D500–D503 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj054

Tetrahymena Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research

Nicholas A. Stover*, Cynthia J. Krieger, Gail Binkley, Qing Dong, Dianna G. Fisk, Robert Nash, Anand Sethuraman, Shuai Weng and J. Michael Cherry

Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA

Received August 15, 2005; Accepted October 4, 2005

ABSTRACT

We have developed a web-based resource (available at www.ciliate.org) for researchers studying the model ciliate organism *Tetrahymena thermophila*. Employing the underlying database structure and programming of the *Saccharomyces* Genome Database, the *Tetrahymena* Genome Database (TGD) integrates the wealth of knowledge generated by the *Tetrahymena* research community about genome structure, genes and gene products with the newly sequenced macronuclear genome determined by The Institute for Genomic Research (TIGR). TGD provides information curated from the literature about each published gene, including a standardized gene name, a link to the genomic locus in our graphical genome browser, gene product annotations utilizing the Gene Ontology, links to published literature about the gene and more. TGD also displays automatic annotations generated for the gene models predicted by TIGR. A variety of tools are available at TGD for searching the *Tetrahymena* genome, its literature and information about members of the research community.

INTRODUCTION

Research on the ciliated protozoan *Tetrahymena thermophila* (here referred to simply as *Tetrahymena*) has been providing remarkable insights into basic biological principles for over 40 years. Among the most notable discoveries made have been Type-I self-splicing introns, the molecular motor dynein, a histone modification 'code' responsible for different chromatin states and the riboprotein complex telomerase (1). More recently, studies of the RNA interference pathway and tubulin modification in *Tetrahymena* cells continue to reveal intriguing biological processes. These studies all owe their success in part to the fascinating cell biology and life cycle of *Tetrahymena*. *Tetrahymena* are large cells (~30 μm × 50 μm), each covered with hundreds of cilia and containing two nuclei: the huge, polyploid macronucleus, which serves as the somatic nucleus for the cell, and the transcriptionally silent micronucleus, which generates gametes in times of stress. *Tetrahymena*'s extreme amplification of the macronuclear chromosome containing the rDNA (~9000 haploid copies per cell) has been a boon to the study of chromosome features, such as telomeres and DNA replication origins, whereas the dense collection of cilia and basal bodies at the cell surface has made it a premier system for dissecting the components of these structures.

The *Tetrahymena* Genome Database (TGD; www.ciliate.org) was established by the *Tetrahymena* Macronuclear Genome Project as a manually curated and updated resource linking past and future *Tetrahymena* research to the newly available genome sequence. By working in close collaboration with the staff of the *Saccharomyces* Genome Database (SGD), we have developed TGD by making only minor modifications to the software and database environment of the SGD project (2). In addition to providing the bioinformatic tools and annotation efforts described below, TGD serves as a community information center, with lists of upcoming meetings, a primer on ciliate biology and contact information for members of the *Tetrahymena* research community.

GENE CURATION AND DISPLAY

To date, over 250 actively studied genes in the *Tetrahymena* macronuclear genome have been named and described by authors in the research community. Each of these published genes is presented on a TGD web page containing basic information we have collected about that gene, plus links to other TGD pages with more information and links out to other resources. For each published gene we have (i) identified a standardized gene name that conforms to the published *Tetrahymena* nomenclature guidelines (3), plus any aliases used for that gene in the literature; (ii) collected relevant

*To whom correspondence should be addressed. Tel: +1 650 724 9959; Fax: +1 650 725 1534; Email: nick@genome.stanford.edu literature citations and provided links to these papers' entries at Pubmed; (iii) written short, free-text descriptions summarizing the knowledge about the gene; and (iv) provided a link to one or more GenBank entries for the gene, to allow users to access the sequence of the gene described in the literature.

