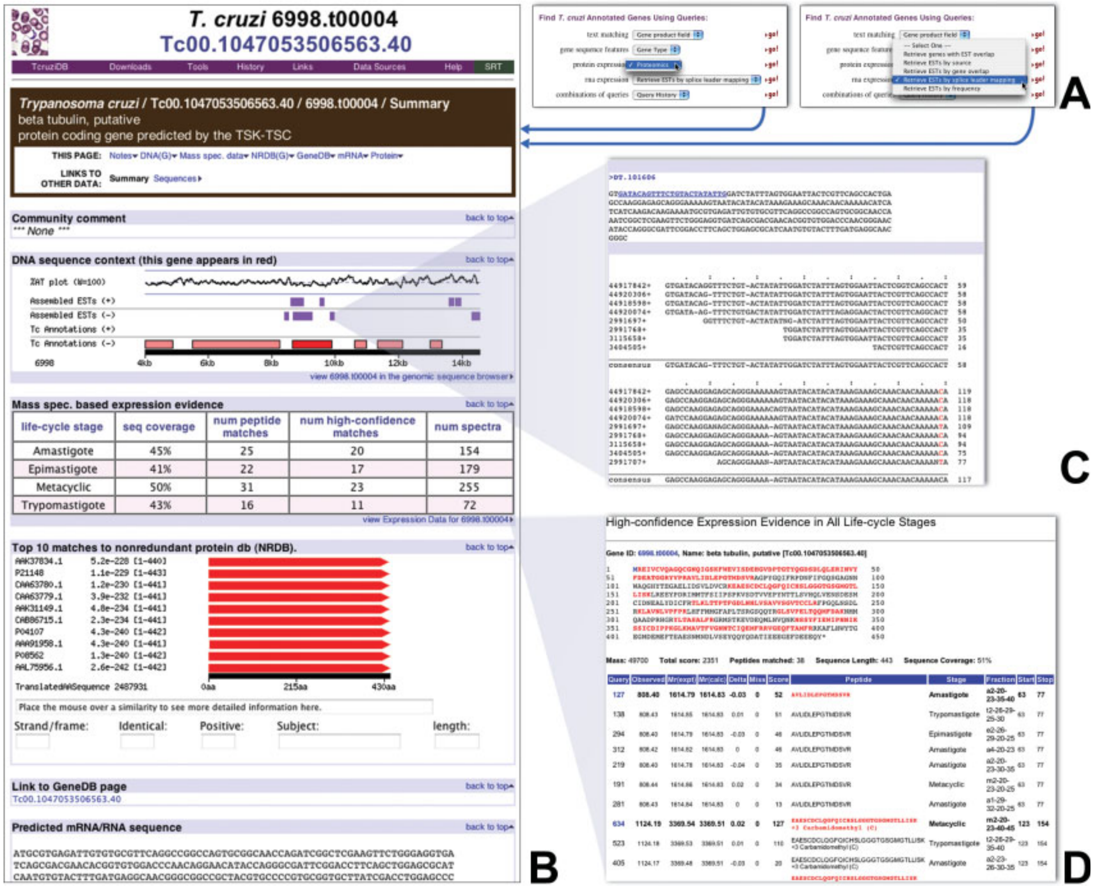
Beginning with TcruziDB release 4.0, a community comment field has been added to the database. Users with additional information on the annotation, function or properties of a predicted gene or feature of the genome sequence are encouraged to submit their comments to the database via email to help@tcruzidb.org. Comments, with author attribution, will be posted in the community comment field.

ANALYSIS TOOLS

The migration of the database to a relational architecture combined with the deposition of new genomic and functional genomic datasets has greatly expanded the types of analyses that can be performed. Queries of annotated features are now available including searches by gene name, gene location, gene type (protein coding, pseudogene, rRNA, slRNA, snRNA and tRNA) and key word descriptions, e.g. 'mucin'. Users can view the genomic context of both genes and contigs (Figure 1B) and these sequences can be retrieved with either feature IDs (e.g. 2383.t00001) or locus tags (e.g. Tc00.1047053439653.10). To facilitate internet navigation by users, a direct link from each annotated gene page in TcruziBD to the same gene as represented in GeneDB (12) is provided.

