Published online 16 November 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D599–D605***
doi: 10.1093/nar/gkz948

ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology

Olivier Arnaiz[1,*], Eric Meyer[2] and Linda Sperling[1,*]

[1]I2BC, Institute of Integrative Biology of the Cell, UMR9198, CNRS, CEA, Univ Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France and [2]IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005 Paris, France

Received September 12, 2019; Revised October 03, 2019; Editorial Decision October 04, 2019; Accepted October 09, 2019

ABSTRACT

ParameciumDB (https://paramecium.i2bc.paris-saclay.fr) is a community model organism database for the genome and genetics of the ciliate *Paramecium*. ParameciumDB development relies on the GMOD (www.gmod.org) toolkit. The ParameciumDB web site has been publicly available since 2006 when the *P. tetraurelia* somatic genome sequence was released, revealing that a series of whole genome duplications punctuated the evolutionary history of the species. The genome is linked to available genetic data and stocks. ParameciumDB has undergone major changes in its content and website since the last update published in 2011. Genomes from multiple *Paramecium* species, especially from the *P. aurelia* complex, are now included in ParameciumDB. A new modern web interface accompanies this transition to a database for the whole *Paramecium* genus. Gene pages have been enriched with orthology relationships, among the *Paramecium* species and with a panel of model organisms across the eukaryotic tree. This update also presents expert curation of *Paramecium* mitochondrial genomes.

INTRODUCTION

ParameciumDB was launched in 2006 as a community model organism database, to accompany publication of the *Paramecium tetraurelia* somatic genome sequence that revealed a series of whole genome duplications had occurred in the lineage (1). Built using the Generic Model Organism Database toolkit (GMOD, http://gmod.org/), ParameciumDB linked the genome to genetic data (2). Since then, an increasing number of genomic datasets and new tools such as BioMart (3) have been integrated (4).

The primary mission of ParameciumDB was to support *Paramecium* research for the community attracted to this unicellular model characterized by good classic and reverse genetic approaches and large cell size (∼120 μM). These properties facilitate study of the complex biological processes, conserved in multicellular organisms, found in these ciliates. For example: polarized organization of the cell cortex featuring some 2000 motile cilia (5); programmed genome rearrangements at every sexual generation (6); transgenerational epigenetic control of the rearrangements involving small RNA pathways and chromatin modifications (7,8). A secondary mission has been to provide documentation, stocks and standard protocols for *Paramecium* husbandry to a broader community including students and educators.

ParameciumDB has undergone major changes since the last update (4). ParameciumDB now integrates genomic sequences for multiple species of *Paramecium*. Multiple genomes not only offer an evolutionary perspective for functional studies of *Paramecium* biology, but also serve the community interested in the mechanisms and evolutionary consequences of polyploidization events, which have occurred repeatedly in the genus and led to the emergence of new species (1,9–11). Along with the increasing amount of genomic data, a new modern web interface has been developed to provide a better user experience and facilitate standard workflows by making it easy to store, align, blast or export sets of genes or proteins. Enriched gene pages provide information about orthologs within and outside the genus. The platform of tools has been expanded and updated to help biologists browse, search and retrieve data.

MULTIPLE GENOMES

Available data

Paramecia, like other ciliates, harbour structurally and functionally distinct nuclei in their unique cytoplasm. A diploid germline micronucleus (MIC) undergoes meiosis and transmits the genetic information to the next sexual generation. A polyploid (800n in *P. tetraurelia*) somatic macronucleus (MAC), streamlined for gene expression through programmed DNA elimination, develops at each sexual generation and determines the phenotype. At

*To whom correspondence should be addressed. Email: linda.sperling@i2bc.paris-saclay.fr
Correspondence may also be addressed to Olivier Arnaiz. Email: olivier.arnaiz@i2bc.paris-saclay.fr

Table 1. Available genomes. For each genome, the species, strain, cellular compartment (macronucleus, MAC, micronucleus, MIC or mitochondrion, MITO), assembly complexity, number of annotated coding genes and bibliographic reference are provided

