D632–D636 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 14 October 2010*
doi:10.1093/nar/gkq918

ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*

Olivier Arnaiz and Linda Sperling*

Centre de Génétique Moléculaire, CNRS FRE3144, Avenue de la Terrasse, 91198 Gif-sur-Yvette, France; Université Paris-Sud 11, 91405 Orsay, France

Received September 2, 2010; Accepted September 24, 2010

ABSTRACT

ParameciumDB is a community model organism database built with the GMOD toolkit to integrate the genome and biology of the ciliate *Paramecium tetraurelia*. Over the last four years, post-genomic data from proteome and transcriptome studies has been incorporated along with predicted orthologs in 33 species, annotations from the community and publications from the scientific literature. Available tools include BioMart for complex queries, GBrowse2 for genome browsing, the Apollo genome editor for expert curation of gene models, a Blast server, a motif finder, and a wiki for protocols, nomenclature guidelines and other documentation. In-house tools have been developed for ontology browsing and evaluation of off-target RNAi matches. Now ready for next-generation deep sequencing data and the genomes of other *Paramecium* species, this open-access resource is available at http://paramecium.cgm.cnrs-gif.fr.

INTRODUCTION

The pioneering work of Tracy Sonneborn established *Paramecium* as a unicellular model of great interest for studies of cellular organization and non-mendelian heredity, over half a century ago. *Paramecium* and other ciliates are the only unicellular eukaryotes that separate germ line and somatic functions. Germ line micronuclei undergo meiosis and transmit the genetic information to the next sexual generation. A somatic macronucleus contains a rearranged version of the genome streamlined for gene expression and is the seat of all transcriptional activity during vegetative growth.

The somatic genome of the *Paramecium* species most used for genetic studies, *Paramecium tetraurelia*, was sequenced and annotated (1), revealing nearly 40 000 protein-coding genes. This surprisingly large number of genes, especially for a unicellular organism, turned out to be the result of a series of at least three whole genome duplications (WGD). These rare and dramatic events, previously documented in plants, animals and fungi, are resolved over evolutionary time by the loss of one duplicate for the majority of genes (2). In *P. tetraurelia*, the most recent WGD (51% of pre-duplication genes still in two copies) occurred just before the explosion of speciation events that gave rise to the *Paramecium aurelia* complex of 15 sibling species (1,3) while the old WGD (8% of genes still in two copies) probably occurred before the divergence of *Paramecium* and *Tetrahymena* (1,4).

Paramecium remains an outstanding model for studies of cilia and ciliary basal bodies (5,6) and of genome rearrangements and their epigenetic control (7,8). The remarkable ease and efficiency of RNAi by feeding in this organism provides a powerful tool for analysis of gene function (9). In addition, *Paramecium* now constitutes a unique system to study the mechanisms of genome evolution consequent to gene duplication, since a very low rate of large-scale genome rearrangements in the *Paramecium* lineage (10) made it possible to identify an unprecedented large number of WGD paralogs of different ages and because the *P. aurelia* species complex provides the opportunity of testing the role of reciprocal gene loss in speciation.

ParameciumDB was inaugurated at the same time as the publication of the somatic genome sequence (1,11). The mission of this open-access resource is to integrate genomic and post-genomic data, improve genome annotations and provide information about the biology of the reference species *P. tetraurelia*. The database strives to be standards-aware, provide downloadable data sets and database dumps and assure curation involving the research community. Since its inauguration, ParameciumDB has acquired new computational and post-genomic data, new tools for viewing, searching and annotating the data and is now ready to accommodate next-generation sequencing (NGS) and the genomes of other *Paramecium* species.

*To whom correspondence should be addressed. Tel: +33 169823209; Fax: +33 169823150; Email: sperling@cgm.cnrs-gif.fr

NEW DATA

ParameciumDB uses the modular, standards-aware Chado relational database schema (12). All data is typed using controlled vocabularies such as the Sequence Ontology (SO) (13) and microarray data is stored using the MIAME-compliant MAGE module. Phenotypes associated with mutant strains and alleles are represented by the entity–quality (EQ) model and the PATO Quality Ontology (14).

