D344–D346 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh049

TcruziDB: an integrated *Trypanosoma cruzi* genome resource

Michael Luchtan[1,2], Chetna Warade[2], D. Brent Weatherly[1], Wim M. Degrave[3], Rick L. Tarleton[1,4] and Jessica C. Kissinger[1,5,*]

[1]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602-2606, USA, [2]Department of Computer Science, University of Georgia, Athens, GA 30602, USA, [3]Department of Biochemistry and Molecular Biology, Oswaldo Cruz Institute, Rio de Janeiro, RJ 21045-900, Brazil, [4]Department of Cellular Biology, University of Georgia, Athens, GA 30602-2607, USA and [5]Department of Genetics, University of Georgia, Athens, GA 30602-7223, USA

Received August 15, 2003; Accepted September 23, 2003

ABSTRACT

TcruziDB (http://TcruziDB.org) is an integrated genome database for the parasitic organism *Trypanosoma cruzi*, the causative agent of Chagas' disease. The database currently incorporates all available sequence data (Genomic, BAC, EST) in a single user-friendly location. The database contains a variety of tools specifically designed for searching unannotated draft sequence via BLAST, keyword searches of pre-computed BLAST results, and protein motif searches. Release 1.0 of the database contains nearly 730 million bp of genome sequence from 1.1 million sequence reads generated by the TIGR–Karolinska–SBRI *Trypanosoma cruzi* Genome Consortium and 15 million bp of clustered EST and genomic sequence obtained from other sources. As annotation, microarray and proteomic data become available, the database will incorporate and integrate these data using the GUS (http://www.gusdb.org) relational framework.

INTRODUCTION

Trypanosoma cruzi is a pathogenic kinetoplastid parasite and the causative agent of Chagas' disease. The initiative to characterize the *T.cruzi* genome grew out of a series of meetings held in 1993 and 1994, and by 1997 initial characterization and the molecular karyotype had been determined for the CL Brener strain (1–3). These data along with emerging EST sequence data were deposited in a web-based home for the project, TcruziDB (3), hosted by the FIOCRUZ in Brazil (http://www.dbbm.fiocruz.br/genome/tcruzi/tcruzi.html). Limited shotgun sequencing of the genome was released in 2000 (4) and a whole-genome shotgun (WGS) sequence has recently been completed by the TIGR, Karolinska and SBRI *Trypanosoma cruzi* Genome Consortium (TKS-TGC). The current version of TcruziDB (http://TcruziDB.org) represents a collaborative effort between researchers in Brazil and the United States to bring the *T.cruzi* research community a comprehensive genome resource which will integrate numerous data types ranging from strain characterization and molecular karyotypes to genome sequence and functional genomic studies. The TKS-TGC has agreed to deposit genome sequence and draft annotation with the database on a regular basis. TcruziDB.org went live in April 2003 and averages hundreds of visits per month.

DATA INVENTORY

TcruziDB release 1.0 contains primarily sequence data. Microarray and proteomic data are available for *T.cruzi*, but in the absence of a draft genome sequence these data cannot yet be incorporated into the database in a meaningful and integrated fashion. The most recent release of genome data from the *T.cruzi* Genome Consortium is the July 11, 2003 data release. This release contains 1 118 787 individual WGS reads totaling 715 236 132 nt of sequence from the CL Brener strain. The database also contains available WGS assemblies, clustered and singleton ESTs totaling ~3 million bp, and ~13 million bases of genomic sequence available in the NCBI GenBank Databases. Open reading frames (ORFs) greater than 50 and 100 amino acids in length have been calculated for all nucleic acid sequences in all six reading frames. All data sets are available for bulk download or analysis using a variety of tools. The database is linked to TcruziDB at the FIOCRUZ in Brazil and this site provides information on Chagas' disease, *T.cruzi* strain types, population structure and information on molecular karyotypes.

ANALYSIS TOOLS

TcruziDB is designed to be a resource where users actively search the available data to find genes of interest. As the data are currently unannotated, three tools are provided to facilitate gene discovery. The first and most familiar tool is BLAST analysis. All data sets, including ORFs, are available for searching individually or in combined 'all available data' sets via BLAST. Results from BLAST searches are linked to a sequence retrieval tool, SRT (Fig. 1A) which permits the user

*To whom correspondence should be addressed. Tel: +1 706 542 6562; Fax: +1 706 542 3582; Email: jkissing@uga.edu