Species	Strain	Compartment	Complexity	N50	# coding genes	Reference
Paramecium biaurelia	V1-4	MAC	76976592	145535	40261	(10)
	V1-4	MITO	39731	39731	46	(22)
Paramecium bursaria	110224	MAC	29155737	96293	17226	(12)
Paramecium caudatum	43c3d	MAC	30525943	306679	18673	(9)
	43c3d	MITO	43620	43620	46	(22)
	C026	MITO	44414	44413	46	(22,24)
	C083	MITO	44180	44175	46	(22,24)
	C104	MITO	44421	44420	46	(22,24)
	GBE	MITO	43663	43657	46	(22,24)
Paramecium decaurelia	223	MAC	71912400	189418	40810	(11)
	223	MITO	40127	40127	46	(22)
Paramecium dodecaurelia	274	MAC	71627707	176048	41085	(11)
	274	MITO	40070	40070	40070	(22)
Paramecium jenningsi	M	MAC	65347890	212635	37098	(11)
	M	MITO	40138	40138	46	(22)
Paramecium multimicronucleatum	MO3c4	MAC	35730210	436665	17834	(11)
	MO4	MITO	38064	38062	42	(22)
	Peniche 3I	MITO	39192	39192	42	(22)
Paramecium novaurelia	TE	MAC	64789109	78722	35534	(11)
	TE	MITO	39671	39671	46	(22)
Paramecium octaurelia	K8	MAC	72980862	439553	38668	(11)
	138	MITO	39802	39802	46	
Paramecium pentaurelia	87	MITO	39865	39865	46	
Paramecium primaurelia	Ir4-2	MAC	71017591	460739	34474	(11)
	AZ9-3	MITO	39763	39763	46	
Paramecium quadecaurelia	N1A	MAC	59122419	223945	33793	(11)
	N1A	MITO	39781	39781	46	(22)
Paramecium sexaurelia	AZ8-4	MAC	68020722	420472	36094	(10)
	AZ8-4	MITO	39745	39745	46	(22)
	128	MITO	39939	39939	46	(22,24)
	130	MITO	39946	39946	46	(22,24)
	133	MITO	39984	39983	46	(22,24)
Paramecium sonneborni	30995	MITO	40274	40274	46	
Paramecium tetraurelia	d4-2	MAC	72094543	410619	39642	(1)
	51	MAC	72102941	412881	40460	(13)
	51	MITO	40040	40040	46	(22,24)
	51	MIC	98489268	37181	NA	(14)
	32	MITO	39835	39835	46	
Paramecium tredecaurelia	209	MAC	65931501	490675	36179	(11)
	209	MITO	39834	39834	46	(22)

the time of the most recent ParameciumDB update (4), only the somatic genome of the *P. tetraurelia* model was available (1). The MAC genomes of *P. caudatum*, *P. sexaurelia* and *P. biaurelia* were published in 2014 (9,10) followed by many more genomes for *aurelia* species, the *P. bursaria* and *P. multimicronucleatum* genomes (11,12). Table 1 provides a list of all the genomes (MAC, MIC and mitochondrial) available in ParameciumDB in 2019. For the MAC genomes, the gene annotations were obtained using a pipeline based on EuGene described in (13). For some species (*P. caudatum*, *P. mulimicronucleatum*, *P. biaurelia*, *P. tetraurelia* and *P. sexaurelia*), earlier annotations are also provided. RNA-Seq data used for annotation is available for most of the species and can be visualized in the ParameciumDB genome browser.

The first germline sequences are becoming available for *P. tetraurelia* (14) (Table 1). One class of germline-limited sequence elements, the Internal Eliminated Sequences (IESs), have been annotated genome-wide (15). The ∼ 45 000 IESs of *P. tetraurelia*, unique non-coding sequences that can interrupt both coding and non-coding regions of the genome and are precisely excised by a domesticated transposase (16,17) are integrated in ParameciumDB. All IESs can be visualized in the genome browser which links to IES pages with additional information. Intragenic IESs also appear on Gene pages. In the future, germline genomes from more species, their IESs and other germline-limited elements such as transposable elements will be included in ParameciumDB.