Since the initial release of the database, new post-genomic data sets have been introduced. Two sets of proteome peptides are currently available in ParameciumDB. The first set is from the infraciliary lattice (ICL), a continuous contractile network that constitutes the innermost layer of the *Paramecium* cortical cytoskeleton. The ICL is composed of small, EF-hand Ca^{2+}-binding proteins known as centrins and large Sfi-related centrin-binding proteins that transduce the Ca^{2+} sensed by the centrins into macroscopic cellular contraction (15,16). The second set of proteome peptides are from isolated cilia (6). The gene expression data currently available in ParameciumDB is from genome-wide transcriptome experiments using a NimbleGen custom microarray platform (17). The experiments identified genes differentially expressed during the sexual cycle of autogamy, during reciliation and during recovery from massive exocytosis that stimulates secretory granule biogenesis. A genome-wide study of the *Paramecium* kinome (18) has likewise been incorporated, as has annotation of non-coding RNA (19). Figure 1 shows how post-genomic data is displayed in the genome browser.

Publications concerning *Paramecium* are now mined from PubMed once a month and curated to establish links to genes and to add the published gene nomenclature to ParameciumDB. Orthologs in 33 species predicted using Inparanoid (20) are shown on every gene page and the detailed view provides protein alignments and links to external databases.

Supplementary Table S1 shows the currently available data, the ontology used to type the data and the study that generated the data, if relevant.

NEW TOOLS

The GMODWeb framework used to render ParameciumDB gene and protein pages has been described (11,21). Since the initial release of the database, the data in ParameciumDB is now used to populate a BioMart data warehouse (22) to allow complex queries. It is possible to download customized reports of the query results as tabulated text files, Excel files with active links to ParameciumDB gene pages, or dumps in standard SQL or GFF3 formats. The sequences can be downloaded in fasta format. Since the *Paramecium* BioMart is available through the EBI portal, *Paramecium* data sets can be accessed by Galaxy (23), a web-based data analysis platform.

ParameciumDB now uses the latest release (version 2.03) of GBrowse (24). In addition to providing a much improved user-experience through a number of major improvements, GBrowse2 makes it possible to visualize next-generation deep sequencing data. The Bio::DB::Sam Perl module port of Samtools (25) allows the browser to render binary files with mapped short reads and even zoom down to the sequences. Although no NGS data is currently available (July 2010) through the public version of ParameciumDB, we have used the same software to make a tool that evaluates RNAi off-target matches. The tool is intended to help researchers design RNAi reagents that will target only the gene(s) of interest, which it does by cutting the sequence provided by the user into overlapping user-specified windows (23 nt, the size of *Paramecium* siRNA, is the default value). The short sequences are then mapped to the genome, with a choice of 0–3 mismatches, using the BWA aligner (26). A binary file is created on the fly with Samtools and displayed by Gbrowse2. This tool is also useful for identification of paralogs/repetitions of chosen coding or non-coding regions of the genome. A track in the genome browser, 'RNAi off-target', was pre-calculated using the same approach but gives the inverse image, as it shows how many 23 nt stretches from elsewhere in the genome map to a given genomic position (Figure 1).

Other new tools include a motif finder and an early version of an ontology browser that links Gene Ontology (GO) terms to genes. Table 1 provides an exhaustive list of the tools available in ParameciumDB.

COMMUNITY INVOLVEMENT

Since ParameciumDB does not have any full time curators, we interfaced the Apollo genome editor (27) to ParameciumDB so that community members can query an annotation instance of the ParameciumDB chado database to view regions of the genome with many kinds of evidence and create new genome annotations. Once a month, the new expert annotations tagged 'finished' are incorporated into ParameciumDB, with the name of the curator and comments. The latest version of ParameciumDB proteins, including the new curated annotations, then becomes available for download as a fasta file and can be chosen as database for Blast searches.

