Nucleic Acids Research Advance Access published September 29, 2008

Nucleic Acids Research, 2008, 1–5
doi:10.1093/nar/gkn631

GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*

Cristina Aurrecoechea[1], John Brestelli[2], Brian P. Brunk[2], Jane M. Carlton[3], Jennifer Dommer[2], Steve Fischer[2], Bindu Gajria[2], Xin Gao[2], Alan Gingle[4], Greg Grant[5], Omar S. Harb[2,*], Mark Heiges[1], Frank Innamorato[2], John Iodice[2], Jessica C. Kissinger[1,6], Eileen Kraemer[7], Wei Li[2], John A. Miller[7], Hilary G. Morrison[8], Vishal Nayak[2], Cary Pennington[1], Deborah F. Pinney[2], David S. Roos[9], Chris Ross[1], Christian J. Stoeckert Jr[2], Steven Sullivan[3], Charles Treatman[2] and Haiming Wang[1]

[1]Center for Tropical & Emerging Global Diseases, University of Georgia, Athens, GA 30602, [2]Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, [3]Department of Medical Parasitology, New York University Langone Medical Center, New York, NY 10010, [4]Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, [5]School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, [6]Department of Genetics, [7]Department of Computer Science, University of Georgia, Athens, GA 30602, [8]Josephine Bay Paul Center [for Comparative Molecular Biology and Evolution], Marine Biological Laboratory, Woods Hole, MA 02543 and [9]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104 USA

Received August 15, 2008; Accepted September 14, 2008

ABSTRACT

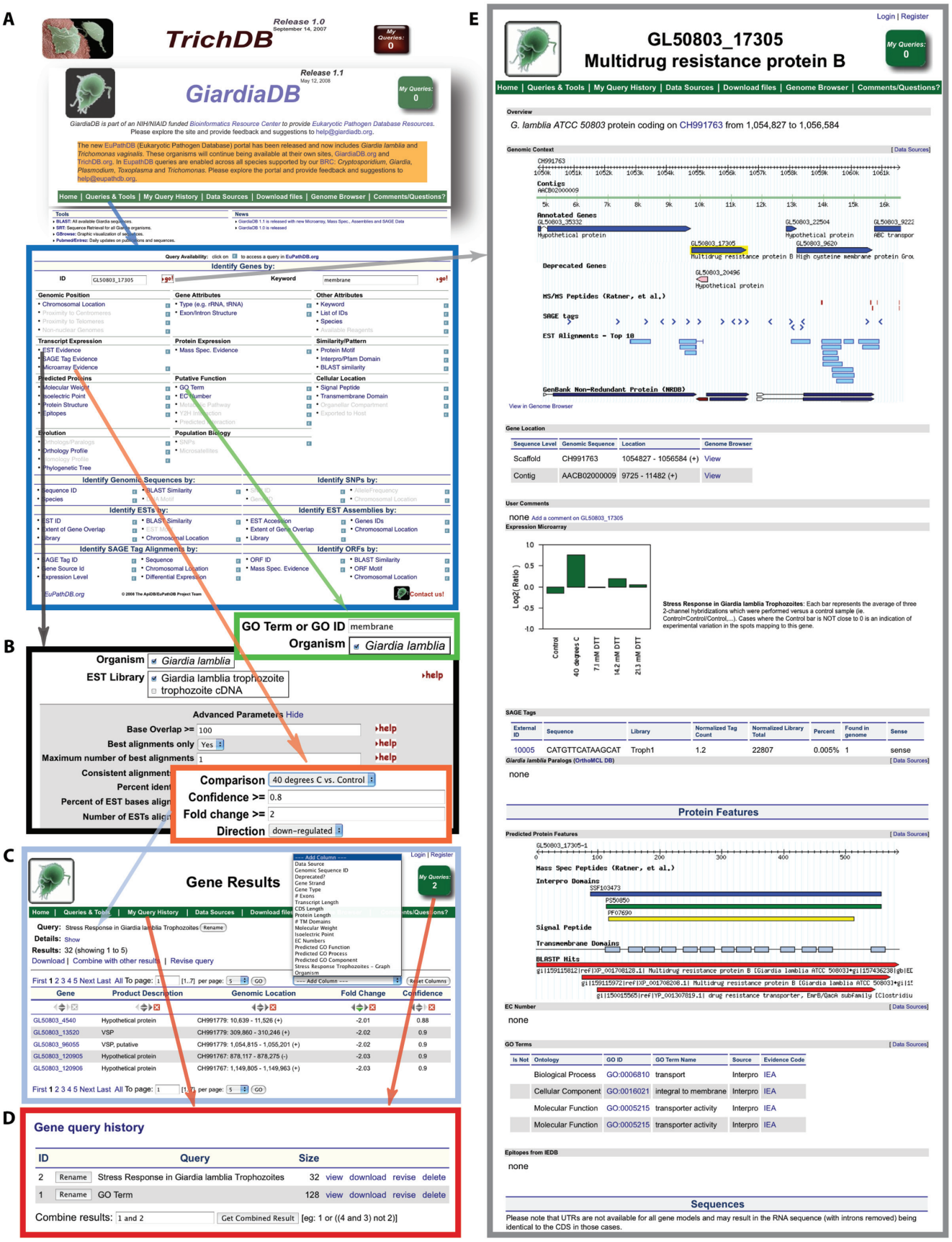
GiardiaDB (http://GiardiaDB.org) and TrichDB (http://TrichDB.org) house the genome databases for *Giardia lamblia* and *Trichomonas vaginalis*, respectively, and represent the latest additions to the EuPathDB (http://EuPathDB.org) family of functional genomic databases. GiardiaDB and TrichDB employ the same framework as other EuPathDB sites (CryptoDB, PlasmoDB and ToxoDB), supporting fully integrated and searchable databases. Genomic-scale data available via these resources may be queried based on BLAST searches, annotation keywords and gene ID searches, GO terms, sequence motifs and other protein characteristics. Functional queries may also be formulated, based on transcript and protein expression data from a variety of platforms. Phylogenetic relationships may also be interrogated. The ability to combine the results from independent queries, and to store queries and query results for future use facilitates complex, genome-wide mining of functional genomic data.