An exponentially growing number of genomic datasets from mechanistic studies of *P. tetraurelia* biological processes are integrated into ParameciumDB. In addition to RNA-Seq developmental time-course data (13), a majority of the functional datasets currently available consist of DNA-Seq data generated by re-sequencing the somatic genome that developed after depletion of factors involved in the genome rearrangements. This can lead to retention of some or all IESs at the next sexual generation. The retention scores (18) for each IES after depletion of a factor can be retrieved from ParameciumDB, on IES pages or using BioMart. The DNA-Seq datasets are mapped to the genome and can be visualized with the genome browser.

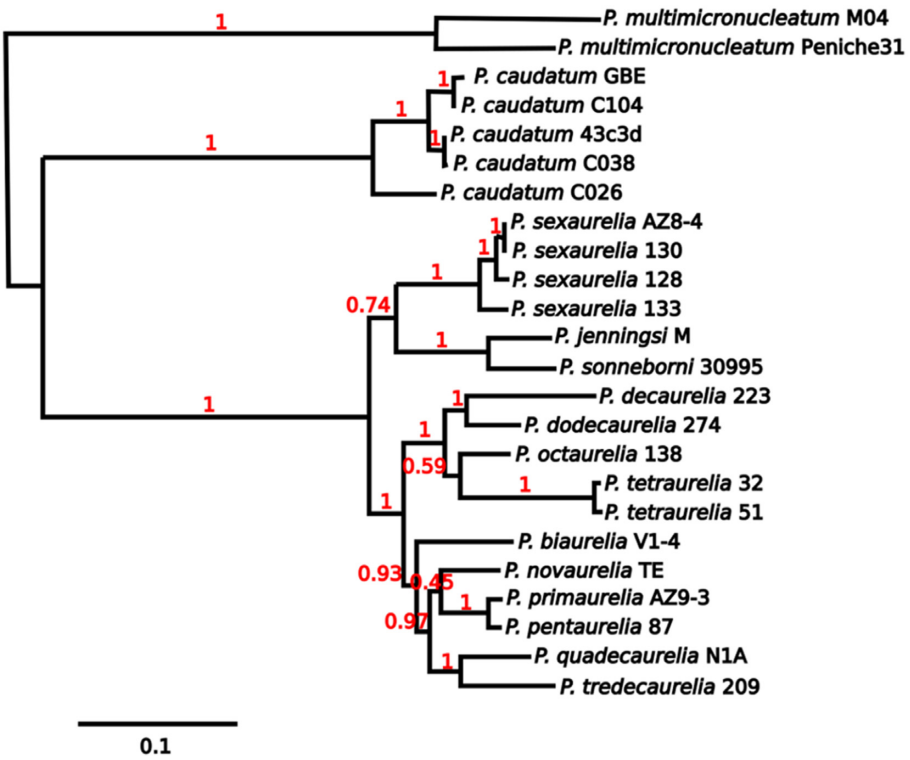

Figure 1. Phylogenetic tree of *Paramecium* species based on the alignment of all 46 mitochondrial proteins. The tree was built by MAFFT alignment of concatenated protein sequences, where all N- and C-termini are correctly aligned; strict Gblocks curation selects 9135 residues aligned in all species. PhyML aLRT analysis was carried out using the service at https://www.phylogeny.fr with default parameters. This analysis places the *caudatum* species as closest relatives of the *aurelia* species. Note the good support for a *sexaurelia-sonneborni-jenningsi* clade.

All genomes, annotations and mapping (Bam files (19)) are available from Downloads.

Curation of mitochondrial genomes

In *Paramecium*, mitochondria are inherited maternally as very little cytoplasm is normally exchanged between mates during conjugation (20). This is consistent with the lack of evidence for mitochondrial recombination first reported by (21) and now confirmed at the molecular level (22). Pioneering work by Cummings and colleagues showed that the *Paramecium* mitochondrial genome is a linear molecule with a telomere at one end and a single-stranded loop at the other (23) and provided the first complete mitochondrial genome sequence, of *P. tetraurelia* (24).