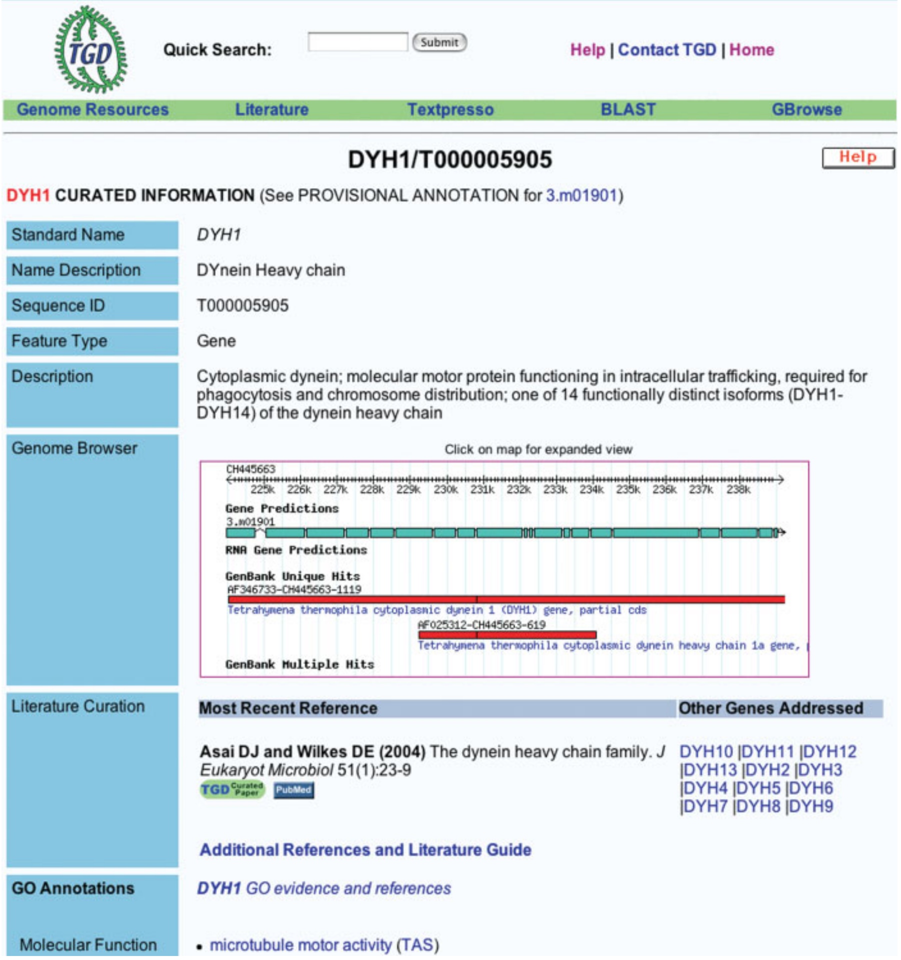
We are currently annotating the molecular function, biological process and cellular localization of each gene product for which data are available, using terms from the Gene Ontology (GO) (4). The GO is a controlled vocabulary used by many model organism databases to describe these gene features in a species-independent fashion, allowing users to easily search, sort and compare genes in diverse organisms. An integral part of performing these annotations is updating the GO with terms necessary to describe the biology of ciliates. TGD has already contributed a number of GO terms and definitions, including new terms for processes related to the RNA interference pathway, nuclear dimorphism, translation termination and other important areas of ciliate biology.

SEQUENCE AND GENE MODELS

The Institute for Genomic Research (TIGR) has sequenced the *Tetrahymena* macronuclear genome, in an effort led by Jonathan Eisen. The ∼106 megabase haploid macronuclear genome is predicted to be arranged into 250–300 chromosomes; the sequence closure effort at TIGR is ongoing, with over 40% of the genomic sequence assembled into complete chromosomes bounded by telomeres at both ends. TIGR has submitted the genome sequences to the Whole Genome Shotgun depository at National Center for Biotechnology Information (NCBI) under accession number AAGF00000000.

TIGR has released a preliminary set of gene model predictions using their genome scaffolds. The number of protein-coding genes predicted in their analysis is 27 400. The gene models currently available have not been manually reviewed and, reflecting their preliminary nature, may contain inaccuracies in their coding sequence boundaries when compared to cDNA sequences. Nonetheless, the preliminary gene models have proven to be very useful to the *Tetrahymena* research community, and studies identifying new genes and gene families based on these gene models are already being published (5–10).

In order to accommodate the ongoing gene model refinements at TIGR, we have adopted a two-tiered gene page, an example of which is shown in Figures 1 and 2. The upper section of the page (Figure 1) displays the information we



DYH1/T000005905

DYH1 **CURATED INFORMATION** (See PROVISIONAL ANNOTATION for 3.m01901)

Standard Name	*DYH1*
Name Description	DYnein Heavy chain
Sequence ID	T000005905
Feature Type	Gene
Description	Cytoplasmic dynein; molecular motor protein functioning in intracellular trafficking, required for phagocytosis and chromosome distribution; one of 14 functionally distinct isoforms (DYH1-DYH14) of the dynein heavy chain

Genome Browser

Click on map for expanded view

Literature Curation

Most Recent Reference	Other Genes Addressed
Asai DJ and Wilkes DE (2004) The dynein heavy chain family. *J Eukaryot Microbiol* 51(1):23-9 TGD Curated Paper PubMed	DYH10 |DYH11 |DYH12 |DYH13 |DYH2 |DYH3 |DYH4 |DYH5 |DYH6 |DYH7 |DYH8 |DYH9

