



Original article

# ***Tetrahymena* Comparative Genomics Database (TCGD): a community resource for *Tetrahymena***

Wentao Yang<sup>1,2</sup>, Chuanqi Jiang<sup>1,2</sup>, Ying Zhu<sup>5</sup>, Kai Chen<sup>1,2</sup>, Guangying Wang<sup>1,2</sup>, Dongxia Yuan<sup>1</sup>, Wei Miao<sup>1,2,3,4,\*</sup> and Jie Xiong<sup>1,\*</sup>

<sup>1</sup>Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China, <sup>2</sup>University of Chinese Academy of Sciences, Beijing 100049, China, <sup>3</sup>State Key Laboratory of Freshwater Ecology and Biotechnology, Wuhan 430072, China, <sup>4</sup>CAS Center for Excellence in Animal Evolution and Genetics, Kunming 650223, China and <sup>5</sup>Nextomics Biosciences Institute, Wuhan 430000, China

\*Corresponding author: Tel: +86 27 68780209; Email: xiongjie@ihb.ac.cn

\*Correspondence may also be addressed to Wei Miao. Tel: +86 27 68780050; Email: miaowei@ihb.ac.cn

Citation details: Yang,W., Jiang,C., Zhu,Y. *et al.* *Tetrahymena* Comparative Genomics Database (TCGD): a community resource for *Tetrahymena*. *Database* (2019) Vol. 2019: article ID baz029; doi:10.1093/database/baz029

Received 14 November 2018; Revised 4 February 2019; Accepted 6 February 2019

## **Abstract**

Ciliates are a large and diverse group of unicellular organisms characterized by having the following two distinct type of nuclei within a single cell: micronucleus (MIC) and macronucleus (MAC). Although the genomes of several ciliates in different groups have been sequenced, comparative genomics data for multiple species within a ciliate genus are not yet available. Here we collected the genome information and comparative genomics analysis results for 10 species in the *Tetrahymena* genus, including the previously sequenced model organism *Tetrahymena thermophila* and 9 newly sequenced species, and constructed a genus-level comparative analysis platform, the *Tetrahymena* Comparative Genomics Database (TCGD). Genome sequences, transcriptomic data, gene models, functional annotation, ortholog groups and synteny maps were built into this database and a user-friendly interface was developed for searching, visualizing and analyzing these data. In summary, the TCGD (<http://ciliate.ihb.ac.cn>) will be an important and useful resource for the ciliate research community.

**Database URL:** <http://ciliate.ihb.ac.cn>

## **Introduction**

Ciliates are a large and diverse group of unicellular organisms (1). Within the cytoplasm of a single cell, ciliates have the following two types of differentiated nuclei that are structurally and functionally distinct: micronucleus (MIC)

and macronucleus (MAC) (2). The MAC has exclusively somatic functions and directs gene expression (3). So far, draft MAC genomes of several ciliates have been sequenced in different groups (e.g. class) of ciliates, such as the *Tetrahymena thermophila* and *Paramecium*

*tetraurelia* in Oligohymenophorea, *Euplotes octocarinatus*, *Stylonychia lemnae* and *Oxytricha trifallax* in Spirotrichea, *Stentor coeruleus* in Heterotrichida (4–9). Individual genome databases have been established for the ciliate community, such as the TGD Wiki (<http://ciliate.org>), ParameciumDB (<http://paramecium.i2bc.paris-saclay.fr/>) and EOGD (<http://ciliates.ihb.ac.cn/database/home/#eo>; 10–12). However, there is still a lack of comparative genomics database for closely related species within a ciliate genus similar to DroSpeGe (<http://arthropods.eugenescience.org/species/>; 13) or PlasmODB (<http://plasmodb.org/plasmo/>; 14).