INTRODUCTION

The amitochondriate protists *Giardia lamblia* (*G. intestinalis; G. duodenalis*) and *Trichomonas vaginalis* are ubiquitous microaerophilic parasites. *Giardia lamblia*, a major source of enteric infection in humans and potential bioterrorism agent (category B priority pathogen), is spread through fecal–oral transmission of highly stable cysts, with manifestations such as diarrhea, cramps, bloating, weight loss and maladsorption in symptomatic cases (1). *Trichomonas vaginalis* is the causative agent of trichomoniasis, and is considered the most common nonviral sexually transmitted disease of humans, with ~170 million cases annually (2). This parasite infects the urogenital epithelia of both the sexes, causing inflammation (although men are usually asymptomatic) and increased risk of HIV infection.

The 12 Mb *G. lamblia* genome (3) and ~160 Mb *T. vaginalis* genome (4) have been deposited in GenBank, and are also accessible at GiardiaDB (http://GiardiaDB.org) and TrichDB (http://TrichDB.org), respectively, along with both manually curated automatically generated annotation, and a variety of functional genomics data. Data can may accessed and queried directly via the individual

*To whom correspondence should be addressed. Tel: +1 215 746 7019; Fax: +1 215 573 3111; Email: oharb@pcbi.upenn.edu
Correspondence may also be addressed to Brian P. Brunk. Tel: +1 215 573 3118; Fax: +1 215 573 3111; Email: brunkb@pcbi.upenn.edu; Jessica C. Kissinger. Tel: +1 706 542 6562; Fax: +1 706 542 3582; Email: jkissing@uga.edu; David S. Roos. Tel: +1 215 898 2118; Fax: +1 215 746 6697; Christian J. Stoeckert. Tel: +1 215 573 4409; Fax: +1 215 573 3111; Email: stoeckrt@pcbi.upenn.edu