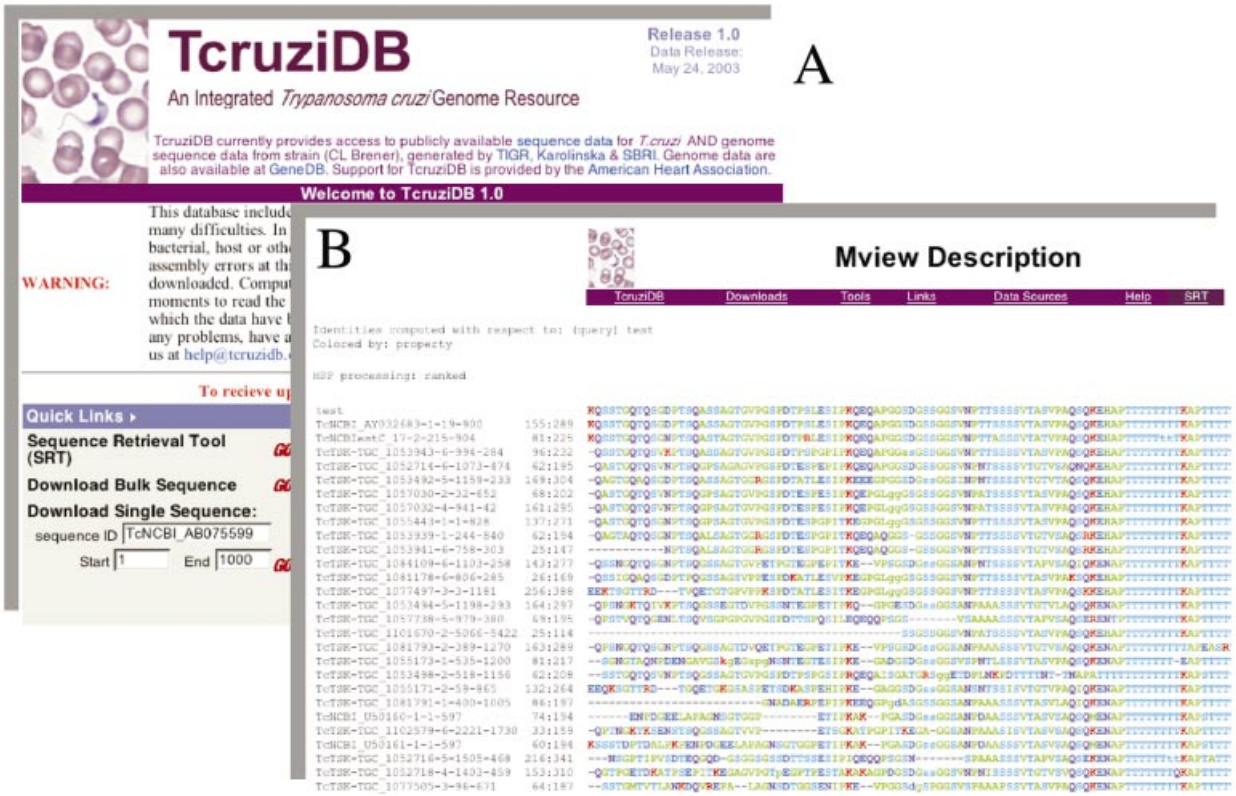

Figure 1. Composite of screen displays demonstrating features of TcruziDB. (**A**) Front page of the website. The lower left-hand corner shows the sequence retrieval tool, SRT. (**B**) Sample multiple sequence alignment generated from BLAST results using the application Mview. A portion of a mucin protein sequence was used to search the database with BLASTP.

to download immediately all or any portion of a sequence discovered in the search. BLAST results may also be viewed in multiple sequence alignment format using the application Mview (5) (Fig. 1).

The second tool is a complement to BLAST searches. In a typical BLAST analysis the user inputs a sequence that is then used to search a database to obtain other similar sequences. To alleviate the task of fetching a sequence to use for BLAST searching, all nucleotide sequences for *T.cruzi* have been used to perform a BLASTX search of the GenBank non-redundant protein database. All words and EC numbers in the definition line of hits with an E-value $\leqslant 0.001$ were collected and indexed with a pointer to the sequence(s) that gave rise to them. Users search the index by entering keywords, like 'mucin' or 'ribo*' and a list of all sequences that produced BLAST hits containing the keywords, along with their BLAST alignments, are returned to the user for inspection.

Another useful way to search for genes of interest is to search for protein motifs. In the absence of predicted genes, the motif search tool permits searches of any of the six-frame ORF data sets with defined PROSITE motifs (6), or a user-defined motif. Instructions are provided describing how to generate a PERL regular expression describing your pattern including ambiguous characters (e.g. eight threonine residues followed by lysine, alanine, proline and four or more threonine residues, a common mucin pattern, Fig. 1B).

Pub-Crawler searches (7) of the NCBI PubMed and sequence databases are performed nightly and users are alerted to new research articles or sequence depositions for *T.cruzi* as well as all other kinetoplastid organisms. Several links are provided to TcruziDB-FIOCRUZ resources and links are provided to other useful community sites including the World Health Organization's Tropical Disease Research Chagas' Disease site, and other kinetoplastid organism databases maintained by the Sanger Institute in GeneDB (http://www.GeneDB.org) and the TKS-TGC website.