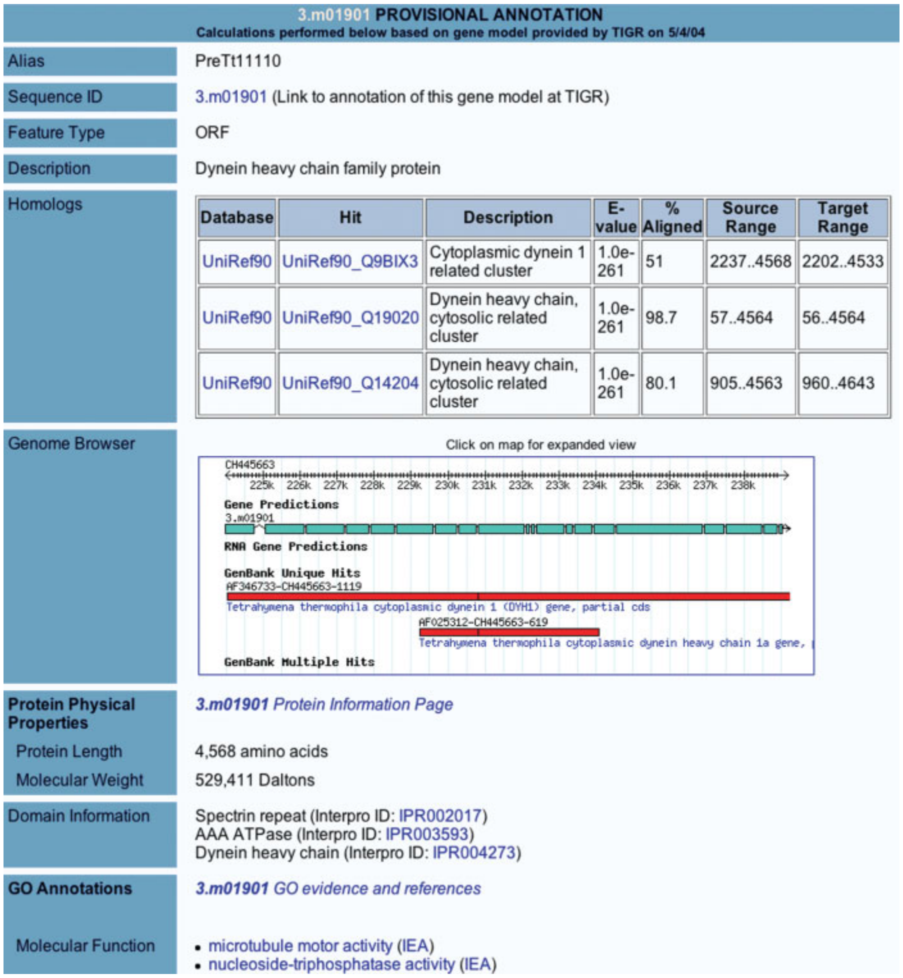
Additional References and Literature Guide

GO Annotations *DYH1 GO evidence and references*

Molecular Function

- microtubule motor activity (TAS)

Figure 1. The upper section of the gene page for the *Tetrahymena* dynein heavy chain *DYH1* presents information about the published gene and its product, including its standard name and aliases, a short description, a graphical display of the gene, GO annotations and links to its literature and gene sequence at GenBank.



3.m01901 PROVISIONAL ANNOTATION

Calculations performed below based on gene model provided by TIGR on 5/4/04

Alias	PreTt11110
Sequence ID	3.m01901 (Link to annotation of this gene model at TIGR)
Feature Type	ORF
Description	Dynein heavy chain family protein

Homologs

Database	Hit	Description	E-value	% Aligned	Source Range	Target Range
UniRef90	UniRef90_Q9BIX3	Cytoplasmic dynein 1 related cluster	1.0e-261	51	2237..4568	2202..4533
UniRef90	UniRef90_Q19020	Dynein heavy chain, cytosolic related cluster	1.0e-261	98.7	57..4564	56..4564
UniRef90	UniRef90_Q14204	Dynein heavy chain, cytosolic related cluster	1.0e-261	80.1	905..4563	960..4643

Genome Browser

Click on map for expanded view

CH445663
225k 226k 227k 228k 229k 230k 231k 232k 233k 234k 235k 236k 237k 238k
Gene Predictions
3.m01901
RNA Gene Predictions
GenBank Unique Hits
AF346733-CH445663-1119
Tetrahymena thermophila cytoplasmic dynein 1 (DYH1) gene, partial cds
AF025312-CH445663-619
Tetrahymena thermophila cytoplasmic dynein heavy chain 1a gene,
GenBank Multiple Hits

Protein Physical Properties — 3.m01901 Protein Information Page

Protein Length	4,568 amino acids
Molecular Weight	529,411 Daltons

Domain Information

Spectrin repeat (Interpro ID: IPR002017)
AAA ATPase (Interpro ID: IPR003593)
Dynein heavy chain (Interpro ID: IPR004273)

GO Annotations — 3.m01901 GO evidence and references

Molecular Function

- microtubule motor activity (IEA)
- nucleoside-triphosphatase activity (IEA)