A
TrichDB
Release 1.0
September 14, 2007
My Queries: 0
GiardiaDB
Release 1.1
May 12, 2008
My Queries: 0
GiardiaDB is part of an NIH/NIAID funded Bioinformatics Resource Center to provide Eukaryotic Pathogen Database Resources.
Please explore the site and provide feedback and suggestions to help@giardiadb.org.
The new EuPathDB (Eukaryotic Pathogen Database) portal has been released and now includes Giardia lamblia and Trichomonas vaginalis. These organisms will continue being available at their own sites, GiardiaDB.org and TrichDB.org. In EupathDB queries are enabled across all species supported by our BRC: Cryptosporidium, Giardia, Plasmodium, Toxoplasma and Trichomonas. Please explore the portal and provide feedback and suggestions to help@eupathdb.org.
Home | Queries & Tools | My Query History | Data Sources | Download files | Genome Browser | Comments/Questions?
Tools
BLAST: All available Giardia sequences
SRT: Sequence Retrieval for all Giardia organisms.
GBrowse: Graphic visualization of sequences.
Pubmed/Entrez: Daily updates on publications and sequences.
News
GiardiaDB 1.1 is released with new Microarray, Mass Spec., Assemblies and SAGE Data
GiardiaDB 1.0 is released
Query Availability: click on to access a query in EuPathDB.org
Identify Genes by:
ID GL50803_17305 go!
Keyword membrane go!
Genomic Position
Chromosomal Location
Proximity to Centromeres
Proximity to Telomeres
Non-nuclear Genomes
Gene Attributes
Type (e.g. rRNA, tRNA)
Exon/Intron Structure
Other Attributes
Keyword
List of IDs
Species
Available Reagents
Transcript Expression
EST Evidence
SAGE Tag Evidence
Microarray Evidence
Protein Expression
Mass Spec. Evidence
Similarity/Pattern
Protein Motif
Interpro/Pfam Domain
BLAST similarity
Predicted Proteins
Molecular Weight
Isoelectric Point
Protein Structure
Epitopes
Putative Function
GO Term
EC Number
Metabolic Pathway
Y2H Interaction
Predicted Interaction
Cellular Location
Signal Peptide
Transmembrane Domain
Organellar Compartment
Exported to Host
Evolution
Orthologs/Paralogs
Orthology Profile
Homology Profile
Phylogenetic Tree
Population Biology
SNPs
Microsatellites
Identify Genomic Sequences by:
Sequence ID
Species
BLAST Similarity
Motif
Identify SNPs by:
SNP ID
Gene
AlleleFrequency
Chromosomal Location
Identify ESTs by:
EST ID
Extent of Gene Overlap
Library
BLAST Similarity
EST Accession
Chromosomal Location
Identify EST Assemblies by:
Genes IDs
Chromosomal Location
Identify SAGE Tag Alignments by:
SAGE Tag ID
Gene Source Id
Expression Level
Sequence
Chromosomal Location
Differential Expression
Identify ORFs by:
ORF ID
Mass Spec. Evidence
BLAST Similarity
ORF Motif
Chromosomal Location
EuPathDB.org
© 2008 The ApiDB/EuPathDB Project Team
Contact us!
GO Term or GO ID membrane
Organism Giardia lamblia
B
Organism Giardia lamblia
EST Library Giardia lamblia trophozoite
trophozoite cDNA
help
Advanced Parameters Hide
Base Overlap >= 100 help
Best alignments only Yes help
Maximum number of best alignments 1 help
Consistent alignment
Percent ident
Percent of EST bases align
Number of ESTs align
Comparison 40 degrees C vs. Control
Confidence >= 0.8
Fold change >= 2
Direction down-regulated
C
Gene Results
Login | Register
My Queries: 2
Home | Queries & Tools | My Query History | Data Sources | Download files | Comments/Questions?
Query: Stress Response in Giardia lamblia Trophozoites Rename
Details: Show
Results: 32 (showing 1 to 5)
Download | Combine with other results | Revise query
First 1 2 3 4 5 Next Last All To page: 1 [1..7] per page: 5 GO