The genus *Tetrahymena* contains more than 40 named species (15). Some species in this genus have a long and glorious history as model organisms, such as *T. thermophila*, a well-studied eukaryotic model organism (16), and *Tetrahymena pyriformis*, the most commonly used ciliated model for toxicology research (17). In 2006, the *T. thermophila* MAC genome was sequenced (4), and its corresponding database was established accessing to the community (TGD Wiki; <http://ciliate.org>). We subsequently constructed a functional genomics database TetraFGD (<http://tfgd.ihb.ac.cn/>; 18). Initialized in 2010, we sequenced and performed a comparative genomics analysis for 10 species in *Tetrahymena*, including the previous sequenced model organism *T. thermophila*, and 9 newly sequenced species. These 10 species are distributed among different subgroups and species complexes within the *Tetrahymena* genus. Among the nine newly sequenced species, three of them (*Tetrahymena malaccensis*, *Tetrahymena borealis* and *Tetrahymena ellioti*) were sequenced by the Broad Institute, and the rest six species (*T. pyriformis*, *Tetrahymena vorax*, *Tetrahymena canadensis*, *Tetrahymena empidokyrea*, *Tetrahymena shanghaiensis* and *Tetrahymena paravorax*) were sequenced by the Institute of Hydrobiology, Chinese Academy of Sciences. To facilitate the usage of these genome data, there is a need to build a comparative genomic database.

Here, we present the *Tetrahymena* Comparative Genomics Database (TCGD), which integrates genomic sequences, transcriptomic data, gene annotations, orthologs and synteny maps of 10 closely related *Tetrahymena* species. A user-friendly web interface has been implemented for data searching and visualization. We believe that this is an important online resource for the ciliate community.

## Materials and methods

### Data collection

The TCGD contains the following five main types of data: (i) sequence data, which include genome assemblies, predicted coding sequences (CDS) and protein sequences.

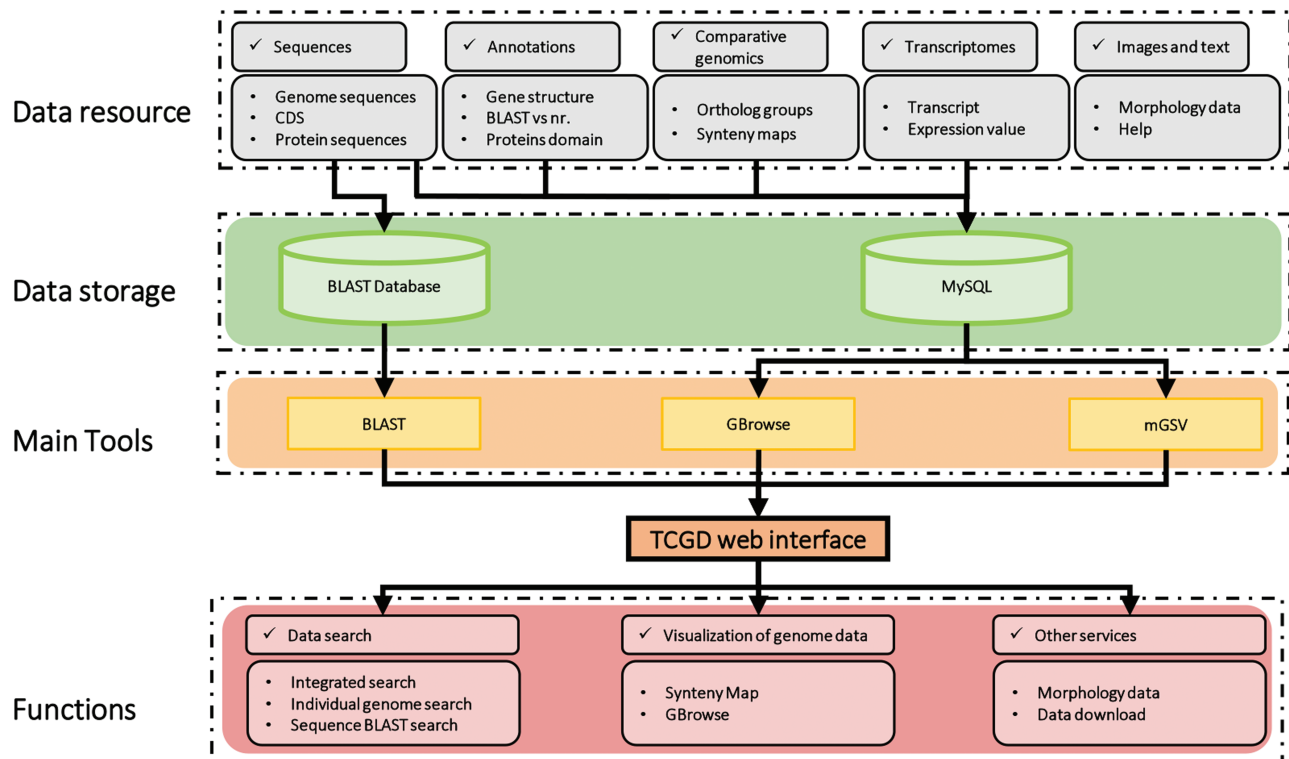
**Table 1.** Basic statistics of MAC genomes of 10 *Tetrahymena* species