Johri *et al.* (22) recently extended our knowledge of *Paramecium* mitochondrial genomes to nine *P. aurelia* species, *P. caudatum* and *P. multimicronucleatum*. We have added a few more *P. aurelia* species (cf. Table 1). These genome sequences were curated, first by polishing them using available Illumina sequencing data (the changes are provided in Supplementary Table S1) then by manual annotation of the coding and non-coding genes as detailed in the legend to Supplementary Figure S1, which schematizes the structure of *P. multimicronucleatum*, *P. caudatum* and *P. tetraurelia* mitochondrial genomes. The main problem found in the annotation of (22) concerns the 5′ annotation of protein-coding genes. Indeed, UUG was not considered as possible initiation codon as it is not listed for ciliates in the NCBI codon usage table (code 4). Use of UUG as initiation codon allowed highly coherent annotation across all species of complete ORFs. Figure 1 shows a phylogenetic tree obtained after concatenation of the mitochondrial proteins. No evidence was found for the multiple tRNAs described in (22), only the three (Tyr, Phe, Trp) tRNAs previously identified in *P. tetraurelia* were found in all species. Our annotation also considerably reduces overlaps between adjacent protein-coding genes. Consistent with the previously established structure with a single-stranded loop at one end and a telomere at the other, we only found telomeres at one end of the molecules. The curated genomes can be viewed in ParameciumDB JBrowse and the genome, gene and protein sequences are available from ParameciumDB Downloads.

MODERN INTERFACE AND TOOLS

Technology components behind ParameciumDB

Table 2 presents the tools and software components used by ParameciumDB. Like the previous versions of ParameciumDB, the current release takes advantage of tools developed by the GMOD project consortium. All genomic and post-genomic data are loaded into a Chado-PosgreSQL database (25) and also into a Bio::DB::SeqFeature::Store-MySQL database (26) for use by the genome browsers. To build the web interface, Perl-Template-toolkit technology was used for server-side dynamic web content generation. The client-side possibilities provided by Ajax-JavaScript