Using wiki media, we recently added the 'parawiki' to ParameciumDB for documentation, protocols and nomenclature guidelines (a complete version of the nomenclature guidelines is provided as Supplementary data). The parawiki is also used to provide information and registration for meetings. With permission from Cold Spring Harbor Laboratory Press, the parawiki includes the complete version of short articles about the biological uses of *Paramecium* that constitute, in an abbreviated from, a chapter in the Emerging Model Organism Series (28). We welcome additional contributions from the community.

FUTURE DIRECTIONS

ParameciumDB is ready to map and display NGS data, which is already being generated to study gene expression, identify mutations, polish the sequence of the *P. tetraurelia*

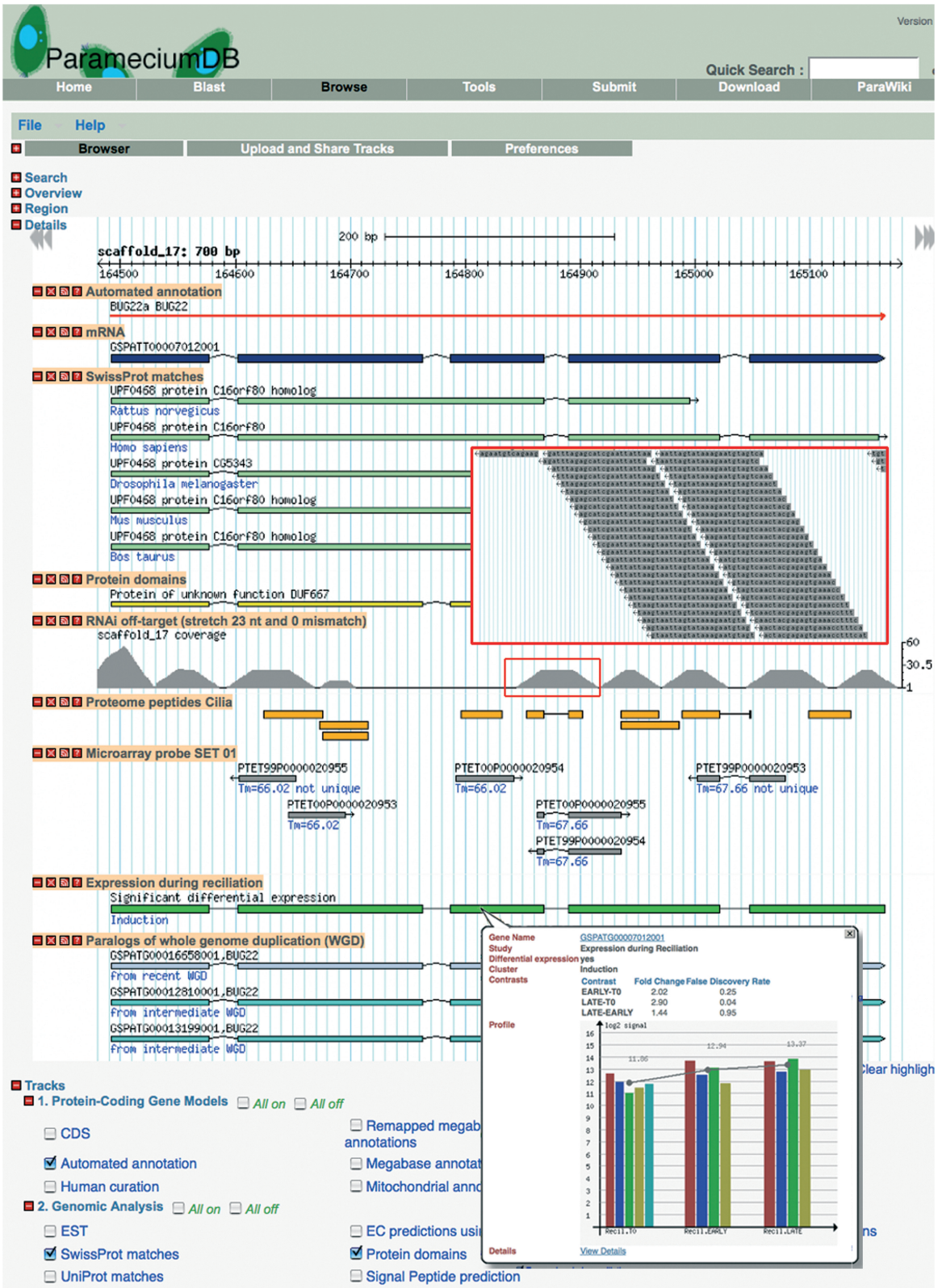

Figure 1. ParameciumDB Genome Browser. The screenshot displays a GBrowse view of the conserved *BUG22a* gene, required for efficient ciliary beating in Paramecium (34). Only a subset of available tracks is shown. Four *BUG22* genes, paralogs of the recent and intermediate WGDs, encode identical Bug22 proteins (cf. 'Paralogs of whole genome duplication' track). The RNAi off-target track and the enlargement (obtained by zooming to a 90 nt window) show the off-target matches to the gene. These matches are from the recent WGD paralog, *BUG22b*, except at the 5′ end of the gene where there are also contributions from the other paralogs, as appreciated when 'expand and label' is chosen to view the track after zooming to the sequence level. Only four of the six microarray probes are unique in the genome owing to the high nucleotide identity between the two paralogs related by the recent WGD. The popup balloon that appears on a mouse click of the 'Expression during reciliation' track, shows details of the microarray signals such as a significant 2.9-fold up-regulation of the gene. The ciliary proteome track shows a number of Bug22p peptides, and a popup balloon (not shown) links to details about the experiment and the peptides.

Table 1. ParameciumDB tools

Service	Software and version	References
Complex queries	BioMart 0.7	Smedley *et al.* (22)
BLAST	NCBI Blast 2.2.15	Camacho *et al.* (30)
Align two sequences	Smith–Waterman, EMBOSS 4.0.0	Rice *et al.* (31)
Multiple alignment	Muscle 3.7	Edgar (32)
Get sequence scaffold	Bio::DB::Fasta, Bioperl-live	Stajich *et al.* (33)
Motif search	patmatdb, EMBOSS 4.0.0	Rice *et al.* (31)