Species	MAC genome size (Mb)	Gene number
<i>T. thermophila</i>	103	26 996
<i>T. malaccensis</i>	106.7	24 866
<i>T. ellioti</i>	90.8	22 925
<i>T. pyriformis</i>	116.1	26 866
<i>T. vorax</i>	114.8	25 238
<i>T. borealis</i>	93.5	20 694
<i>T. canadensis</i>	103.4	25 188
<i>T. empidokyrea</i>	84.9	20 847
<i>T. shanghaiensis</i>	95.6	21 982
<i>T. paravorax</i>	108.4	25 551

Four MAC genomes including *T. borealis*, *T. ellioti*, *T. malaccensis* and *T. thermophila* were sequenced by the Broad Institute and can be accessed through the TGD Wiki (<http://ciliate.org/index.php/home/downloads>; 19). The other six genomes, sequenced by the Institute of Hydrobiology, Chinese Academy of Sciences, can be accessed through this database. The MAC genome sizes of 10 *Tetrahymena* species range from 84.9 Mb (*T. empidokyrea*) to 116.1 Mb (*T. pyriformis*; Table 1). In general, more than 200 000 protein-coding genes were predicted in all these 10 species (Table 1); (ii) gene annotations, which include homology-based gene annotations (e.g. best BLAST hit), protein domains, gene ontology (GO) and KEGG ortholog information. Protein domains were annotated using InterProScan (Version 5.19-58.0), which integrates information from ProDom, PANTHER, PROSITE, Pfam and SMART (20). In addition, *in silico* annotation on coiled-coil, signal peptides and transmembrane helices were also included in the database; (iii) synteny maps, which were detected using the MCScanX toolkit (21), based on ortholog groups generated by OrthoMCL (Version 2.0.9); (iv) transcriptomic data, which include the RNA-seq data at growth and starvation stages for each species; and (v) morphological data, including silver staining and scanning electron microscopy images.

### TCGD implementation

The schematic structure of TCGD is shown in Figure 1. The TCGD was built on the Linux operating system with Apache web server. All data are stored in a MySQL relational database management system. The TCGD web interface was developed with JavaScript/HTML to integrate all data resources for user-friendly searching and visualization. GBrowse software was used for visualization of genomic sequences, gene model and transcriptomic data (22).



**Figure 1.** Schematic structure of the TCGD. A flow diagram shows the database architecture. Genome sequences, CDS and protein sequences were formatted as a BLAST database. Sequences, annotation information, comparative genomics data and transcriptomic data were stored in the MySQL database. GBrowse and mGSV were used for visualization of genome data and synteny map. Search and visualization allowed user to easily access the data resources in TCGD.

Multi-Genome Synteny Viewer (mGSV Version 2.1; 23), a web-based tool, was adapted to display the genomic features and their relative order in the genomes of the 10 *Tetrahymena* species. In addition, a standard NCBI BLAST server was set up to enable users to search for similar sequences and retrieve homologous genome components or regions in TCGD.

## Using the TCGD

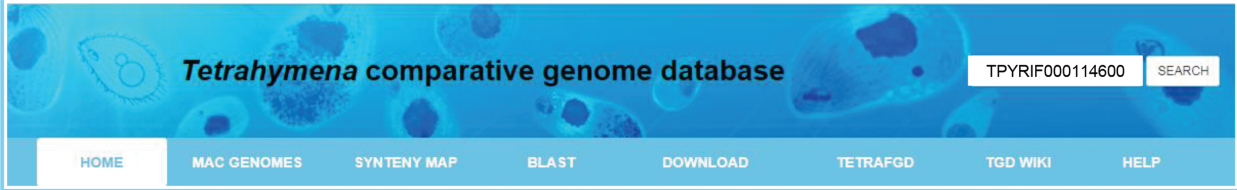
The TCGD has a user-friendly web interface. Two main functions, including the search and visualization, have been implemented to facilitate data accessing.