Reset Columns
Add Column
Data Source
Genomic Sequence ID
Deprecated?
Gene Strand
Gene Type
Exons
Transcript Length
CDS Length
Protein Length
TM Domains
Molecular Weight
Isoelectric Point
EC Numbers
Predicted GO Function
Predicted GO Process
Predicted GO Component
Stress Response Trophozoites - Graph
Organism
Gene | Product Description | Genomic Location | Fold Change | Confidence
GL50803_4540 | Hypothetical protein | CH991779: 10,639 - 11,526 (+) | -2.01 | 0.88
GL50803_13520 | VSP | CH991779: 309,860 - 310,246 (+) | -2.02 | 0.9
GL50803_96055 | VSP, putative | CH991779: 1,054,815 - 1,055,201 (+) | -2.02 | 0.9
GL50803_120905 | Hypothetical protein | CH991767: 878,117 - 878,275 (-) | -2.03 | 0.9
GL50803_120906 | Hypothetical protein | CH991767: 1,149,805 - 1,149,963 (+) | -2.03 | 0.9
First 1 2 3 4 5 Next Last All To page: 1 [1..7] per page: 5 GO
D
Gene query history
ID | Query | Size
2 Rename Stress Response in Giardia lamblia Trophozoites 32 view download revise delete
1 Rename GO Term 128 view download revise delete
Combine results: 1 and 2 Get Combined Result [eg: 1 or ((4 and 3) not 2)]
E
Login | Register
GL50803_17305
Multidrug resistance protein B
My Queries: 0
Home | Queries & Tools | My Query History | Data Sources | Download files | Genome Browser | Comments/Questions?
Overview
G. lamblia ATCC 50803 protein coding on CH991763 from 1,054,827 to 1,056,584
Genomic Context [Data Sources]
CH991763
Contigs
AACB02000009
Annotated Genes
GL50803_35332
Hypothetical protein
GL50803_22504
Hypothetical protein
GL50803_17305
Multidrug resistance protein B
GL50803_9620
High cysteine membrane protein
Deprecated Genes
GL50803_20496
Hypothetical protein
MS/MS Peptides (Ratner, et al.)
SAGE tags
EST Alignments - Top 10
GenBank Non-Redundant Protein (NRDB)
View in Genome Browser
Gene Location
Sequence Level | Genomic Sequence | Location | Genome Browser
Scaffold | CH991763 | 1054827 - 1056584 (+) | View
Contig | AACB02000009 | 9725 - 11482 (+) | View
User Comments
none Add a comment on GL50803_17305
Expression Microarray
Log2(Ratio)
Control
40 degrees C
7.1 mM DTT
14.2 mM DTT
21.3 mM DTT
Stress Response in Giardia lamblia Trophozoites: Each bar represents the average of three 2-channel hybridizations which were performed versus a control sample (ie. Control/Control,...). Cases where the Control bar is NOT close to 0 is an indication of experimental variation in the spots mapping to this gene.
SAGE Tags
External ID | Sequence | Library | Normalized Tag Count | Normalized Library Total | Percent | Found in genome | Sense
10005 | CATGTTCATAAGCAT | Troph1 | 1.2 | 22807 | 0.005% | 1 | sense
Giardia lamblia Paralogs (OrthoMCL DB) [Data Sources]
none
Protein Features
Predicted Protein Features [Data Sources]
GL50803_17305-1
Mass Spec Peptides (Ratner, et al.)
Interpro Domains
SSF103473
PS50850
PF07690
Signal Peptide
Transmembrane Domains
BLASTP Hits
EC Number [Data Sources]
none
GO Terms [Data Sources]
Is Not | Ontology | GO ID | GO Term Name | Source | Evidence Code
Biological Process | GO:0006810 | transport | Interpro | IEA
Cellular Component | GO:0016021 | integral to membrane | Interpro | IEA
Molecular Function | GO:0005215 | transporter activity | Interpro | IEA
Molecular Function | GO:0005215 | transporter activity | Interpro | IEA
Epitopes from IEDB
none
Sequences
Please note that UTRs are not available for all gene models and may result in the RNA sequence (with introns removed) being identical to the CDS in those cases.

