ToxoDB: an integrated *Toxoplasma gondii* database resource

Bindu Gajria[1], Amit Bahl[1], John Brestelli[2], Jennifer Dommer[2], Steve Fischer[2], Xin Gao[2], Mark Heiges[3], John Iodice[2], Jessica C. Kissinger[3,4,*], Aaron J. Mackey[1], Deborah F. Pinney[2], David S. Roos[1], Christian J. Stoeckert Jr[2], Haiming Wang[3] and Brian P. Brunk[2]

[1]Department of Biology, [2]Department of Genetics, Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA, [3]Center for Tropical & Emerging Global Diseases and [4]Department of Genetics, University of Georgia, Athens, GA, USA

ABSTRACT

ToxoDB (http://ToxoDB.org) is a genome and functional genomic database for the protozoan parasite *Toxoplasma gondii*. It incorporates the sequence and annotation of the *T. gondii* ME49 strain, as well as genome sequences for the GT1, VEG and RH (Chr Ia, Chr Ib) strains. Sequence information is integrated with various other genomic-scale data, including community annotation, ESTs, gene expression and proteomics data. ToxoDB has matured significantly since its initial release. Here we outline the numerous updates with respect to the data and increased functionality available on the website.

INTRODUCTION

Toxoplasma gondii is an intracellular apicomplexan parasite capable of infecting humans. Infection is typically asymptomatic in healthy individuals, but may lead to congenital birth defects and encephalitis in immuno-suppressed individuals (1,2). ToxoDB, initially released in May 2001, has been substantially updated in both content and functionality since last described in January 2003 (3). ToxoDB provides access to the genome sequence and annotation of the *T. gondii* ME49 strain. It also incorporates the genomic sequence of multiple other strains. The parasite genome is ∼63 Mb in size and consists of 14 chromosomes (4).

The initial ToxoDB release was not supported by a relational database and thus the site had restricted functionality and little capability to integrate diverse data types such as gene expression data and single nucleotide polymorphism data (SNPs) with genomic sequence. Since initial publication, ToxoDB has been completely rebuilt using a common architecture similar to another apicomplexan database project, PlasmoDB (5). Both sites, along with CryptoDB, are component sites of ApiDB, the Apicomplexan Bioinformatics Resource Center (6). Many of the new methods of data loading, querying and presentation that are mentioned here have been applied to all of the ApiDB sites to provide a common research platform and facilitate data access among this group of related organisms. ApiDB (http://apidb.org/) serves as an 'umbrella' site for cross-species comparisons. Researchers can mine for *Toxoplasma* genes at ApiDB directly or via their orthologous relationship(s) to genes in other apicomplexan species.

CONTENT OF THE CURRENT RELEASE

Data

ToxoDB provides access to the genome sequence and annotation of *T. gondii* (ME49 strain) and the genomic sequence of the GT1, VEG and RH (Chr Ia and Chr Ib) strains. Annotation is also available for the apicoplast genome. The current database version (Release 4.2) also contains manual annotation (solicited in the initial genome annotation and entered by users as user comments), ESTs, TIGR Gene Indices clustered ESTs, SAGE tags, SNPs, cosmid and BAC ends, microarray and proteomics studies, all of which have been mapped to the genome (7,8). The database contains the results of automated analyses including gene predictions (using various algorithms), open reading frames (ORFs) greater than 50 aa and protein feature predictions

Table 1. Data and analyses that have been integrated into ToxoDB and the number of genes that are impacted

Data type	Data source	Number of genes
Genes	TIGR	8032
Community annotation	Various contributors	1610
Orthologs	Generated from OrthoMCL	4616
GO terms	TIGR; InterPro	3136
EC numbers	TIGR	800
SNPs	John Boothroyd Laboratory; David Roos Laboratory	7322
Microarray	David Roos Laboratory	7664
ESTs	dbEST, TIGR Gene Indices	6080
SAGE tags	TgSAGEDB (14)	6284
Proteomics	Johnathan Wastling Laboratory; John Murray Laboratory	2435
Epitopes	IEDB	10
Metabolic pathways	KEGG Pathway	614

[signal peptides, transmembrane domains, hydrophobicity plots, AA content and InterPro domains (9)], Gene Ontology function predictions, and BLAST similarities to the NCBI non-redundant protein database (Table 1).