### Data searching

Multiple search functions have been implemented to enable researchers to obtain useful information (Figure 1). Firstly, an integrated search box (located in the top right corner of the home page) enables whole database searching through the categories ‘Gene ID,’ ‘Scaffold ID,’ or ‘key words’ (Figure 2A). Gene ID is recommended. The general naming rule for genes in nine species (except for *T. thermophila*) is a prefix indicating the species plus a suffix consisting of eight-digit numbers. The prefixes are

TBOREA for *T. borealis*, TELLIO for *T. elliotti*, TEMPID for *T. empidokyrea*, TMALAC for *T. malaccensis*, TPARAV for *T. paravorax*, TPYRIF for *T. pyriformis*, TSHANG for *T. shanghaiensis*, TSP for *T. canadensis* and TVORAX for *T. vorax*. The suffixes indicate the order of genes in the assembled scaffolds. After searching with a specified Gene ID, the hit will be displayed (Figure 2B), with a hyperlink (Figure 2C) to a page containing detailed information of the gene. The following five types of information are presented on this page (Figure 3): (i) a brief description of the gene, including the species, putative annotation based on NCBI BLAST, its location, the assembled scaffold and a hyperlink with coordinates to the synteny map (Figure 3A); (ii) a snapshot with hyperlink shows the gene structure and could further forward the user to GBrowse (Figure 3B); (iii) protein domain, GO and KEGG functional annotations (Figure 3C); (iv) homolog information for all 10 species based on OrthoMCL ortholog groups (Figure 3D); and (v) the predicted CDS and protein sequences (Figure 3E). In addition, a ‘MAC Genomes’ tab navigation bar in the home page directs the user to the MAC genome database for each species (Figure 1). The TCGD also allows BLAST search of sequence against either individual genomes or all 10 genomes.

**A**



**B**

Search result of All

TPYRIF00114600

⋮

*T. pyriformis*

Assembly	0
Gene	1
Annotation	0
Publication	0
⋮	⋮

**C**

*T. pyriformis*

type: gene query: TPYRIF00114600 start: end: [DOWNLOAD](#)

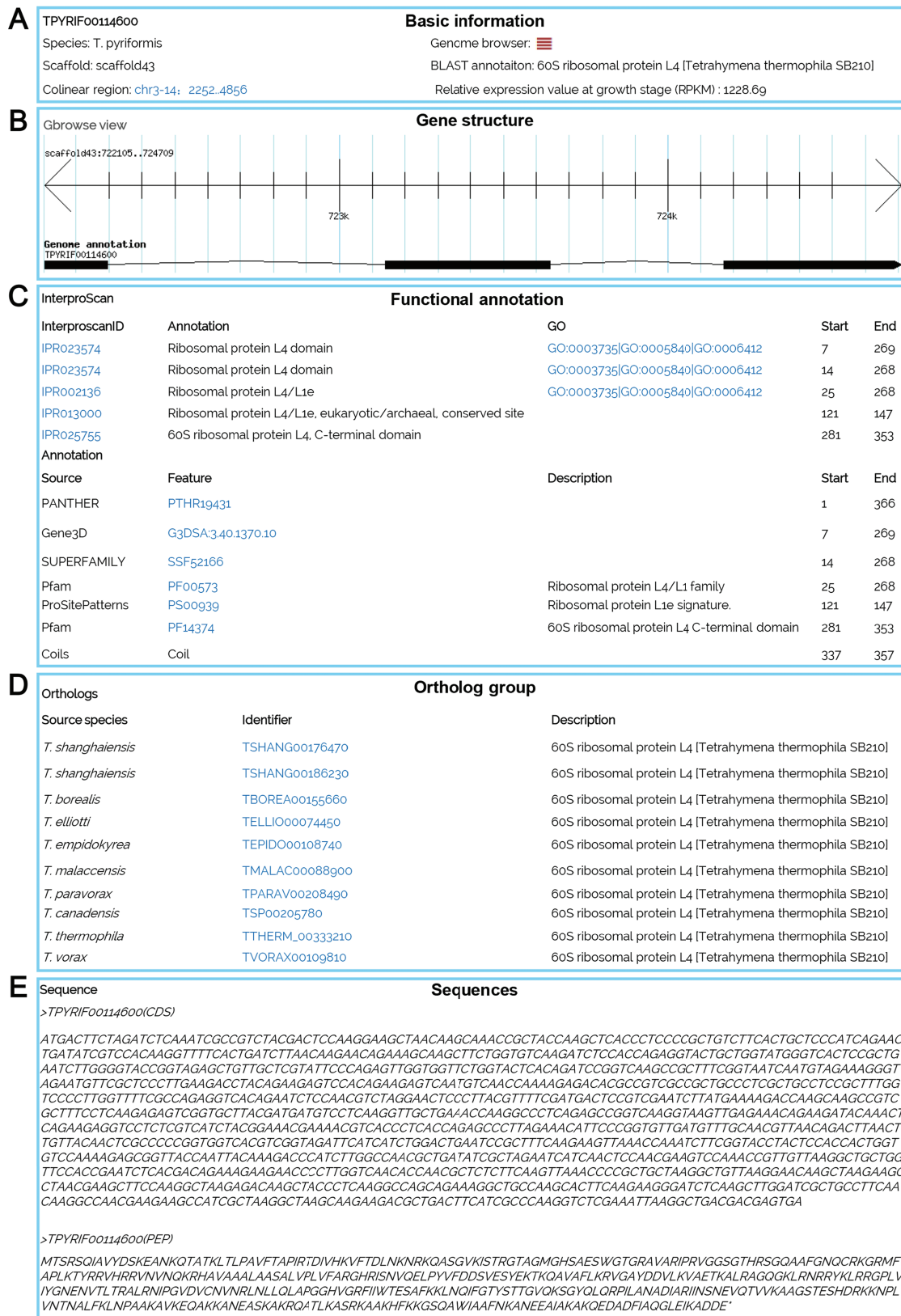
Species	Blast	Gene ID	Alias Name	Scaffold	Start	End
<i>T. pyriformis</i>	60S ribosomal protein L4 [Tetrahymena thermophila SB210]	TPYRIF00114600		scaffold43	722105	724709