The integration of proteomic and EST datasets in TcruziDB permits users to construct combined queries for *T.cruzi* genes based on the available evidence of expression, both at the RNA and protein levels. This is particularly important for *T.cruzi*, but also for other trypanosomatid genomes (for which proteomic and EST data are lacking or limited) since only ∼50% of the predicted protein coding genes have been assigned a putative function. The integration of these functional analysis datasets with the genome annotation permits users to decide if a hypothetical protein encoding gene prediction can be now called a real protein (although perhaps still of unknown function), and at the same time provides information on developmental expression of the transcript and/or protein. Links to a detailed report of the analysis data is available for each gene having expression evidence.

The queries for both types of expression data are based on the mapping of either proteomic data [as described in (3)] or EST assemblies [as described in (13)] against the genome, and



Find *T. cruzi* Annotated Genes Using Queries:

text matching	Gene product field	›go!
gene sequence features	Gene Type	›go!
protein expression	✓ Proteomics	›go!
rna expression	Retrieve ESTs by splice leader mapping	›go!
combinations of queries	Query History	›go!

Find *T. cruzi* Annotated Genes Using Queries:

-- Select One --
Retrieve genes with EST overlap
Retrieve ESTs by source
Retrieve ESTs by gene overlap
✓ Retrieve ESTs by splice leader mapping
Retrieve ESTs by frequency

T. cruzi 6998.t00004
Tc00.1047053506563.40

TcruziDB Downloads Tools History Links Data Sources Help SRT

Trypanosoma cruzi / Tc00.1047053506563.40 / 6998.t00004 / Summary
beta tubulin, putative
protein coding gene predicted by the TSK-TSC

THIS PAGE: Notes▾ DNA(G)▾ Mass spec. data▾ NRDB(G)▾ GeneDB▾ mRNA▾ Protein▾
LINKS TO OTHER DATA: Summary Sequences▸

Community comment back to top▴
*** *None* ***

DNA sequence context (this gene appears in red) back to top▴

view 6998.t00004 in the genomic sequence browser▸

Mass spec. based expression evidence back to top▴

life-cycle stage	seq coverage	num peptide matches	num high-confidence matches	num spectra
Amastigote	45%	25	20	154
Epimastigote	41%	22	17	179
Metacyclic	50%	31	23	255
Trypomastigote	43%	16	11	72

view Expression Data for 6998.t00004▸

Top 10 matches to nonredundant protein db (NRDB). back to top▴

AAK37834.1	5.2e-228 [1-440]
P21148	1.1e-229 [1-443]
CAA63780.1	1.2e-230 [1-441]
CAA63779.1	3.9e-232 [1-441]
AAK31149.1	4.8e-234 [1-441]
CAB86715.1	2.3e-234 [1-441]
P04107	4.3e-240 [1-442]
AAA91958.1	4.3e-240 [1-441]
P08562	1.3e-240 [1-442]
AAL75956.1	2.6e-242 [1-442]

TranslatedAASequence 2487931

Place the mouse over a similarity to see more detailed information here.

Strand/frame: Identical: Positive: Subject: length:

Link to GeneDB page back to top▴
Tc00.1047053506563.40

Predicted mRNA/RNA sequence back to top▴

```
ATGCGTGAGATTGTGTGCGTTCAGGCCGGCCAGTGCGGCAACCAGATCGGCTCGAAGTTCTGGGAGGTGA
TCAGCGACGAACACGGTGTGGACCCAACAGGAACATACCAGGGCGATTCGGACCTTCAGCTGGAGCGCAT
CAATGTGTACTTTGATGAGGCAACGGGCGGCCGCTACGTGCCCCGTGCGGTGCTTATCGACCTGGAGCCC
```

```
>DT.101606

GTGATACAGTTTCTGTACTATATTGGATCTATTTAGTGGAATTACTCGTTCAGCCACTGA
GCCAAGGAGAGCAGGGAAAAAGTAATACATACATAAAGAAAGCAAACAACAAAAACATCA
TCATCAAGACAAGAAAATGCGTGAGATTGTGTGCGTTCAGGCCGGCCAGTGCGGCAACCA
AATCGGCTCGAAGTTCTGGGAGGTGATCAGCGACGAACACGGTGTGGACCCAACGGGAAC
ATACCAGGGCGATTCGGACCTTCAGCTGGAGCGCATCAATGTGTACTTTGATGAGGCAAC
GGGC


                 .    :    .    :    .    :    .    :    .    :    .    :
44917842+   GTGATACAGGTTTCTGT-ACTATATTGGATCTATTTAGTGGAATTACTCGGTCAGCCACT  59
44920306+   GTGATACAG-TTTCTGT-ACTATATTGGATCTATTTAGTGGAATTACTCGTTCAGCCACT  58
44918598+   GTGATACAG-TTTCTGT-ACTATATTGGATCTATTTAGTGGAATTACTCGTTCAGCCACT  58
44920074+   GTGATA-AG-TTTCTGTGACTATATTGGATCTATTTAGAGGAACTACTCGTTCAGGCACT  58
2991697+          GGTTTCTGT-ACTATATNG-ATCTATTTAGTGGAATTACTCGTTCAGCCACT  50
2991768+                       TGGATCTATTTAGTGGAATTACTCGTTCAGCCACT  35
3115658+                       TGGATCTATTTAGTGGAATTACTCGTTCAGCCACT  35
3404505+                                    TACTCGTTCAGCCACT  16

consensus   GTGATACAG-TTTCTGT-ACTATATTGGATCTATTTAGTGGAATTACTCGTTCAGCCACT  58

                 .    :    .    :    .    :    .    :    .    :    .    :
44917842+   GAGCCAAGGAGAGCAGGGAAAAAGTAATACATACATAAAGAAAGCAAACAACAAAAACA  119
44920306+   GAGCCAAGGAGAGCAGGGAAAAAGTAATACATACATAAAGAAAGCAAACAACAAAAACA  118
44918598+   GAGCCAAGGAGAGCAGGGAAAACAGTAATACATACATAAAGAAAGCAAACAACAAAAACA  118
44920074+   GATCCAAGGAGAGCAGGGAAAAAGTAATACATACATAAAGAAAGCAAACAACAAAAACA  118
2991697+    GAGCCAAGGANAGCAGGGAAAA-AGTAATACATACATAAAGAAAGCAAACAACAAAAATA  109
2991768+    GAGCCAAGGAGAGCAGGGAAAA-AGTAATACATACATAAAGAAAGCAAACAACAAAAACA  94
3115658+    GAGCCAAGGAGAGCAGGGAAAA-AGTAATACATACATAAAGAAAGCAAACAACAAAAACA  94
3404505+    GAGCCAAGGAGAGCAGGGAAAA-AGTAATACATACATAAAGAAAGCAAACAACAAAAACA  75
2991707+             AGCAGGGAAAN-ANTAATACATACATAAAGAAAGCAAACAACAAAANTA  77

consensus   GAGCCAAGGAGAGCAGGGAAAA-AGTAATACATACATAAAGAAAGCAAACAACAAAAACA  117
```

High-confidence Expression Evidence in All Life-cycle Stages

Gene ID: 6998.t00004, Name: beta tubulin, putative [Tc00.1047053506563.40]

```
1      MREIVCVQAGQCGNQIGSKFWEVISDEHGVDPTGTYQGDSDLQLERINVY    50
51     FDEATGGRYVPRAVLIDLEPGTMDSVRAGPYGQIFRPDNFIFGQSGAGNN    100
101    WAQGHYTEGAELIDSVLDVCRKEAESCDCLQGFQICHSLGGGTGSGMGTL    150
151    LISKLREEYPDRIMMTFSIIPSPKVSDTVVEPYNTTLSVHQLVENSDESM    200
201    CIDNEALYDICFRTLKLTTPTFGDLNHLVSAVVSGVTCCLRFPGQLNSDL    250
251    RKLAVNLVPFPRLHFFMMGFAPLTSRGSQQYRGLSVPELTQQMFDAKNMM    300
301    QAADPRHGRYLTASALFRGRMSTKEVDEQMLNVQNKNSSYFIEWIPNNIK    350
351    SSICDIPPKGLKMAVTFVGNNTCIQEMFRRVGEQFTAMFRRKAFLHWYTG    400
401    EGMDEMEFTEAESNMNDLVSEYQQYQDATIEEEGEFDEEEQY*           450
```

Mass: 49700 Total score: 2351 Peptides matched: 38 Sequence Length: 443 Sequence Coverage: 51%