In addition, we have used the OrthoMCL algorithm to group genes from *T. gondii* with orthologous genes from 86 other eukaryotic and prokaryotic genomes (10). A mapping of immune epitopes identified in *Toxoplasma* provided by the Immune Epitope Database and Analysis Resource (IEDB) (11) has been integrated. Affymetrix probes mapped to the genome are visible in GBrowse, as are SNPs generated from nucmer alignments of sequences from the GT1, VEG and RH (Chr Ia and Ib) strains against the reference ME49 sequence. Two expression experiments utilizing a *Toxoplasma* Affymetrix array have also been deposited in ToxoDB. Users gain access to these new data types in record pages and by queries using the powerful query interface (see Data-Mining section).

Database architecture

As a part of the complete restructure of the ToxoDB resource, the practice of using flat files as a means of data storage was abandoned in early 2006. We now use GUS 3.5, and load data into an underlying Oracle database in a systematic fashion. GUS (Genomics Unified Schema) is an open source project (www.gusdb.org) with a rich relational schema including sequence annotation, expression data and proteomics using controlled vocabularies and ontologies (12).

ToxoDB also employs the GUS WDK (Web Development Kit, www.gusdb.org/wdk), to access the database from the internet dramatically improving the way the website operates. This transformation has added considerable increased functionality for database users and conforms to the model used by all ApiDB projects, making it possible for us to generate future database releases in short cycles.

DATA-MINING TOOLS

ToxoDB currently provides 40 different queries of the data and several ancillary tools for analyzing, retrieving or viewing the data such as BLAST, Pathway Tools and an installation of the GMOD project Genome Browser (13). The ToxoDB 'Query & Tools' page has been restructured to make all queries available at a glance. Most of the individual queries have been reorganized into categories such as 'Position', 'Expression' and 'Function' to make them more intuitive to the average researcher. Enhanced functionality for the queries has also been added. For example, the ToxoDB keyword search has been significantly improved, offering the user control over which fields in the database are searched, including the official annotation, synonyms, user-supplied comments, domain names, BLAST similarities, etc. Many queries, such as 'Find SNPs based on Gene ID', now allow a gene ID list as input [either typed (or copied) by hand or uploaded from a file] facilitating analyses on large groups of genes. The results from all queries can be sorted based on various criteria (columns in the returned data set) and users can also add additional criteria for display (e.g. add columns to display protein features, GO annotation, expression characteristics for gene results, etc.) and sort on them as well. Once the appropriate selection of data types to display has been achieved, users can integrate these search results with other search results using the 'Query History' page, or the data can be downloaded in multiple formats for further analysis by the researcher (Figure 1).

ToxoDB uses the GBrowse genome browser (www.gmod.org) (13) to display gene models, EST alignments, SNPs, SAGE tags, etc. GBrowse enables visualization of the parasite genome and gene models, custom restriction-site identification, open reading frame identification, and facilitates download of data in various formats. Different data sets or analyses are displayed as individual tracks within the genome browser. There are approximately 50 GBrowse tracks available in the current version of ToxoDB. All genome sequences [ME49, GT1, VEG and RH (Chr Ia, Ib)] are also available in BLAST-searchable databases and for download in FASTA, GenBank and EMBL formats.

ToxoDB users may now register and log in to the site. Doing so enables a researcher to add comments to genes and genomic sequences. It also lets users save query results permanently. Queries in the Query History page can be organized (re-named or deleted) as well as combined with other results (Figure 1). This is a very powerful feature that allows users to refine their results so that precise sets of genes can be discovered.