**Figure 2.** Search functions implemented into TCGD. (A) An integrated search box. (B) Screenshot of search result interface for gene TPYRIF00114600 through the integrated search box. (C) A brief gene description of TPYRIF00114600, including the species the gene belongs to, putative annotation based on NCBI BLAST hits and the gene location.

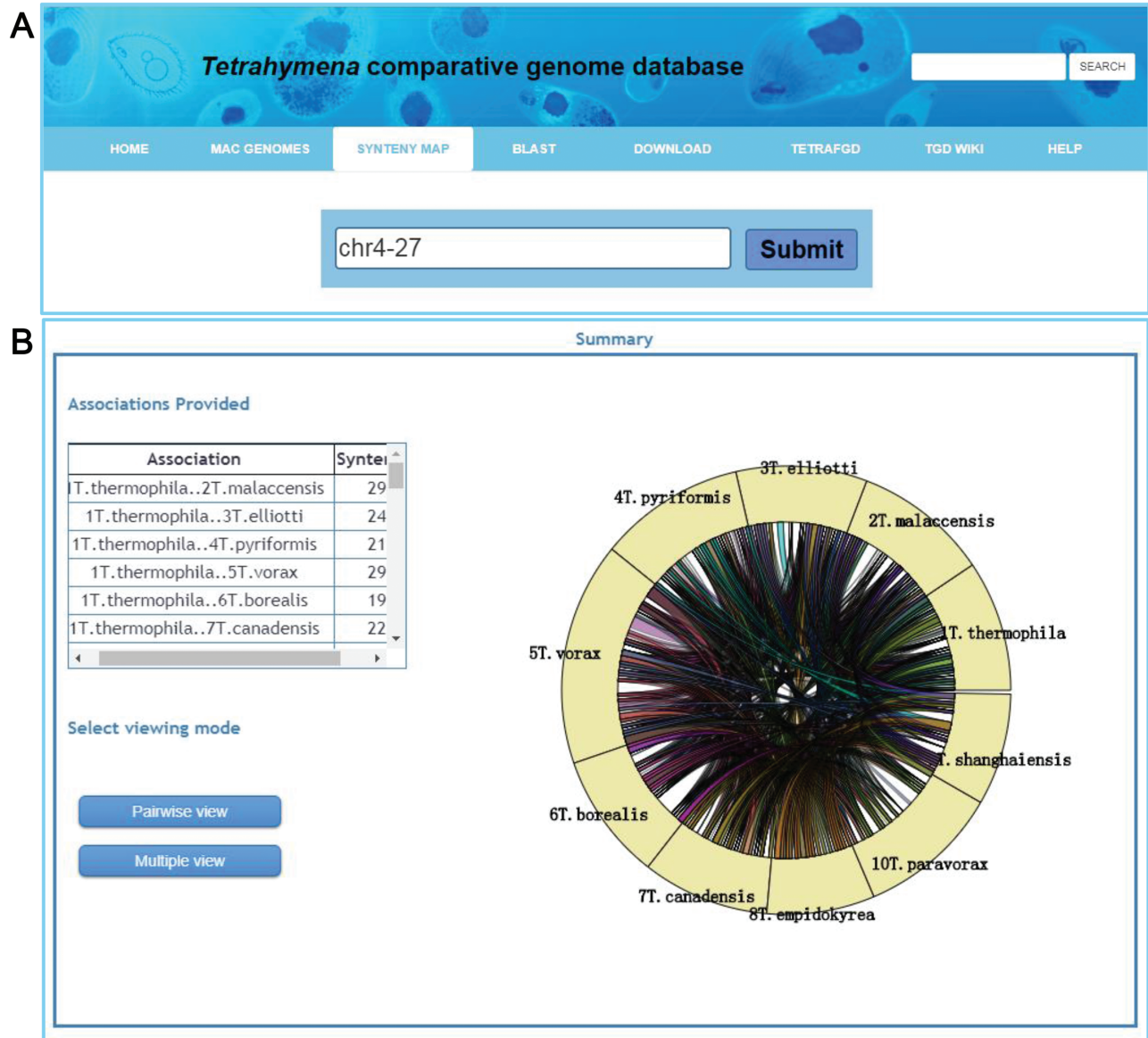
## Visualization of genomic data

An important function for comparative genomics analysis in TCGD is visualization of the synteny maps between genomes of all 10 species. In TCGD, mGSV tool was implemented for browsing the synteny maps. For this, we divided genome-wide collinear relationships into blocks that were generated based on MAC chromosomes of the model species, *T. thermophila*. In *T. thermophila*, the newest assembly suggested that the MAC has 181 chromosomes (including the rDNA minichromosome). We first aligned all assembled scaffolds in other species onto the 181 MAC chromosomes of *T. thermophila* using LASTZ, and then ordered the assembled scaffolds according to the alignment location in *T. thermophila* MAC chromosomes. In some cases, only one homologous scaffold in other species has been found for two *T. thermophila* MAC chromosomes, meaning that a single chromosome in another species will be represented by two in *T. thermophila*. In these cases, we

artificially merged the two *T. thermophila* chromosomes to form a block. A total of 173 blocks (in general, each one representing a MAC chromosome) were obtained and synteny maps were generated for incorporating them into mGSV. All 173 collinear blocks were named according to their