Figure 2. The lower section of the gene page for *DYH1* presents computational annotation of gene model 3.m01901, the preliminary gene model corresponding to the *DYH1* gene. The information displayed includes a short description of the 3.m01901 gene product provided by TIGR, a link to its gene model page at TIGR, its top three BLASTP hits against the UniRef90 protein database, a graphical display of the gene model, protein physical properties, protein domains predicted using InterProScan and automatic GO annotations based on its predicted protein domains.

have curated from the literature about a gene, plus a link to its published sequence at GenBank. The lower section of the page (Figure 2) displays the automatic annotations generated by TIGR and TGD for the corresponding gene model. As TIGR releases updates, information curated for particular genes can be linked to alternate gene models as appropriate. If there is no published information for a gene corresponding to a given gene model, the gene page will only display information about the gene model. We have found this to be an effective method for displaying the hypothetical gene models and automatic, non-reviewed annotations of a newly sequenced genome, while simultaneously presenting and maintaining the integrity of information published about a slightly different coding sequence.

TOOLS

We have created a server and an interface at our website for BLAST (11) and BLAT (12) searches against a number of relevant sequence datasets. Searching against the *Tetrahymena* macronuclear genome scaffold sequences produces a graphical view of the target sequence region in the genome, in addition to the sequence alignment for this region. This allows the user to see annotations of predicted gene models found in the region, directly from the BLAST/BLAT results page. The graphic is hyperlinked to GBrowse, a graphical genome browser utility available from the Generic Model Organism Database (GMOD) Construction Set (13), which TGD uses to display *Tetrahymena* genome data. GBrowse allows a rich display of annotations to the genome sequence using a combination of text and icons, which is particularly useful for showing large-scale data in the context of the genome sequence. Proteomic analyses have identified the composition of different organelles and structures (6,14), and microarray expression analyses are anticipated by the ciliate research community.

TIGR and TGD have combined to provide a basic set of annotations for each preliminary gene model. TGD determined the protein domain composition of each model using the InterProScan utility (15), and performed a BLASTP comparison of TIGR's preliminary gene models against the UniRef90 protein database (16). Gene models shown in GBrowse are labeled with the domains or top BLASTP hit determined by TGD, together with the gene product description provided by TIGR's automatic annotation. Expanded domain, homolog and gene product description data are shown on the preliminary model section of each gene page, plus computationally determined GO annotations based on its protein domain composition. A link to TIGR's analyses of the gene models is found on each of the gene pages. Information in TGD can be accessed using TGD's Quick Search utility, which allows users to query keywords in the following fields: gene names and aliases, gene descriptions, GO terms, predicted domains, homologs, paper abstracts, colleagues and authors of references in TGD.

TGD provides full-text searching of *Tetrahymena*-related texts available in electronic format via Textpresso, a full-text literature search and information extraction tool available from GMOD (17). In addition to keyword searching, Textpresso allows users to search for the coincidence of keywords and terms from a number of defined categories, in a single sentence or article. Textpresso at TGD has a number of ciliate-related terms added to these categories (i.e. species names and nuclear terms) and searches papers that use the word '*Tetrahymena*' in their keywords or abstract. Currently over 1100 full-text articles, 3200 abstracts and 5000 titles can be searched at TGD.

SUMMARY

TGD serves the *Tetrahymena* research community by linking the many years of published *Tetrahymena* research data to the recently completed macronuclear genome sequence. The resources we have developed are freely available at www.ciliate.org. TGD's close relationship with the *Saccharomyces* Genome Database allows it to quickly adopt tools and displays created by SGD to deliver a wide variety of yeast data. Please contact the TGD curators at ciliate-curator@genome.stanford.edu with any comments or suggestions.