SYSTEM DESIGN AND IMPLEMENTATION

Release 1.0 of TcruziDB is a flat file database which combines bioinformatics applications such as WUBLAST [W. Gish (1996–2003) http://blast.wustl.edu] and PubCrawler (7) with custom software developed and kindly provided by the PlasmoDB (8,9) and GUS database projects and subsequently modified to suit this project. The website software is constructed from CGI scripts written in PERL and pages are served using the Apache web server. The sequence retrieval tool, SRT, employs a daemon that continually listens for, and processes, sequence requests. Any sequence discovered during searches of the database can be downloaded in part or entirety in FASTA format. A data download site facilitates the download of large files. Live and development versions of

the database are maintained at all times to ensure smooth operation of the site.

FUTURE PLANS

With the imminent release of the draft *T.cruzi* genome sequence and first pass annotation, a large influx of new data is anticipated from numerous sources including expression profiling and proteomics. At this juncture, the underlying architecture will be changed to the GUS relational database software (10) (http://www.gusdb.org) that currently supports the PlasmoDB database. Several schema modifications will be necessary to capture the complex nature of this genome. We will be adding features and queries related to splice leader sequences, homologous chromosomes, polycistronic transcripts and allelic variation. Following the migration users will be able to perform complex queries of the database (e.g. 'list all genes annotated as transialidases' or 'list all genes expressed only in trypomastigotes' or 'list all genes with signal peptides and proteomic evidence'). The results from all such searches will be stored in the history for the user's session and can be combined with one another using Boolean operators to perform complex queries. All results will be downloadable in FASTA or tab-delimited format for easy loading into spreadsheet programs.

TcruziDB will display the official annotation provided by the TKS-TGC in addition to a variety of automated analyses including BLAST similarities, tandem repeats and protein features such as signal peptides, transmembrane domains and low-complexity regions. A community comment field will be added to each gene record to provide a forum for community input. Links to the *Trypanosoma brucei* and *Leishmania* databases located in GeneDB will be enhanced to facilitate comparative kinetoplastid research.

CITING TcruziDB AND ORIGINAL DATA SOURCES

Publications and presentations benefiting from the use of this database should cite the original data source(s), the data release date(s) and the database. A table of data sources, release dates, source contact information and publications related to data contained within the database are linked to the front page of TcruziDB. Some of the genome data contained within the site are subject to a data usage agreement. If you have downloaded the entire genome and accepted this agreement please follow the terms of the agreement regarding acknowledgement and publication. TcruziDB should be cited via the URL (http://TcruziDB.org) and this publication.

ACKNOWLEDGEMENTS

Preliminary genomic data were accessed via http://www.tigr.org/tdb/e2k1/tca1/. We thank these researchers for making data available prior to publication of the completed genome sequence. We also thank the GUS and PlasmoDB developers and numerous researchers who have collaborated with and contributed to TcruziDB by depositing data, making software available, and by making useful suggestions on how to improve this community resource. This work was supported by an award from the American Heart Association. Most of the genomic data were provided by the TSK-TSC supported by NIH grants AI45038, AI45061 and AI45039.

REFERENCES

1. Zingales,B., Rondinelli,E., Degrave,W., Da Silveira,J.F., Levin,M.J., Le Paslier,D., Modabber,F., Dobrokhotov,B., Swindle,J., Kelly,J.M. *et al.* (1996) The *Trypanosoma cruzi* Genome Consortium. *Parasitol. Today*, **13**, 16–22.
2. Cano,M.I., Gruber,A., Vazquez,M., Cortes,A., Levin,M.J., Gonzalez,A., Degrave,W., Rondinelli,E., Zingales,B., Ramirez,J.L. *et al.* (1995) Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* Genome Project. *Mol. Biochem. Parasitol.*, **71**, 273–278.
3. Degrave,W., de Miranda,A.B., Amorim,A., Brandao,A., Aslett,M. and Vandeyar,M. (1997) TcruziDB, an integrated database and the WWW information server for the *Trypanosoma cruzi* genome project. *Mem. Inst. Oswaldo Cruz*, **92**, 805–809.
4. Aguero,F., Verdun,R.E., Frasch,A.C. and Sanchez,D.O. (2000) A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: general structure, large gene and repetitive DNA families and gene discovery. *Genome Res.*, **10**, 1996–2005.
5. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
6. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.
7. Hokamp,K. and Wolfe,K. (1999) What's new in the library? What's new in GenBank? Let PubCrawler tell you. *Trends Genet.*, **15**, 471–472.
8. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
9. Kissinger,J.C., Brunk,B.P., Crabtree,J., Fraunholz,M.J., Gajria,B., Milgram,A.J., Pearson,D.S., Schug,J., Bahl,A., Diskin,S.J. *et al.* (2002) The *Plasmodium* genome database. *Nature*, **419**, 490–492.
10. Davidson,S.B., Crabtree,J., Brunk,B., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,J.,C.J. (2001) K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems J.*, **40**, 512–531.