order on the MIC chromosomes of the *T. thermophila* MIC. The principle for searching the collinear block is that a ‘chrX-Y’ style ID should be provided, in which ‘X’ represents the MIC chromosome number (five MIC chromosomes, numbered 1–5) and ‘Y’ represents the order on the MIC chromosome. A list of genes in each block was also provided. Besides the block ID, a Gene ID can also be used for searching the synteny map (Figure 4A). After selecting a block or searching by Gene ID, a circular layout containing all conserved genes in the block among 10 species is shown (Figure 4B), with two visualization modes available to browse the synteny maps ‘Pairwise view’ and ‘Multiple view’ (Figure 4B).



**Figure 3.** A gene details page for TPYRIF00114600, showing five types of data. (A) Basic information on the gene, such as the species, putative annotation based on NCBI BLAST hits, and the gene location. (B) A snapshot of the gene structure with a hyperlink to GBrowse. (C) Annotation with InterProScan for protein domains, GO and KEGG function. (D) Homolog information for all 10 species based on OrthoMCL ortholog groups. (E) The predicted CDS and protein sequences.



**Figure 4.** Visualization of a synteny map in TCGD. (A) The ‘collinear block ID’ or ‘Gene ID’ is inserted into the search box to acquire a synteny map for 10 *Tetrahymena* species. (B) A summary page shows genome associations and the number of genes for each genome pair for collinear block ID chr4-27. A circular diagram shows a general overview of the associations. To obtain the full synteny display, users can choose to enter either ‘Pairwise view’ or ‘Multiple view’ mode.