ACKNOWLEDGEMENTS

We would like to thank Jonathan Eisen for sharing the *Tetrahymena* genome data at www.tigr.org/tdb/e2k1/ttg/ prior to publication, and for support and funding of the TGD project. We would also like to thank the staff members at SGD, the staff at the *Candida* Genome Database and the entire *Tetrahymena* research community for their aid and support. TGD is part of the *Tetrahymena* Macronuclear Genome Sequencing Project and is funded by NIGMS Grant 5R01GM067012-03 to Jonathan Eisen at TIGR. Sequencing of the *Tetrahymena* macronuclear genome received additional funding from NSF Grant 0240361 to Jonathan Eisen. Funding to pay the Open Access publication charges for this article was provided by NIGMS Grant 5R01GM067012-03.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,K. and Gorovsky,M.A. (2005) *Tetrahymena thermophila*. *Curr. Biol.*, **15**, R317–R318.
2. Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E., Hong,E.L., Nash,R. *et al.* (2005) Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD). *Nucleic Acids Res.*, **33**, D374–D377.
3. Allen,S.L. (2000) Genetic nomenclature rules for *Tetrahymena thermophila*. *Methods Cell Biol.*, **62**, 561–563.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
5. Janke,C., Rogowski,K., Wloga,D., Regnard,C., Kajava,A.V., Strub,J.M., Temurak,N., van Dijk,J., Boucher,D., van Dorsselaer,A. *et al.* (2005) Tubulin polyglutamylase enzymes are members of the TTL domain protein family. *Science*, **308**, 1758–1762.
6. Bowman,G.R., Smith,D.G., Michael Siu,K.W., Pearlman,R.E. and Turkewitz,A.P. (2005) Genomic and proteomic evidence for a second family of dense core granule cargo proteins in *Tetrahymena thermophila*. *J. Eukaryot. Microbiol.*, **52**, 291–297.
7. Mochizuki,K. and Gorovsky,M.A. (2005) A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev.*, **19**, 77–89.
8. Stemm-Wolf,A.J., Morgan,G., Giddings,T.H.,Jr, White,E.A., Marchione,R., McDonald,H.B. and Winey,M. (2005) Basal body duplication and maintenance require one member of the *Tetrahymena thermophila* centrin gene family. *Mol. Biol. Cell*, **16**, 3606–3619.
9. Stover,N.A., Cavalcanti,A.R., Li,A.J., Richardson,B.C. and Landweber,L.F. (2005) Reciprocal fusions of two genes in the formaldehyde detoxification pathway in ciliates and diatoms. *Mol. Biol. Evol.*, **22**, 1539–1542.
10. Williams,S.A. and Gavin,R.H. (2005) Myosin genes in *Tetrahymena*. *Cell Motil. Cytoskeleton*, **61**, 237–243.
11. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
12. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
13. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. and Lewis,S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
14. Smith,J.C., Northey,J.G., Garg,J., Pearlman,R.E. and Siu,K.W. (2005) Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of *Tetrahymena thermophila*. *J. Proteome Res.*, **4**, 909–919.
15. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
16. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
17. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.

literature citations and provided links to these papers' entries at Pubmed; (iii) written short, free-text descriptions summarizing the knowledge about the gene; and (iv) provided a link to one or more GenBank entries for the gene, to allow users to access the sequence of the gene described in the literature.

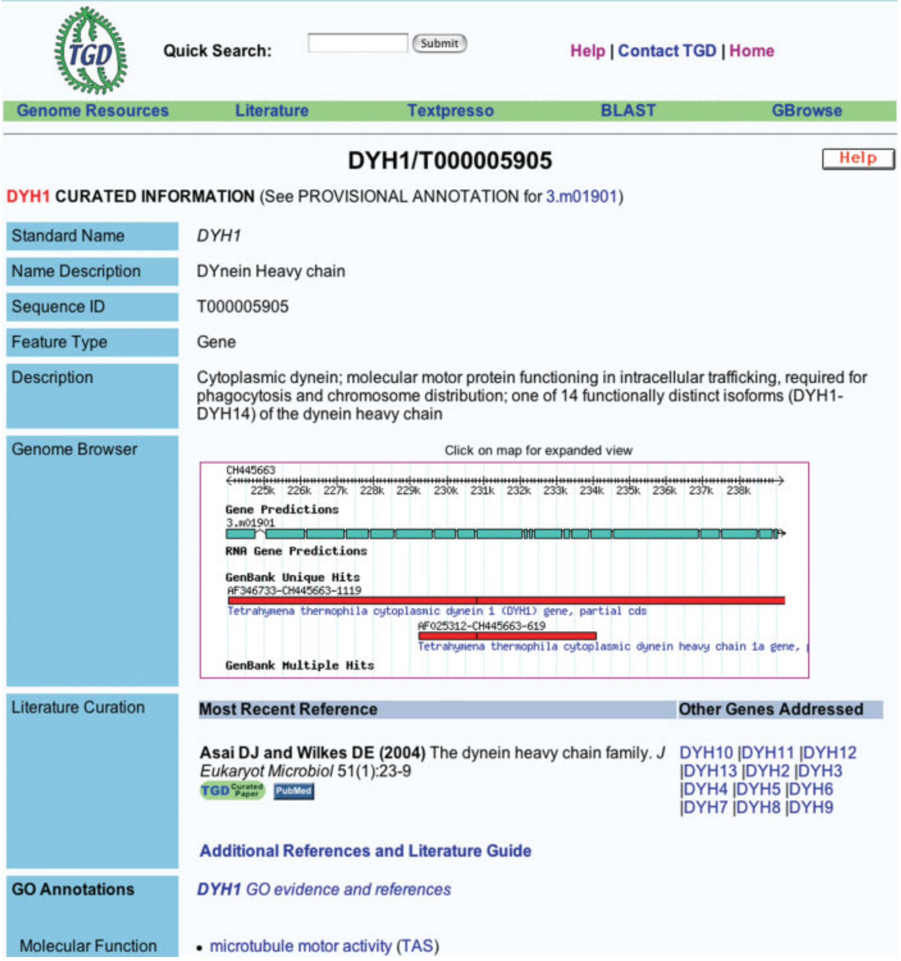
We are currently annotating the molecular function, biological process and cellular localization of each gene product for which data are available, using terms from the Gene Ontology (GO) (4). The GO is a controlled vocabulary used by many model organism databases to describe these gene features in a species-independent fashion, allowing users to easily search, sort and compare genes in diverse organisms. An integral part of performing these annotations is updating the GO with terms necessary to describe the biology of ciliates. TGD has already contributed a number of GO terms and definitions, including new terms for processes related to the RNA interference pathway, nuclear dimorphism, translation termination and other important areas of ciliate biology.