Genome annotation editor	Apollo 1.11	Lewis *et al.* (27)
RNAi off-target	BWA 0.5.8, Samtools 0.1.7	Li and Durbin (26), Li *et al.* (25)
Genome Browser	GBrowse 2.03	Stein *et al.* (24)
Browse genetic data	Template toolkit 2.19	http://template-toolkit.org
Browse proteome data	Template toolkit 2.19	http://template-toolkit.org
Browse transcriptome data	Template toolkit 2.19	http://template-toolkit.org
Browse stock tubes	Template toolkit 2.19	http://template-toolkit.org
Browse chromosome synteny	In-house	Duret *et al.* (10)
Ontology browser	Obrowser 0.01	In-house, uses jquery and jquery.tree
Parawiki	Mediawiki 1.6.10	http://www.mediawiki.org

The different tools available in ParameciumDB for querying, viewing, analyzing and sharing data are shown. The references concern the software used (22,24–27,30–33), except for the 'Browse chromosome synteny' tool, which shows the recent WGD relationships between scaffolds determined by nucleotide alignment from the study described in (10).

reference somatic chromosomes and determine the sequence of the germ line genome. We also plan to develop ParameciumDB as a resource for comparative studies. The high-throughput and relatively low cost of NGS is making it possible to consider sequencing the genomes of other *Paramecium* species, such as the members of the *P. aurelia* sibling complex. Indeed, *Paramecium biaurelia* sequencing is already well advanced (M. Lynch, personal communication). ParameciumDB, which uses the same Chado database scheme originally designed by FlyBase to prepare for the Drosophila 12 genomes project (29), is ready for incorporation of the genomes of more species. However, a big challenge will be to integrate and/or develop tools for viewing and analyzing synteny within and between genomes. GBrowse_syn (http://gmod.org/wiki/GBrowse_syn) might provide a good solution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the GMOD community for the software components used by ParameciumDB and support. They are indebted to Mark Gibson, Jonathan Crabtree and Scott Cain for help with the interface of Apollo to ParameciumDB. They thank all members of the Cohen and Bétermier labs at the Centre de Génétique Moléculaire for testing and feedback and the *Paramecium* research community for contributed data and suggestions for improvements. ParameciumDB is developed in the context of the CNRS European Research Group 'Paramecium Genome Dynamics and Evolution'.