genome sites or through the Eukaryotic Pathogen DataBase portal (EuPathDB: http://EuPathDB.org) (5), which also accommodates other eukaryotic pathogen databases including CryptoDB (*Cryptosporidium spp.*) (6), PlasmoDB (*Plasmodium spp.*) (7) and ToxoDB (*Toxoplasma gondii*) (8).

DATA CONTENT OF CURRENT RELEASES

GiardiaDB

GiardiaDB release 1.1 is based on the 12 Mb genome of the WBC6 clinical isolate of *G. lamblia*. The sequence is distributed among 306 contigs, assembled on 92 scaffolds (supercontigs), with an average depth of coverage of 11×. A total of 4976 genes have been annotated, including 4889 protein coding genes, 61 tRNAs and 17 rRNAs. In addition, 1611 genes have been flagged as 'deprecated' (demoted) in this release of GiardiaDB as they appear unlikely to represent true genes, based on incompatibility with longer gene models, published data, or alternative models for which functional evidence is available.

Transcript and proteomic expression data sets are both available for analysis through GiardiaDB. These include expressed sequence tag (EST) evidence from the trophozoite life stage (3, and data deposited in dbEST; http://www.ncbi.nlm.nih.gov/dbEST/), and ten serial analysis of gene expression (SAGE) data sets, representing time points distributed throughout the *Giardia* parasite life cycle: trophozoites, encystation, cyst and excystation (9). Understanding various parasite life stages will be critical for vaccine and therapy development strategies, making the SAGE time series particularly valuable. SAGE data may be queried for evidence of gene expression at a particular life stage(s), and also based on relative levels of expression at different stages (differential expression). As part of the Pathogen Functional Genomics Resource Center (http://pfgrc.tigr.org/index.php/microarray/available_microarrays.html), the J. Craig Venter Institute (JCVI) has developed a microarray based on the WBC6 genome which is available to *Giardia* research groups. Results from one completed microarray study identifying genes that are up- or downregulated in response to stress (heat or DTT) are available in the current database (A. Hehl *et al.*, unpublished). Trophozoites during log-phase growth are also represented by mass spectrometry-based proteomics data, with peptide assignments made using the SEQUEST (10) and DTASelect (D. Ratner *et al.*, unpublished data). These data have been used to further validate gene calls and restore a small number of genes from 'deprecated' status.

Because *Giardia* has provided an extraordinarily valuable window into the evolution of eukaryotic cells, GiardiaDB also provides precomputed phylogenetic trees for 1441 genes. Giardial genes and homologs in other eukaryotic organisms were aligned using MUSCLE (11) and phylogenetic relationships inferred using MrBayes (12). While such trees cannot be computed dynamically, these precomputed trees provide a starting point for further analysis.

TrichDB

TrichDB release 1.0 is based on the 160 Mb genome sequence of the G3 isolate of *T. vaginalis* (4) distributed among many contigs and scaffolds. The annotated genome is comprised of 59 672 protein-coding genes (only 65 of which contain introns), 1136 RNA-coding genes (668 rRNA and 438 tRNA) and 38 201 'repeat genes' (protein-coding genes present in high copy number). Transcript expression includes data from 11 *T. vaginalis* EST libraries, courtesy of TvXpress at the Chang Gung Bioinformatics Center (http://tvxpress.cgu.edu.tw/). These represent ESTs from four isolates, grown under different conditions, and from various cell cycle stages.

AVAILABLE QUERIES AND DATA-MINING TOOLS

TrichDB and GiardiaDB are both accessed via the standard EuPathDB web interface, providing a wide variety of tools for genomic database mining. In addition to BLAST (13) and pattern/motif similarity searches, users can identify genes based on genomic position; common name or keyword; gene attributes (such as gene type, or number of exons); evidence of transcript expression including ESTs (both TrichDB and GiardiaDB), SAGE tags, microarray and proteomics (GiardiaDB only); gene product annotation (such as GO function, or EC enzyme number); and predicted cellular location (based on signal peptide and transmembrane predictions). Figure 1A and B illustrates a set of queries supported in GiardiaDB (TrichDB is very similar). Query results are returned as a tabular list, with columns that users can sort, or manipulate by adding or removing attributes to be displayed (Figure 1C). Clicking on any gene identifier links to the gene record page, providing all the information associated with the gene of interest. The right-hand panel of Figure 1E shows a representative gene record page from GiardiaDB.