SEQUENCE AND GENE MODELS

The Institute for Genomic Research (TIGR) has sequenced the *Tetrahymena* macronuclear genome, in an effort led by Jonathan Eisen. The ∼106 megabase haploid macronuclear genome is predicted to be arranged into 250–300 chromosomes; the sequence closure effort at TIGR is ongoing, with over 40% of the genomic sequence assembled into complete chromosomes bounded by telomeres at both ends. TIGR has submitted the genome sequences to the Whole Genome Shotgun depository at National Center for Biotechnology Information (NCBI) under accession number AAGF00000000.

TIGR has released a preliminary set of gene model predictions using their genome scaffolds. The number of protein-coding genes predicted in their analysis is 27 400. The gene models currently available have not been manually reviewed and, reflecting their preliminary nature, may contain inaccuracies in their coding sequence boundaries when compared to cDNA sequences. Nonetheless, the preliminary gene models have proven to be very useful to the *Tetrahymena* research community, and studies identifying new genes and gene families based on these gene models are already being published (5–10).

In order to accommodate the ongoing gene model refinements at TIGR, we have adopted a two-tiered gene page, an example of which is shown in Figures 1 and 2. The upper section of the page (Figure 1) displays the information we



DYH1/T000005905

DYH1 CURATED INFORMATION (See PROVISIONAL ANNOTATION for 3.m01901)

Standard Name	*DYH1*
Name Description	DYnein Heavy chain
Sequence ID	T000005905
Feature Type	Gene
Description	Cytoplasmic dynein; molecular motor protein functioning in intracellular trafficking, required for phagocytosis and chromosome distribution; one of 14 functionally distinct isoforms (DYH1-DYH14) of the dynein heavy chain
Genome Browser	Click on map for expanded view

CH445663
225k 226k 227k 228k 229k 230k 231k 232k 233k 234k 235k 236k 237k 238k
Gene Predictions
3.m01901
RNA Gene Predictions
GenBank Unique Hits
AF346733-CH445663-1119
Tetrahymena thermophila cytoplasmic dynein 1 (DYH1) gene, partial cds
AF025312-CH445663-619
Tetrahymena thermophila cytoplasmic dynein heavy chain 1a gene,
GenBank Multiple Hits

Literature Curation

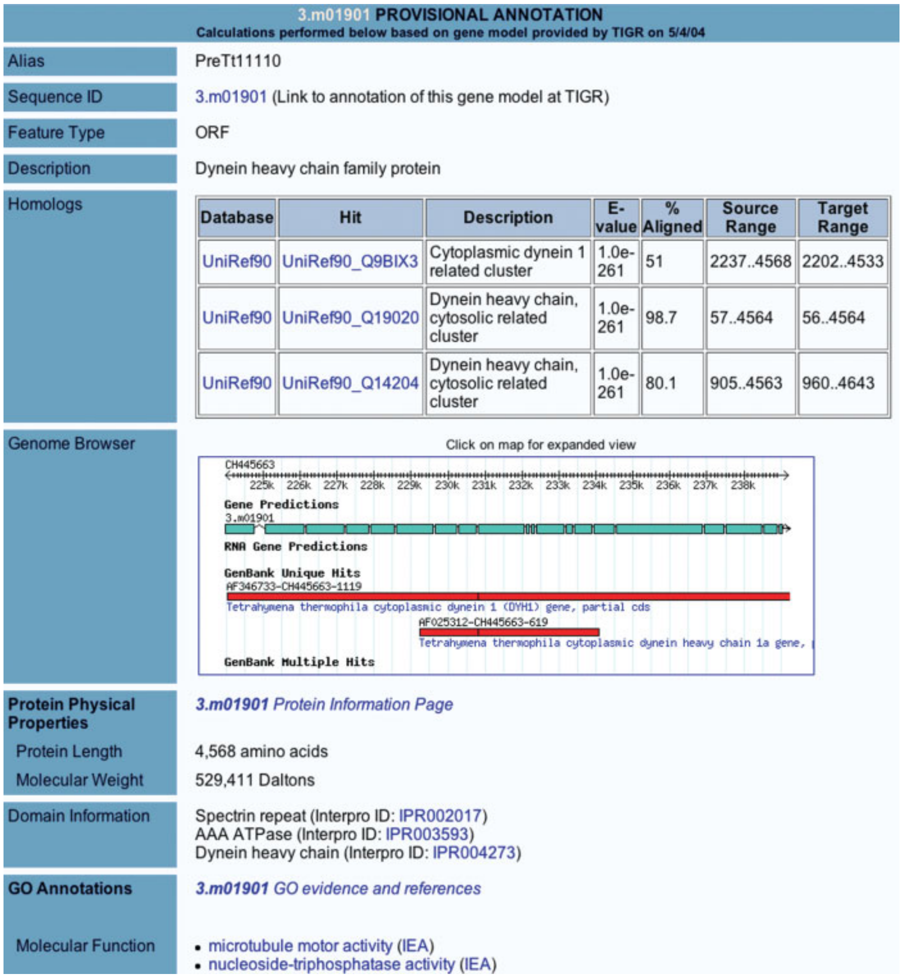
Most Recent Reference	Other Genes Addressed
Asai DJ and Wilkes DE (2004) The dynein heavy chain family. *J Eukaryot Microbiol* 51(1):23-9 TGD Curated Paper PubMed	DYH10 |DYH11 |DYH12 |DYH13 |DYH2 |DYH3 |DYH4 |DYH5 |DYH6 |DYH7 |DYH8 |DYH9