The results may be downloaded using ToxoDB's improved reporting facility. It supports summary reports (Excel compatible tab delimited text), GFF, FASTA and a detailed report that includes almost all available data for each gene in the users result table. Use of this facility as well as many others on the site are now described in short video tutorials that are accessible from the database home page.
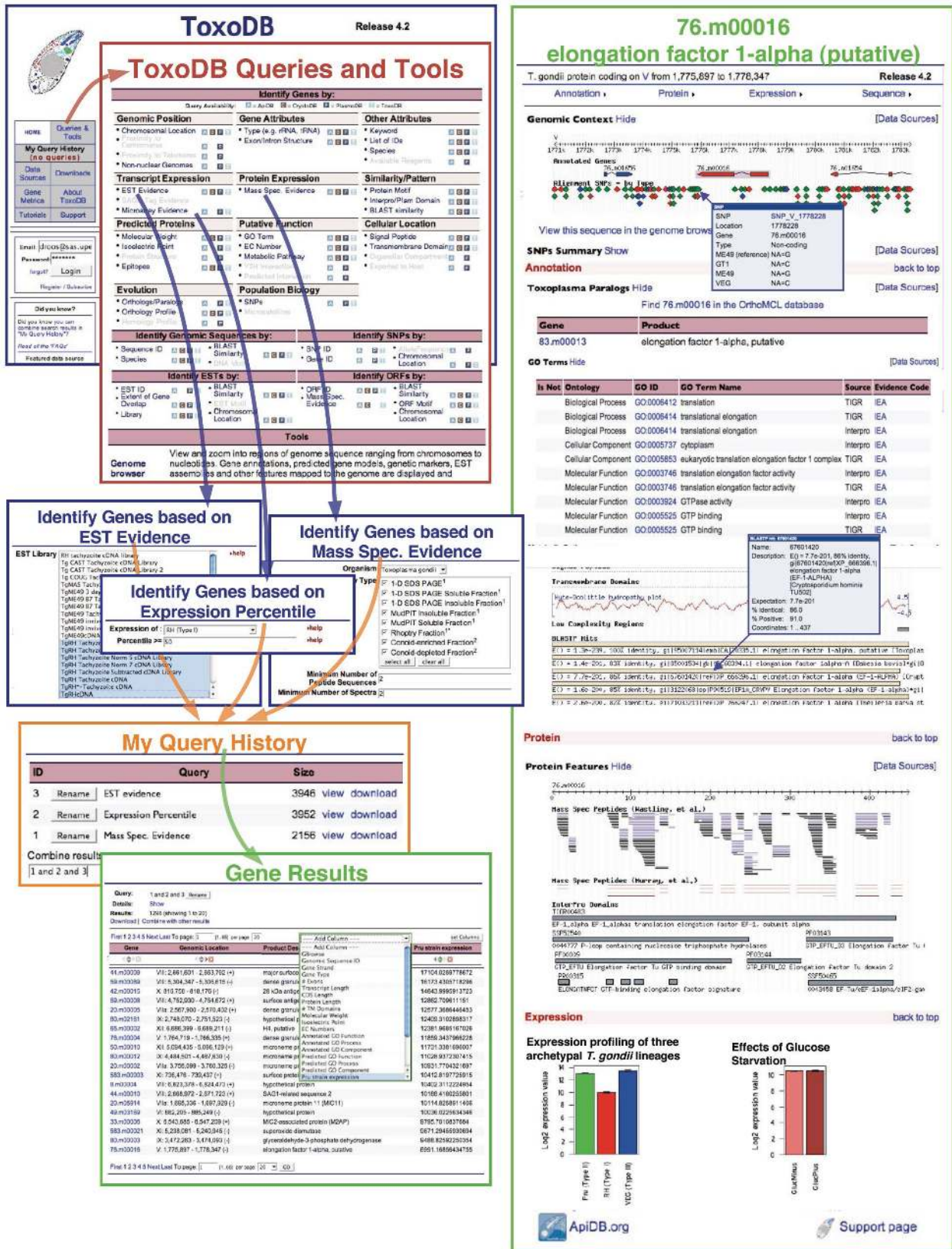
Figure 1. Screenshots showing the flow of a query in ToxoDB. From the Query & Tools page, users can go to particular queries for expression evidence (EST or Mass Spec Evidence), to the Results page where they can sort, manage (add or delete) columns of data and open gene pages. The Query History page permits users to manipulate previous queries including combining them and/or downloading the resulting data. Individual genes are listed on the Gene Results page and each gene has its own gene page, illustrated here by the gene encoding elongation factor 1-alpha. The gene page summarizes all information that is available for a gene including gene model predictions, SNPs, BLAST similarities, protein domains, ESTs, proteomic evidence of expression and microarray expression analyses.