The EuPathDB infrastructure also provides a set of tools leveraging these basic queries, enabling users to perform higher level operations on their query results. For example, users may use the query history functionality to combine results using the Boolean operators (AND, OR, NOT) (Figure 1D), allowing them to identify genes that possess a specified combination of attributes, such as

Figure 1. Screenshots of queries available through GiardiaDB (similar queries are also available for TrichDB). (**A**) starting with the home page of GiardiaDB (or TrichDB), users may access 'Queries & Tools' via the green navigation bar. (**B**) Highlighted queries focus on EST evidence (black), microarray expression (orange) or GO terms (green), yielding results shown in a 'Gene Results' page (**C**), which can be organized by the addition (or deletion) of data columns (menu inset), or sorted based on any column of interest. Hot-links provide direct access to individual gene record pages of interest. (**D**) The results obtained from any queries run by the user are stored under 'Query History' (green navigation bar), which may be combined using Boolean operations available. (**E**) a sample gene record page, corresponding to the Multidrug Resistance Protein B (GL50803_17305), identified through the Boolean combination shown in (D). Gene record pages may be accessed from several locations, including the Gene Result page C (by clicking on gene IDs of interest), by entering the gene ID on the home page (A, gray arrow), or by clicking on gene glyphs shown in genomic context at the top the gene record page (permits sequential 'walking' along the chromosome to neighboring genes, or more distant genes when viewing within the Genome Browser).

putative kinase genes expressed in trophozoites for which either proteomics or EST evidence is available. Investigators interested in drug target discovery, may wish to search for genes with EST, microarray or proteomics evidence for expression, that appear likely to encode small soluble proteins assigned EC numbers or GO terms associated with catalytic activity, and lack evident orthologs in humans or other mammals. Similar queries against TrichDB might be further refined by asking for protein coding genes that are not highly repeated. The resulting candidate list could be expanded or restricted based on the addition of additional criteria, or parameter refinement.

At any point, users may download the results of any query in various formats, including a detailed report contains all of the data stored for each gene record, enabling further bioinformatics analysis. FASTA format allows users to simply retrieve transcript and/or protein sequences, as along with flanking genomic sequences if desired. For example, users might wish to identify a set of genes and download all sequences that lie within 1000-bp upstream of each. Optional registration and logins enable users to retain their query history over time, so that these results can be further refined, combined with additional queries, or re-run at a later date. Registered users may also submit comments on any gene or sequence entity in the database, providing support for community annotation of these parasite genomes. User comments (labeled as such) are immediately available to other users of the database, and indexed for retrieval using keyword searches.

THE EuPathDB PORTAL

EuPathDB provides a unified query interface for TrichDB and GiardiaDB, as well as other pathogen databases including CryptoDB (supporting three species of *Cryptosporidium*), PlasmoDB (six *Plasmodium* species) and ToxoDB (three *Toxoplasma gondii* strains, and the closely related species *Neospora caninum*). In support of functional and evolutionarily relevant studies, the organism parameter facilitates searches for 'anaerobic protists' (*Cryptosporidium, Giardia*, *Trichomonas*) or 'apicomplexans' (*Crytosporidium, Plasmodium*, *Toxoplasma*), in addition to 'all organisms' or any user-defined subsets of species. All queries available on the component websites are available in EuPathDB, enabling users to leverage orthologous relationships between organisms to identify genes based on data types that may not be available for their organism of primary interest. (Ortholog functionality not available for *T. vaginalis* at the time of manuscript submission, but scheduled for the autumn 2008 release of TrichDB.)

FUTURE DIRECTIONS

GiardiaDB is expected to grow substantially over the coming, year as next-generation sequencing and assembly technologies are now being applied to three new genomes, including a second assemblage A isolate and two assemblage B isolates (14). The ability to query across related genomes will likely identify both a core set of *Giardia* genes, and genes that appear to be strain-specific. Proteomic data sets corresponding to various life cycle stages and subcellular fractions (particularly the ESV excretory/secretory vesicles) are also anticipated (Gillin *et al*., Tachezy *et al*., personal communication).