Additional References and Literature Guide

GO Annotations *DYH1 GO evidence and references*

Molecular Function
- microtubule motor activity (TAS)

Figure 1. The upper section of the gene page for the *Tetrahymena* dynein heavy chain *DYH1* presents information about the published gene and its product, including its standard name and aliases, a short description, a graphical display of the gene, GO annotations and links to its literature and gene sequence at GenBank.



3.m01901 PROVISIONAL ANNOTATION
Calculations performed below based on gene model provided by TIGR on 5/4/04

Alias: PreTt11110

Sequence ID: 3.m01901 (Link to annotation of this gene model at TIGR)

Feature Type: ORF

Description: Dynein heavy chain family protein

Homologs:

Database	Hit	Description	E-value	% Aligned	Source Range	Target Range
UniRef90	UniRef90_Q9BIX3	Cytoplasmic dynein 1 related cluster	1.0e-261	51	2237..4568	2202..4533
UniRef90	UniRef90_Q19020	Dynein heavy chain, cytosolic related cluster	1.0e-261	98.7	57..4564	56..4564
UniRef90	UniRef90_Q14204	Dynein heavy chain, cytosolic related cluster	1.0e-261	80.1	905..4563	960..4643

Genome Browser: Click on map for expanded view

Protein Physical Properties: 3.m01901 Protein Information Page

Protein Length: 4,568 amino acids

Molecular Weight: 529,411 Daltons

Domain Information:
Spectrin repeat (Interpro ID: IPR002017)
AAA ATPase (Interpro ID: IPR003593)
Dynein heavy chain (Interpro ID: IPR004273)

GO Annotations: 3.m01901 GO evidence and references

Molecular Function:

- microtubule motor activity (IEA)
- nucleoside-triphosphatase activity (IEA)

Figure 2. The lower section of the gene page for *DYH1* presents computational annotation of gene model 3.m01901, the preliminary gene model corresponding to the *DYH1* gene. The information displayed includes a short description of the 3.m01901 gene product provided by TIGR, a link to its gene model page at TIGR, its top three BLASTP hits against the UniRef90 protein database, a graphical display of the gene model, protein physical properties, protein domains predicted using InterProScan and automatic GO annotations based on its predicted protein domains.

have curated from the literature about a gene, plus a link to its published sequence at GenBank. The lower section of the page (Figure 2) displays the automatic annotations generated by TIGR and TGD for the corresponding gene model. As TIGR releases updates, information curated for particular genes can be linked to alternate gene models as appropriate. If there is no published information for a gene corresponding to a given gene model, the gene page will only display information about the gene model. We have found this to be an effective method for displaying the hypothetical gene models and automatic, non-reviewed annotations of a newly sequenced genome, while simultaneously presenting and maintaining the integrity of information published about a slightly different coding sequence.

TOOLS

We have created a server and an interface at our website for BLAST (11) and BLAT (12) searches against a number of relevant sequence datasets. Searching against the *Tetrahymena* macronuclear genome scaffold sequences produces a graphical view of the target sequence region in the genome, in addition to the sequence alignment for this region. This allows the user to see annotations of predicted gene models found in the region, directly from the BLAST/BLAT results page. The graphic is hyperlinked to GBrowse, a graphical genome browser utility available from the Generic Model Organism Database (GMOD) Construction Set (13), which TGD uses to display *Tetrahymena* genome data. GBrowse allows a rich display of annotations to the genome sequence using a combination of text and icons, which is particularly useful for showing large-scale data in the context of the genome sequence. Proteomic analyses have identified the composition of different organelles and structures (6,14), and microarray expression analyses are anticipated by the ciliate research community.