The ‘Pairwise view’ mode shows the synteny maps between adjacent genomes (Figure 5A). Multiple pull-down menus located at the top of the synteny browser enable users to choose specific genomes to display in any order they wish. These pull-down menus can be added or removed using the ‘Insert’ or ‘Delete’ option so that each genome can be shown more than once or removed if necessary. Buttons at the top left corner allow users to change the view for all genomes displayed by zooming in/out, moving left/right. The display for each genome consists of two parts: the control panel on the left and the synteny display on the right. The left control panel allows users to zoom in/out,

move left/right or select specific regions for display. Users can also filter the conserved genes based on the length of conserved regions, as recorded in the synteny files. In the synteny display window, each genome is represented as a horizontal ruler with tick marks showing its genome position and with genes displayed as colored arrowed blocks. The shapes and colors of the gene display can be easily changed via the control panel. When the mouse pointer moves to a gene, the gene name appears. A detailed description of each gene is obtained by clicking on the gene. Each conserved gene is colored differently in the default setting, but users can change the displayed regions



**Figure 5.** Visualization of synteny map in 'Pairwise view' mode and 'Multiple view' mode. (A) In 'Pairwise view' mode, genes are shown as lines between adjacent genomes. Genomes can be rearranged, removed or shown more than once. Genome control panels on the left side of the interface allow the genome viewing range to be adjusted. Master controls at the top apply to all genomes. By using the control panel on the left, users can choose the shape and color of genes. Regions of visible synteny can be filtered based on the numerical criteria specified for genes. (B) In 'Multiple view' mode, conserved genes across all selected genomes are shown. The regions associated with one or more specific genome pairs can be hidden using the buttons above the synteny display. Genomes can also be rearranged or removed, and each genome is displayed only once.

to a uniform color via the ‘Colors’ option. We have also provided the coordinates of each in the gene detail page; user can easily find the exact location of a gene in the synteny map using these coordinates.

In ‘Multiple view’ mode, all synteny maps are shown for each pair of genomes in all selected species (Figure 5B). Users can switch on/off the display of conserved genes between any pair of genomes. As in ‘Pairwise view’ mode, any genome can be added or removed from the display. A single filter panel at the top of the synteny view can be used to filter the conserved genes. By default, all genomes are shown in the order in which they are specified in the synteny files. In both viewing mode, an ‘Optimize order’ option is provided to rearrange the order based on an algorithm developed in mGSV.

In addition, GBrowse was built into the TCGD to visualize genomic data for individual species. Users can click the ‘MAC Genomes’ tab to select a species and then visualize the genome sequence through selecting the navigation tab ‘GENOME BROWSER’ on the homepage of each species. Three tracks [putative gene model, RNA-seq coverage plot and expressing value (RPKM)] are shown by default.

### Other services in the TCGD

The TCGD also provides morphological data in the form of silver staining and scanning electron microscopy images, which allow users to familiarize themselves with the morphology of these species.

## Discussion

*T. thermophila* is a well-studied model system in ciliates. After sequencing of the *T. thermophila* MAC genome in 2006, two important *Tetrahymena* databases have been established: the TGD Wiki, which contains genomic information; and TetraFGD, a functional genomics database that enables researchers to access gene expression, gene network and proteomics data. Sequencing of the MAC genomes of other nine species provided an opportunity to build the TCGD (similar to the *Drosophila* and *Plasmodium* databases). The TCGD aims to provide basic genomic information and comparative genomics analysis of 10 *Tetrahymena* species. Users can obtain comprehensive information on genes of interest through the TCGD. We believe that the TCGD represents an important database for the *Tetrahymena* and ciliate research community.

*Tetrahymena* have both a MIC and MAC in a single cell. Although the MAC genomes of multiple species have been sequenced, so far the MIC genome of only *T. thermophila* has been sequenced using the Illumina platform. As DNA sequencing technology develops, PacBio

and Nanopore sequencing platforms could be used to obtain more complete MIC genome sequences of other *Tetrahymena* species. Future efforts of the TCGD will focus on incorporating the MIC genomes of more *Tetrahymena* species and on facilitating MIC and MAC comparative genomics analysis between different species.

## Availability of supporting data

The TCGD is freely accessible as a web application at <http://ciliate.ihb.ac.cn>. All data, including the genome sequences, CDS, protein sequences, functional annotation files and ortholog groups, are available for download, e.g. <http://ciliate.ihb.ac.cn/tcgd/database/download/#th>.