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Peptide	Stage	Fraction	Start	Stop
127	808.40	1614.79	1614.83	-0.03	0	52	AVLIDLEPGTMDSVR	Amastigote	a2-20-23-35-40	63	77
138	808.43	1614.85	1614.83	0.01	0	51	AVLIDLEPGTMDSVR	Trypomastigote	t2-26-29-25-30	63	77
294	808.40	1614.79	1614.83	-0.03	0	46	AVLIDLEPGTMDSVR	Epimastigote	e2-26-29-20-25	63	77
312	808.42	1614.82	1614.83	0	0	46	AVLIDLEPGTMDSVR	Amastigote	a4-20-23	63	77
219	808.40	1614.78	1614.83	-0.04	0	35	AVLIDLEPGTMDSVR	Amastigote	a2-20-23-30-35	63	77
191	808.44	1614.86	1614.83	0.02	0	34	AVLIDLEPGTMDSVR	Metacyclic	m2-20-23-20-25	63	77
281	808.43	1614.84	1614.83	0	0	13	AVLIDLEPGTMDSVR	Amastigote	a1-29-32-20-25	63	77
634	1124.19	3369.54	3369.51	0.02	0	127	EAESCDCLQGFQICHSLGGGTGSGMGTLLISK +3 Carbamidomethyl (C)	Metacyclic	m2-20-23-40-45	123	154
523	1124.18	3369.53	3369.51	0.01	0	110	EAESCDCLQGFQICHSLGGGTGSGMGTLLISK +3 Carbamidomethyl (C)	Trypomastigote	t2-26-29-35-40	123	154
405	1124.17	3369.48	3369.51	-0.03	0	20	EAESCDCLQGFQICHSLGGGTGSGMGTLLISK +3 Carbamidomethyl (C)	Amastigote	a2-23-26-30-35	123	154
							EAESCDCLQGFQICHSLGGGTGSGMGTLLISK				

Figure 1. New features of TcruziDB. (**A**) New queries for RNA and protein expression data. (**B**) New gene record page displaying expression data, community comments and link to GeneDB. (**C**) Sample assembly/EST cluster alignment. A predicted splice leader is highlighted in blue. An assembly alignment column indicating an SNP is highlighted in red. (**D**) New proteomic data page. The location of identified peptides within the coding sequence is shown in red. Quality values for each observed peptide are provided.

can be modified to select for proteins and/or genes that are expressed in different life cycle stages of *T.cruzi*. In the case of proteomic data, users can additionally adjust their queries to select cases of proteins showing (i) a specified minimum coverage of the sequence with peptide mass-spectrometry data, (ii) a minimum number of peptide spectra that should match the protein and (iii) a minimum number of high-quality peptide spectra being matched (Figure 1A).

EST queries permit users to select for ESTs showing (i) a user-specified length of sequence overlap between annotated gene features and EST sequences, (ii) a specified minimum length of EST sequence that is required to be aligned to the genome and (iii) expression during a particular lifecycle stage. Also, new queries have been implemented to let users search for genes with experimental evidence of *trans*-splicing, based on the mapping of EST assemblies onto the genome and the presence of the *T.cruzi* spliced leader (miniexon) on these transcripts. EST assemblies can be viewed graphically with splice-leader and/or SNPs indicated if present (Figure 1C).

In addition to the new query functions described above, traditional analysis tools such as BLAST, user-defined protein motif searches and text searching are provided.

FUTURE PLANS

For the *T.cruzi* research community, the challenge now lies in turning the wealth of data already available into potential molecular drug and diagnostic markers. Both the new features described herein and the planned additions and improvements are focused to maintain TcruziDB as an integral bioinformatics analysis platform for the trypanosomatid research community.

TcruziDB will be migrated to the most recent version of GUS, ver. 3.5, and a new Web front-end will be installed using the recently released Web Development Kit (WDK) (http://gusdb.org). This infrastructure enhancement will greatly improve our database regeneration time and thus reduce the time needed between database updates. Database update cycles will be improved greatly by the implementation of new automated pipelines for the routine population of the database and common analyses such as BLAST comparisons and the identification of protein features such as signal peptides, GPI-anchors, transmembrane domains and protein motifs. As new data are deposited by the community into the database, they will be added to the analysis pipeline, integrated with existing data and presented to the community as rapidly as possible.

Given the availability of two other kinetoplastid genome sequences, *Trypanosoma brucei* (14) and *Leishmania major* (15), analyses of orthologous gene relationships and hyperlinks between the several existing database resources, TcruziDB and GeneDB (12) are important for researchers. Currently, TcruziDB provides direct gene-to-gene hyperlinks to GeneDB. Orthologous genes as determined by the sequencing consortium are currently provided on gene pages at GeneDB. We will be adding an additional link from the TcruziDB gene pages to the OrthoMCL—orthologous gene database (this issue). This database contains orthologous gene determinations for 55 organisms, representing all domains of life, including *T.cruzi* and the other kinetoplastid organisms.