TIGR and TGD have combined to provide a basic set of annotations for each preliminary gene model. TGD determined the protein domain composition of each model using the InterProScan utility (15), and performed a BLASTP comparison of TIGR's preliminary gene models against the UniRef90 protein database (16). Gene models shown in GBrowse are labeled with the domains or top BLASTP hit determined by TGD, together with the gene product description provided by TIGR's automatic annotation. Expanded domain, homolog and gene product description data are shown on the preliminary model section of each gene page, plus computationally determined GO annotations based on its protein domain composition. A link to TIGR's analyses of the gene models is found on each of the gene pages. Information in TGD can be accessed using TGD's Quick Search utility, which allows users to query keywords in the following fields: gene names and aliases, gene descriptions, GO terms, predicted domains, homologs, paper abstracts, colleagues and authors of references in TGD.

TGD provides full-text searching of *Tetrahymena*-related texts available in electronic format via Textpresso, a full-text literature search and information extraction tool available from GMOD (17). In addition to keyword searching, Textpresso allows users to search for the coincidence of keywords and terms from a number of defined categories, in a single sentence or article. Textpresso at TGD has a number of ciliate-related terms added to these categories (i.e. species names and nuclear terms) and searches papers that use the word '*Tetrahymena*' in their keywords or abstract. Currently over 1100 full-text articles, 3200 abstracts and 5000 titles can be searched at TGD.

SUMMARY

TGD serves the *Tetrahymena* research community by linking the many years of published *Tetrahymena* research data to the recently completed macronuclear genome sequence. The resources we have developed are freely available at www.ciliate.org. TGD's close relationship with the *Saccharomyces* Genome Database allows it to quickly adopt tools and displays created by SGD to deliver a wide variety of yeast data. Please contact the TGD curators at ciliate-curator@genome.stanford.edu with any comments or suggestions.

ACKNOWLEDGEMENTS

We would like to thank Jonathan Eisen for sharing the *Tetrahymena* genome data at www.tigr.org/tdb/e2k1/ttg/ prior to publication, and for support and funding of the TGD project. We would also like to thank the staff members at SGD, the staff at the *Candida* Genome Database and the entire *Tetrahymena* research community for their aid and support. TGD is part of the *Tetrahymena* Macronuclear Genome Sequencing Project and is funded by NIGMS Grant 5R01GM067012-03 to Jonathan Eisen at TIGR. Sequencing of the *Tetrahymena* macronuclear genome received additional funding from NSF Grant 0240361 to Jonathan Eisen. Funding to pay the Open Access publication charges for this article was provided by NIGMS Grant 5R01GM067012-03.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,K. and Gorovsky,M.A. (2005) *Tetrahymena thermophila*. *Curr. Biol*., **15**, R317–R318.
2. Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E., Hong,E.L., Nash,R. *et al*. (2005) Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD). *Nucleic Acids Res*., **33**, D374–D377.
3. Allen,S.L. (2000) Genetic nomenclature rules for *Tetrahymena thermophila*. *Methods Cell Biol*., **62**, 561–563.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
5. Janke,C., Rogowski,K., Wloga,D., Regnard,C., Kajava,A.V., Strub,J.M., Temurak,N., van Dijk,J., Boucher,D., van Dorsselaer,A. *et al*. (2005) Tubulin polyglutamylase enzymes are members of the TTL domain protein family. *Science*, **308**, 1758–1762.
6. Bowman,G.R., Smith,D.G., Michael Siu,K.W., Pearlman,R.E. and Turkewitz,A.P. (2005) Genomic and proteomic evidence for a second family of dense core granule cargo proteins in *Tetrahymena thermophila*. *J. Eukaryot. Microbiol*., **52**, 291–297.
7. Mochizuki,K. and Gorovsky,M.A. (2005) A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev*., **19**, 77–89.
8. Stemm-Wolf,A.J., Morgan,G., Giddings,T.H.,Jr, White,E.A., Marchione,R., McDonald,H.B. and Winey,M. (2005) Basal body duplication and maintenance require one member of the *Tetrahymena thermophila* centrin gene family. *Mol. Biol. Cell*, **16**, 3606–3619.
9. Stover,N.A., Cavalcanti,A.R., Li,A.J., Richardson,B.C. and Landweber,L.F. (2005) Reciprocal fusions of two genes in the formaldehyde detoxification pathway in ciliates and diatoms. *Mol. Biol. Evol*., **22**, 1539–1542.
10. Williams,S.A. and Gavin,R.H. (2005) Myosin genes in *Tetrahymena*. *Cell Motil. Cytoskeleton*, **61**, 237–243.
11. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol*., **266**, 460–480.
12. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
13. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. and Lewis,S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
14. Smith,J.C., Northey,J.G., Garg,J., Pearlman,R.E. and Siu,K.W. (2005) Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of *Tetrahymena thermophila*. *J. Proteome Res*., **4**, 909–919.
15. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
16. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
17. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*., **2**, e309.