The next release of TrichDB is expected to incorporate several new data sets and data upgrades including proteomic, phosphoproteomic, microRNA and microarray data, new EST libraries, and transposable element annotation and categorization. MicroRNA and EST data are also expected for other trichomonads including *T. tenax, T. foetus* and *Pentatrichomonas hominis*. Additionally, we anticipate loading and providing access to the genomic sequence and annotation for a second strain of *T. vaginalis* (TO16). The scheduled incorporation of *T. vaginalis* into the OrthoMCL database of orthologous proteins (http://OrthoMCL.org) (15) will allow users to leverage orthology relationships between *T. vaginalis* and other protozoan parasites.

ACKNOWLEDGEMENTS

The authors wish to thank members of the *Giardia* and *Trichomonas* research communities for their willingness to share genomic-scale datasets, often prior to publication, and for numerous comments and suggestions that have helped to improve the functionality of GiardiaDB and TrichDB. We also thank past and present staff associated with the ApiDB-BRC project, and our research laboratory colleagues whose contributions have facilitated the creation and maintenance of this database resource.

FUNDING

Federal funds from the National Institute of Allergy and Infectious Diseases; Department of Health and Human Services, National Institutes of Health (HHSN266200400037C). Funding for open access charge: National Institutes of Health (HHSN266200400037C).

Conflict of interest statement. None declared.

REFERENCES

1. Kucik,C.J., Martin,G.L. and Sortor,B.V. (2004) Common intestinal parasites. *Am. Fam. Physician*, **69**, 1161–1168.
2. Global Prevalence and Incidence of Selected Curable Sexually Transmitted Infections: Overview and Estimates (World Health Organization, Geneva, 2001). http://www.who.int/docstore/hiv/GRSTI/006.htm (July 2008, last date accessed).
3. Morrison,H.G., McArthur,A.G., Gillin,F.D., Aley,S.B., Adam,R.D., Olsen,G.J., Best,A.A., Cande,W.Z., Chen,F., Cipriano,M.J. *et al*. (2007) Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. *Science*, **317**, 1921–1926.
4. Carlton,J.M., Hirt,R.P., Silva,J.C., Delcher,A.L., Schatz,M., Zhao,Q., Wortman,J.R., Bidwell,S.L., Alsmark,U.C., Besteiro,S. *et al*. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science*, **315**, 207–212.
5. Aurrecoechea,C., Heiges,M., Wang,H., Wang,Z., Fischer,S., Rhodes,P., Miller,J., Kraemer,E., Stoeckert,, C.J.,Jr., Roos,D.S. *et al*. (2007) ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res*., **35**, D427–D430.

6. Heiges,M., Wang,H., Robinson,E., Aurrecoechea,C., Gao,X., Kaluskar,N., Rhodes,P., Wang,S., He,C.Z., Su,Y. *et al.* (2006) CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res.*, **34**, D419–D422.
7. 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.
8. Gajria,B., Bahl,A., Brestelli,J., Dommer,J., Fischer,S., Gao,X., Heiges,M., Iodice,J., Kissinger,J.C., Mackey,A.J. *et al.* (2008) ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic Acids Res.*, **36**, D553–D556.
9. Palm,D., Weiland,M., McArthur,A.G., Winiecka-Krusnell,J., Cipriano,M.J., Birkeland,S.R., Pacocha,S.E., Davids,B., Gillin,F., Linder,E. *et al.* (2005) Developmental changes in the adhesive disk during Giardia differentiation. *Mol. Biochem. Parasitol.*, **141**, 199–207.
10. Yates,J.R. III, Eng,J.K., McCormack,A.L. and Schieltz,D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, **67**, 1426–1436.
11. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
12. Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Thompson,R.C. and Monis,P.T. (2004) Variation in Giardia: implications for taxonomy and epidemiology. *Adv. Parasitol.*, **58**, 69–137.
15. 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.