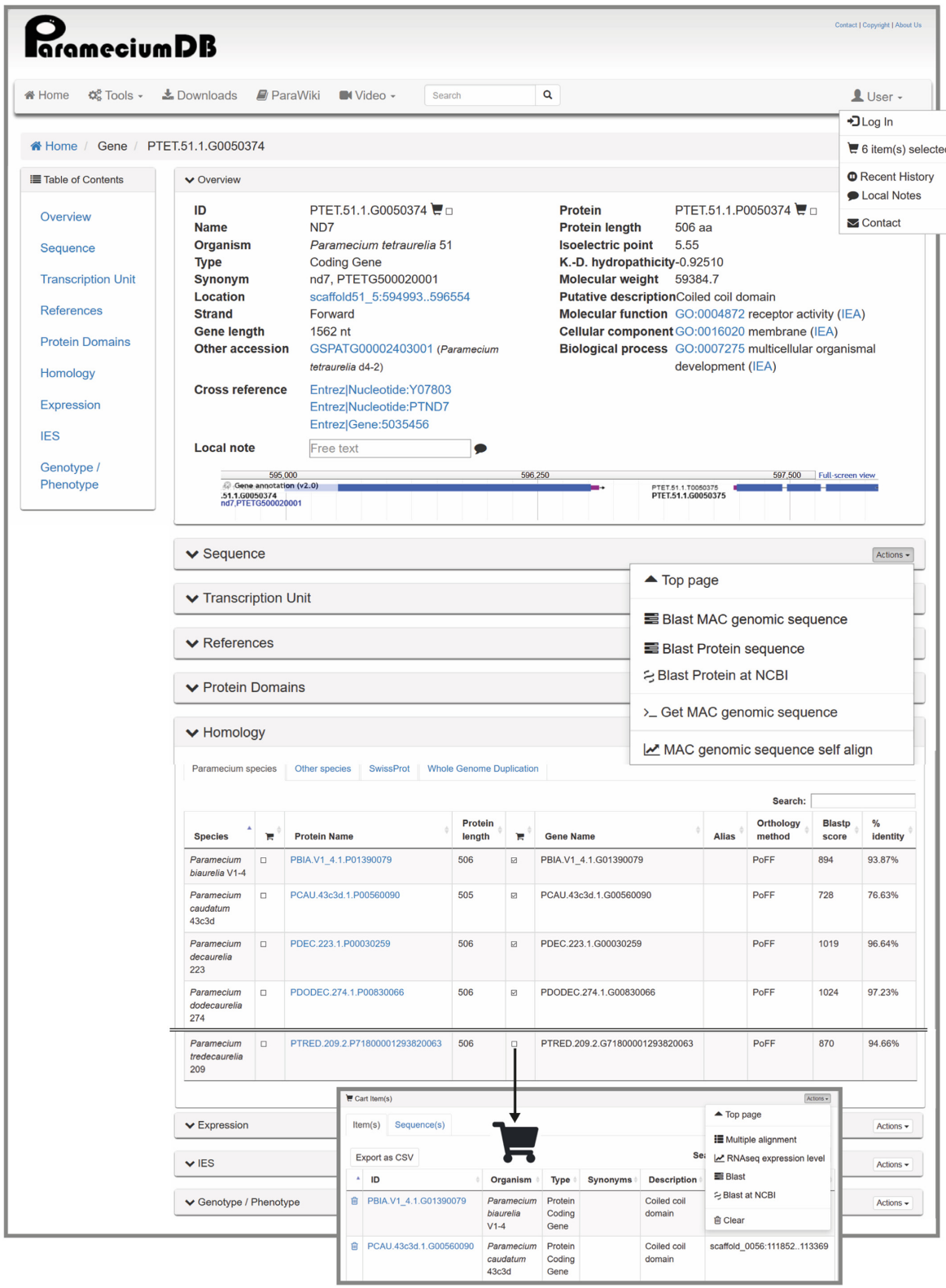

Figure 2. Gene page example. Gene page of *ND7*, showing the different collapsable sections available for this gene, indicated in the Table of Contents: Overview with general description and cross references, Sequence, Transcription unit structure, Bibliographic references, Protein domain predictions, Homology within *Paramecium* species and with other organisms, Expression data (RNA-seq, microarray, proteomic), intragenic IES information and information about *ND7* alleles, stocks and mutant phenotypes. Selecting proteins adds them to the shopping cart as shown for the Homology section. The shopping cart can be visualized from the User dropdown menu (top right of page) as shown by the inset screenshot at the bottom of the figure. Actions are available for the shopping cart items: multiple alignment (with MUSCLE (34)), expression profiles, submit to a blast at ParameciumDB or NCBI. The items can also be exported, in fasta format by choosing the 'Sequence(s)' tab.

Table 2. Tools and software components

	Tool	Software	Description	Reference
Database	chado	V1.31; PostgreSQL	Data storage	(25)
	lucene	2.3.3.4	Quick search indexation	https://lucene.apache.org/
	bio-db-seqfeature-store	MySQL (mariadb v5.5.60), BioPerl (v1)	Data storage	(26)
Interface	Template toolkit	v2.24	Web content system	http://www.template-toolkit.org
	Web design	Jquery (v2.1.4), Bootstrap (v3.3.5), font-awesome (v4.5.0)	Javascript and CSS libraries	https://jquery.com, https://getbootstrap.com, https://fontawesome.com/
Search	Blast	ncbi-blast (v2.3.0+)	Search for a sequence in a nucleotide or protein database, using NCBI Basic Local Alignment Search Tool	(27)
	BioMart	Biomart (v0.9)	Advanced query and data retrieval interface, powered by BioMart software	(3)
	Get Sequence	Samtools (v0.1.19)	Retrieve a single nucleotide or protein sequence from a ParameciumDB database	(19)
	ID converter	In-house	Find corresponding gene IDs between different gene annotation versions	
	Search motif	EMBOSS (v6.6.0.0)	Searches sequences with a sequence motif using patmatdb program from the EMBOSS package	(30)
Browse	JBrowse	Jbrowse (v1.12.1)	A fast, Interactive genome browser	(29)
	GBrowse	Gbrowse2 (v2.54)	The Generic Genome Browser	(28)
	Chromosome Synteny	Circos (v0.69-3)	Visualize gene paralogy or orthology relationships between scaffolds, based on protein similarity	(32)
	Stock Tubes	In-house	Visualize available strains, genotypes and phenotypes	
	Publications	In-house	Paramecium publications, linked to PubMed and to ParameciumDB gene pages	
	Codon Usage	In-house	Codon usage tables and tool to find alternate codons for design of synthetic genes resistant to RNAi	
	ParaWiki	Mediawiki (v1.27.1)	Community pages with information about Paramecium biology, protocols, meetings and more	https://www.mediawiki.org
Analysis	RNAi Off-target	BWA (v0.7.12)	Determine whether a sequence has off-target siRNA matches.	(31)
	Multiple alignment	MUSCLE (v3.8.425)	Construct a multiple alignment of nucleotide or protein sequences with MUSCLE.	(34)