## Author contributions

W.M. and J.X. conceived the project. W.T.Y., Y.Z., K.C., G.Y.W. and D.X.Y. performed the data analysis and constructed the database. C.Q.J. collected the morphological pictures. W.M., J.X. and W.T.Y. wrote the paper.

## Acknowledgements

We would like to thank Prof. Quanfeng Dong (University of North Texas) for suggestions on installing the mGSV in the server. The research was supported by the Wuhan Branch, Supercomputing Centre, Chinese Academy of Sciences, China.

## Funding

Natural Science Foundation of China (31525021, 91631303 to W.M.) and (31672281 to J.X.); Youth Innovation Promotion Association, Chinese Academy of Sciences (to J.X.).

*Conflict of interest.* None declared.

## References

1. Adl,S.M., Leander,B.S., Simpson,A.G. *et al.* (2007) Diversity, nomenclature, and taxonomy of protists. *Syst. Biol.*, **56**, 684–689.
2. Lynn,D. (2008) *The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature (3rd ed)*. Springer Science & Business Media, Dordrecht.
3. Herrick,G. (1994) Germline-soma relationships in ciliated protozoa: the inception and evolution of nuclear dimorphism in one-celled animals. *Semin. Dev. Biol.*, **5**, 3–12.
4. Eisen,J.A., Coyne,R.S., Wu,M. *et al.* (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, **4**, e286.
5. Aury,J.-M., Jaillon,O., Duret,L. *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171.



6. Wang,R., Xiong,J., Wang,W. *et al.* (2016) High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. *Sci. Rep.*, **6**, 21139.
7. Aeschlimann,S.H., Jonsson,F., Postberg,J. *et al.* (2014) The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome. *Genome Biol. Evol.*, **6**, 1707–1723.
8. Swart,E.C., Bracht,J.R., Magrini,V. *et al.* (2013) The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol.*, **11**, e1001473.
9. Slabodnick,M.M., Ruby,J.G., Reiff,S.B. *et al.* (2017) The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. *Curr. Biol.*, **27**, 569–575.
10. Arnaiz,O. and Sperling,L. (2011) *ParameciumDB* in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. *Nucleic Acids Res.*, **39**, D632–D636.
11. Stover,N.A., Punia,R.S., Bowen,M.S. *et al.* (2012) *Tetrahymena Genome Database Wiki*: a community-maintained model organism database. *Database*, 2012, bas007.
12. Wang,R.L., Miao,W., Wang,W. *et al.* (2018) *EOGD*: the *Euplotes octocarinatus* genome database. *BMC Genomics*, **19**, 63.
13. Gilbert,D.G. (2007) *DroSpeGe*: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res.*, **35**, D480–D485.
14. Aurrecochea,C., Brestelli,J., Brunk,B.P. *et al.* (2009) *PlasmoDB*: a functional genomic database for malaria parasites. *Nucleic Acids Res.*, **37**, D539–D543.
15. Lynn,D.H. and Doerder,F.P. (2012) The life and times of *Tetrahymena*. *Methods Cell Biol.*, **109**, 9–27.
16. Nanney,D.L. and Simon,E.M. (2000) Laboratory and evolutionary history of *Tetrahymena thermophila*. *Methods Cell Biol.*, **62**, 3–25.
17. Sauvant,M.P., Pepin,D. and Piccinni,E. (1999) *Tetrahymena pyriformis*: A tool for toxicological studies. A review. *Chemosphere*, **38**, 1631–1669.
18. Xiong,J., Lu,Y., Feng,J. *et al.* (2013) *Tetrahymena functional genomics database (TetraFGD)*: an integrated resource for *Tetrahymena* functional genomics. *Database*, 2013, bat008.
19. Hamilton,E.P., Kapusta,A., Huvos,P.E. *et al.* (2016) Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *Elife*, **5**, e19090.
20. Mitchell,A., Chang,H.Y., Daugherty,L. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
21. Wang,Y., Tang,H., Debarry,J.D. *et al.* (2012) *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
22. Stein,L.D., Mungall,C., Shu,S. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
23. Revanna,K.V., Munro,D., Gao,A. *et al.* (2012) A web-based multi-genome synteny viewer for customized data. *BMC Bioinformatics*, **13**, 190.