One of the largest future challenges facing the database and the *T.cruzi* research community is the representation and characterization of *T.cruzi* haplotype information. Towards this end, we will be adding the increasing wealth of data that are available for the *T.cruzi* Esmeraldo strain and add haplotype designations to existing sequence data as they are determined by the sequencing consortium and the *T.cruzi* research community.

ACKNOWLEDGEMENTS

We would like to thank Mark Heiges for valuable assistance and the *Trypanosoma cruzi* research community for their continued support and encouragement. Funding for this project was provided by an American Heart Association award to J.C.K. F.A. was supported by the Biotechnology for Latin America and the Caribbean Programme (United Nations University) and the National Agency for the Promotion of Science and Technology (PICT 01-10819, Argentina). D.B.W. was supported by NIH grant PO-AI44979 to R. Tarleton. Funding to pay the Open Access publication charges for this article was provided by American Heart Association award 0330338N.

Conflict of interest statement. None declared.

REFERENCES

1. El-Sayed,N.M., Myler,P.J., Bartholomeu,D.C., Nilsson,D., Aggarwal,G., Tran,A.N., Ghedin,E., Worthey,E.A., Delcher,A.L., Blandin,G. *et al.* (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*, **309**, 409–415.
2. Luchtan,M., Warade,C., Weatherly,D.B., Degrave,W.M., Tarleton,R.L. and Kissinger,J.C. (2004) TcruziDB: an integrated *Trypanosoma cruzi* genome resource. *Nucleic Acids Res.*, **32**, D344–D346.
3. Atwood,J.A.III, , Weatherly,D.B., Minning,T.A., Bundy,B., Cavola,C., Opperdoes,F.R., Orlando,R. and Tarleton,R.L. (2005) The *Trypanosoma cruzi* proteome. *Science*, **309**, 473–476.
4. Cerqueira,G.C., DaRocha,W.D., Campos,P.C., Zouain,C.S. and Teixeira, S.M. (2005) Analysis of expressed sequence tags from *Trypanosoma cruzi* amastigotes. *Mem Inst Oswaldo Cruz.*, **100**, 385–389.
5. Aguero,F., Abdellah,K.B., Tekiel,V., Sanchez,D.O. and Gonzalez,A. (2004) Generation and analysis of expressed sequence tags from *Trypanosoma cruzi* trypomastigote and amastigote cDNA libraries. *Mol. Biochem. Parasitol.*, **136**, 221–225.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
7. Brandao,A., Urmenyi,T., Rondinelli,E., Gonzalez,A., de Miranda,A.B. and Degrave,W. (1997) Identification of transcribed sequences (ESTs) in the *Trypanosoma cruzi* genome project. *Mem. Inst. Oswaldo Cruz.*, **92**, 863–866.
8. Verdun,R.E., Di Paolo,N., Urmenyi,T.P., Rondinelli,E., Frasch,A.C. and Sanchez,D.O. (1998) Gene discovery through expressed sequence Tag sequencing in *Trypanosoma cruzi*. *Infect. Immun.*, **66**, 5393–5398.
9. Porcel,B.M., Tran,A.N., Tammy,M., Nyarady,Z., Rydäker,M., Urmenyi,T.P., Rondinelli,E., Pettersson,U., Andersson,B. and Aslund,L. (2000) Gene survey of the pathogenic protozoan *Trypanosoma cruzi*. *Genome Res.*, **10**, 1103–7.
10. Sotomaior,V.S., Freund,A., Ribas,P.D., Ogatta,S.F.Y., Dallagiovana,B., Avila,A.R., Monteiro,V.S., Buck,G.A., Goldenberg,S. and Krieger,M.A. (2000) Using of mRNA differential display to select genes differentially expressed during *Trypanosoma cruzi* metaciclogenesis. *Memorias do Instituto Oswaldo Cruz*, **95**, 273.
11. Davidson,S.B., Crabtree,J., Brunk,B., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,C.J.Jr (2001) K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Syst. J.*, **40**, 512–531.
12. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.
13. Li,L., Crabtree,J., Fischer,S., Pinney,D., Stoeckert,C.J.Jr, , Sibley,L.D. and Roos,D.S. (2004) ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res.*, **32**, D326–D328.
14. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E., Haas,B. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416–422.
15. Ivens,A.C., Peacock,C.S., Worthey,E.A., Murphy,L., Aggarwal,G., Berriman,M., Sisk,E., Rajandream,M.A., Adlem,E., Aert,R. *et al.* (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, **309**, 436–442.