The tools and software components used by ParameciumDB are organized by category (Database, Interface, Search, Browse, Analysis). Each tool has a brief description. For each software component, the version is provided as well as bibliographic reference or URL.

JQuery libraries were fully exploited to improve user experience through fluid and efficient web navigation. In addition, bootstrap and fontawesome CSS libraries were used to achieve a more contemporary style for the web pages.

Since the latest update of ParameciumDB (4), all of the tools have been maintained to assure continuity for the user, with recent versions of the underlying software as given in Table 2. Special effort focused on improvement of some crucial tools. A wrapper for the NCBI-BLAST+ (27) tool was developed to allow users to submit multiple query sequences against multiple indexed databases and improve the output interface to support multiple visualization and download choices. For example, from a BLASTP output, it is possible to choose a species and either recover the sequence of the target protein or be redirected to the corresponding Gene page. The new version of BioMart (v0.9) (3) provides an advanced query interface and data retrieval system. Note that the GBrowse2 (28) and JBrowse (29) coexist in ParameciumDB, although the use of JBrowse is recommended because it supports much faster browsing of data tracks with easier navigation. It is possible to search for a protein motif using an EMBOSS/patmatdb wrapper (30). The Codon Usage tool can translate a coding sequence or retro-translate a protein sequence into a coding sequence, using either optimal or most frequent codons. One popular use of this tool is to help design synthetic genes resistant to RNA interference. The RNAi off-target tool uses

the BWA (31) mapper to check a sequence for possible off-target matches.

Gene pages, the heart of the web interface

All pages of ParameciumDB have a header with hyperlinks leading to the main sections of the website (cf. Figure 2). Beyond the classic link to the Home page, the Tools link reveals a dropdown menu with the available tools (see Table 2). The Download link is used to retrieve all the genomic data and annotations contained in ParameciumDB, as well as a daily SQL dump of the database. The ParaWiki and the Video links give access to various information about *Paramecium* biology and the use of ParameciumDB. Note that video tutorials are available. The names of genes, accessions and putative functions are indexed and searchable with the quick search field. The User menu gives access to a shopping cart, search history and local notes. All this information is stored in the user's browser memory using Web-Storage technology. The Shopping Cart can store genes or proteins and facilitates their export and analysis. It is possible to become a registered user of ParameciumDB allowing access to unpublished data but none of the features or data presented in this article require a login account.

All gene pages present a Table of Contents facilitating access to the different types of data available for the gene of interest. In Figure 2, the Overview section includes important properties of the gene and its protein: name, synonyms, cross references, gene and protein sizes, species to which it belongs, genomic location and electronically inferred functional annotation. Below, a JBrowse inset shows a graphical representation of the gene. The genomic, coding and protein sequences are accessible in the Sequence section. Thanks to the Action button, the user can easily initiate a homology search using BLAST, at ParameciumDB or at NCBI. The Transcription Unit section can provide additional information about transcript structure (introns, exons, 5′ UTR, 3′ UTR, TSS, polyA_site) (13). Any publications related to the gene are mentioned in the References section. Curation of the literature related to *Paramecium* biology is performed every month using PubMed, linking genes and bibliographic references. The Protein Domains section groups the results of domain predictions on protein sequences (pre-calculated results of HMMscan on the PfamA library, dynamic analysis using the NCBI Conserved Domain Database, prediction of signal peptides or transmembrane helices).