FUTURE DIRECTIONS

The last two years were spent on major infrastructure and design elements for ToxoDB. Our future growth will be in the area of increased data acquisition and integration with existing and future data sets. Specifically, we are planning to load and integrate many expression data sets (RNA expression and protein expression) that are just becoming available. We also expect to load and integrate other array-based data sets such as ChIP on Chip and array CGH. As new data are added, we will be adding additional queries and tools to view these data. An area of significant future development will be improving the ability of users to compare the various different sequenced parasite strains visually and download sequence alignments between them.

ACKNOWLEDGEMENTS

REFERENCES

1. Remington,J.S. and Desmonts,G. (1989) Toxoplasmosis. In Remington,J.S. and Klein,J.O. (eds), *Infectious Diseases of the Fetus and Newborn Infant*. W. B. Saunders, Philadelphia, PA, pp. 89–195.
2. Luft,B.J. and Remington,J.S. (1992) Toxoplasmic encephalitis in AIDS patients. *Clin. Infect. Dis.*, **15**, 211–222.
3. Kissinger,J.C., Gajria,B., Li,L., Paulsen,I.T. and Roos,D.S. (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.*, **31**, 234–236.
4. Khan,A., Taylor,S., Su,C., Mackey,A.J., Boyle,J., Cole,R., Glover,D., Tang,K., Paulsen,I.T. *et al.* (2005) Composite genome map and recombination parameters derived from three archetypal lineages of *Toxoplasma gondii*. *Nucleic Acids Res.*, **33**, 2980.
5. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C. *et al.* (2003) PlasmoDB. The *Plasmodium* genome resource: a database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
6. Aurrecoechea,C., Heiges,M., Wang,H., Wang,Z., Fischer,S., Rhodes,P., Miller,J., Kraemer,E., Stoeckert,C.J. *et al.* (2007) ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res.*, **35**, D427–D430.
7. Hu,K., Johnson,J., Florens,L., Fraunholz,M., Suravajjala,S., DiLullo,C., Yates,J., Roos,D.S. and Murray,J.M. (2006) Cytoskeletal components of an invasion machine – the apical complex of *Toxoplasma gondii*. *PLoS Pathog.*, **2**, e13.
8. Bradley,P.J., Ward,C., Cheng,S.J., Alexander,D.L., Coller,S., Coombs,G.H., Dunn,J.D., Ferguson,D.J., Sanderson,S.J. *et al.* (2005) Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii*. *J. Biol. Chem.*, **280**, 34245–34258.
9. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
10. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
11. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.*, **3**, e91.
12. Davidson,S., Crabtree,J., Brunk,B.P., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,C.J. Jr (2001) K2/Klesli and GUS: experiments in integrated access to genomic data sources. *IBM Syst. J.*, **40**, 512–531.
13. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
14. Radke,J.R., Behnke,M.S., Mackey,A.J., Radke,J.B., Roos,D.S. and White,M.W. (2005) The transcriptome of *Toxoplasma gondii*. *BMC Biol.*, **3**, 26.