Orthology between proteins of multiple *Paramecium* species is one of the novelties of this version of ParameciumDB. The orthology calculations between paramecium proteins described by (11) are presented in a table in the Homology section. The orthology relationships can be displayed, with circos (32), at the chromosome level using the Chromosome Synteny tool. As in the previous version of ParameciumDB, Inparanoid (33) analysis provides the orthology links between paramecium proteins and a variety of other organisms (*A. thaliana, C. reinhardtii, C. intestinalis, D. rerio, D. discoideum, E. coli, G. lamblia, H. sapiens, M. musculus, S. cerevisiae, S. pombe, C. elegans, D. melanogaster, P. falciparum, T. thermophila, O. trifallax, I. multifiliis, S. Lemnae, S. coeruleus*). In addition, whole genome duplication paralogs are available in the Whole Genome Duplication tab.

Transcriptome studies by microarray or RNA-seq on *Paramecium tetraurelia* are included in the Expression section. Proteomics data can also be recovered from this section. In the IES section, IESs overlapping the gene of interest are listed with available functional data (see Available Data). Mutant alleles and stocks are available for some genes, indicated in the Genotype/Phenotype section.

FUTURE OF PARAMECIUMDB

We hope that the inclusion of more and more *Paramecium* genomes will make ParameciumDB as useful to evolutionary biologists and population geneticists as it has been to cell and molecular biologists over the past decade. Beyond the goal of maintaining the database with a very small staff and little specific funding—which is only possible thanks to many wonderful open source software projects and to large well-funded generalist databases hosted by e.g. SBI, NCBI, EMBL-EBI—we hope to integrate more and more somatic and germline genomes and their annotations. We also plan to interface, integrate or develop better tools to facilitate comparison of genes, proteins and chromosomes. ParameciumDB can also provide added-value through expert human curation, as highlighted here for *Paramecium* mitochondrial genomes.

DATA AVAILABILITY

https://paramecium.i2bc.paris-saclay.fr.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the French National Sequencing Center (Genoscope CEA) and the France Génomique Paramecium Sequencing Project coordinated by Sandra Duharcourt for help with the mitochondrial genomes for *P. sonneborni*, *P. primaurelia* and *P. pentaurelia*. We are grateful to Michael Lynch, Tom Doak and Jean-François Goût for sharing *Paramecium* genomes. We thank France Koll for finding bugs and making numerous suggestions to improve ParameciumDB. The Informatics and Scientific Calculation Service of the I2BC has consistently supported ParameciumDB by providing infrastructure, help with purchase and configuration of servers and advice on deployment. We continue to benefit enormously from the GMOD project software and support.

FUNDING

ParameciumDB is supported by intramural funding from the CNRS; Agence Nationale de la Recherche [ANR-14-CE10-0005-04 'PIGGYPACK', ANR-18-CE12-0005-01 'LaMarque']. Funding for open access charge: Agence Nationale de la Recherche [ANR-14-CE10-0005-04 'PIGGYPACK'].

Conflict of interest statement. None declared.

REFERENCES

1. Aury,J.-M., Jaillon,O., Duret,L., Noel,B., Jubin,C., Porcel,B.M., Ségurens,B., Daubin,V., Anthouard,V., Aiach,N. *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**, 171–178.
2. Arnaiz,O., Cain,S., Cohen,J. and Sperling,L. (2007) ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Res.*, **35**, D439–D444.
3. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
4. 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.
5. Tassin,A.-M., Lemullois,M. and Aubusson-Fleury,A. (2015) Paramecium tetraurelia basal body structure. *Cilia*, **5**, 6.
6. Betermier,M. and Duharcourt,S. (2014) Programmed rearrangement in ciliates: paramecium. *Microbiol. Spectr.*, **2**, doi:10.1128/microbiolspec.MDNA3-0035-2014.
7. Coyne,R.S., Lhuillier-Akakpo,M. and Duharcourt,S. (2012) RNA-guided DNA rearrangements in ciliates: is the best genome defence a good offence? *Biol. Cell*, **104**, 309–325.
8. Frapporti,A., Miró Pina,C., Arnaiz,O., Holoch,D., Kawaguchi,T., Humbert,A., Eleftheriou,E., Lombard,B., Loew,D., Sperling,L. *et al.* (2019) The Polycomb protein Ezl1 mediates H3K9 and H3K27 methylation to repress transposable elements in Paramecium. *Nat. Commun.*, **10**, 2710.
9. McGrath,C.L., Gout,J.-F., Doak,T.G., Yanagi,A. and Lynch,M. (2014) Insights into three whole-genome duplications gleaned from the Paramecium caudatum genome sequence. *Genetics*, **197**, 1417–1428.
10. McGrath,C.L., Gout,J.-F., Johri,P., Doak,T.G. and Lynch,M. (2014) Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res.*, **24**, 1665–1675.
11. Gout,J.-F., Johri,P., Arnaiz,O., Doak,T.G., Bhullar,S., Couloux,A., Guérin,F., Malinsky,S., Sperling,L., Labadie,K. *et al.* (2019) Universal trends of post-duplication evolution revealed by the genomes of 13 *Paramecium*species sharing an ancestral whole-genome duplication. bioRxiv doi: https://doi.org/10.1101/573576, 11 March 2019, preprint: not peerreviewed.
12. He,M., Wang,J., Fan,X., Liu,X., Shi,W., Huang,N., Zhao,F. and Miao,M. (2019) Genetic basis for the establishment of endosymbiosis in Paramecium. *ISME J.*, **13**, 1360–1369.
13. Arnaiz,O., Van Dijk,E., Bétermier,M., Lhuillier-Akakpo,M., de Vanssay,A., Duharcourt,S., Sallet,E., Gouzy,J. and Sperling,L. (2017) Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression. *BMC Genomics*, **18**, 483.
14. Guérin,F., Arnaiz,O., Boggetto,N., Denby Wilkes,C., Meyer,E., Sperling,L. and Duharcourt,S. (2017) Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements. *BMC Genomics*, **18**, 327.
15. Arnaiz,O., Mathy,N., Baudry,C., Malinsky,S., Aury,J.-M., Denby Wilkes,C., Garnier,O., Labadie,K., Lauderdale,B.E., Le Mouël,A. *et al.* (2012) The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLos Genet.*, **8**, e1002984.
16. Baudry,C., Malinsky,S., Restituito,M., Kapusta,A., Rosa,S., Meyer,E. and Bétermier,M. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev.*, **23**, 2478–2483.
17. Dubois,E., Mathy,N., Régnier,V., Bischerour,J., Baudry,C., Trouslard,R. and Bétermier,M. (2017) Multimerization properties of PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements. *Nucleic Acids Res.*, **45**, 3204–3216.
18. Denby Wilkes,C., Arnaiz,O. and Sperling,L. (2016) ParTIES: a toolbox for Paramecium interspersed DNA elimination studies. *Bioinforma. Oxf. Engl.*, **32**, 599–601.
19. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.
20. Sonneborn,T.M. (1974) Paramecium aurelia. In: King,R (ed). *Handbook of Genetics*. Plenum Press, NY, Vol. **11**, pp. 469–594.
21. Adoutte,A., Knowles,J.K. and Sainsard-Chanet,A. (1979) Absence of detectable mitochondrial recombination in Paramecium. *Genetics*, **93**, 797–831.
22. Johri,P., Marinov,G.K., Doak,T.G. and Lynch,M. (2019) Population genetics of paramecium mitochondrial genomes: recombination, mutation spectrum, and efficacy of selection. *Genome Biol. Evol.*, **11**, 1398–1416.
23. Pritchard,A.E. and Cummings,D.J. (1984) Structural and functional analysis of the origin of replication of mitochondrial DNA from Paramecium aurelia: I. Inverted complements form the terminal loop. *Curr. Genet.*, **8**, 477–482.
24. Pritchard,A.E., Seilhamer,J.J., Mahalingam,R., Sable,C.L., Venuti,S.E. and Cummings,D.J. (1990) Nucleotide sequence of the mitochondrial genome of Paramecium. *Nucleic Acids Res.*, **18**, 173–180.
25. Mungall,C.J., Emmert,D.B. and FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinforma. Oxf. Engl.*, **23**, i337–i346.
26. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
27. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
28. Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.*, **14**, 162–171.
29. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
30. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet. TIG*, **16**, 276–277.
31. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.*, **25**, 1754–1760.
32. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
33. Sonnhammer,E.L.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
34. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.