FUNDING

ParameciumDB funding is provided by the Centre National de la Recherche Scientifique and the Agence National de la Recherche (project ParaDice ANR-08-BLAN-0233). Funding for open access charge: Agence National de la Recherche (ANR).

Conflict of interest statement. None declared.

REFERENCES

1. Aury,J., 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. Sémon,M. and Wolfe,K.H. (2007) Consequences of genome duplication. *Curr. Opin. Genet. Dev.*, **17**, 505–512.
3. Catania,F., Wurmser,F., Potekhin,A.A., Przybos,E. and Lynch,M. (2009) Genetic diversity in the *Paramecium aurelia* species complex. *Mol. Biol. Evol.*, **26**, 421–431.
4. Martens,C., Vandepoele,K. and Van de Peer,Y. (2008) Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl Acad. Sci. USA*, **105**, 3427–3432.
5. Beisson,J. and Wright,M. (2003) Basal body/centriole assembly and continuity. *Curr. Opin. Cell Biol.*, **15**, 96–104.
6. Arnaiz,O., Malinowska,A., Klotz,C., Sperling,L., Dadlez,M., Koll,F. and Cohen,J. (2009) Cildb: a knowledgebase for centrosomes and cilia. *Database*, **2009**, doi:10.1093/database/bap022.
7. Bétermier,M. (2004) Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate *Paramecium*. *Res. Microbiol.*, **155**, 399–408.
8. Duharcourt,S., Lepère,G. and Meyer,E. (2009) Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. *Trends Genet.*, **25**, 344–350.
9. Galvani,A. and Sperling,L. (2002) RNA interference by feeding in *Paramecium*. *Trends Genet.*, **18**, 11–12.
10. Duret,L., Cohen,J., Jubin,C., Dessen,P., Goût,J., Mousset,S., Aury,J., Jaillon,O., Noël,B., Arnaiz,O. *et al.* (2008) Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Res.*, **18**, 585–596.

11. 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.
12. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
13. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
14. Mungall,C.J., Gkoutos,G.V., Smith,C.L., Haendel,M.A., Lewis,S.E. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.
15. Gogendeau,D., Beisson,J., de Loubresse,N.G., Le Caer,J., Ruiz,F., Cohen,J., Sperling,L., Koll,F. and Klotz,C. (2007) An Sfi1p-like centrin-binding protein mediates centrin-based Ca2+ -dependent contractility in *Paramecium tetraurelia*. *Eukaryot. Cell*, **6**, 1992–2000.
16. Gogendeau,D., Klotz,C., Arnaiz,O., Malinowska,A., Dadlez,M., de Loubresse,N.G., Ruiz,F., Koll,F. and Beisson,J. (2008) Functional diversification of centrins and cell morphological complexity. *J. Cell Sci.*, **121**, 65–74.
17. Arnaiz,O., Goût,J.-F., Bétermier,M., Bouhouche,K., Cohen,J., Duret,L., Kapusta,A., Meyer,E. and Sperling,L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*. *BMC Genomics*, in press.
18. Bemm,F., Schwarz,R., Förster,F. and Schultz,J. (2009) A kinome of 2600 in the ciliate *Paramecium tetraurelia*. *FEBS Lett.*, **583**, 3589–3592.
19. Chen,C., Zhou,H., Liao,J., Qu,L. and Amar,L. (2009) Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of *Paramecium tetraurelia*. *RNA*, **15**, 503–514.
20. O'Brien,K.P., Remm,M. and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
21. O'Connor,B.D., Day,A., Cain,S., Arnaiz,O., Sperling,L. and Stein,L.D. (2008) GMODWeb: a web framework for the Generic Model Organism Database. *Genome Biol.*, **9**, R102.
22. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart—biological queries made easy. *BMC Genomics*, **10**, 22.
23. Goecks,J., Nekrutenko,A. and Taylor,J., Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
24. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
25. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
26. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
27. Lewis,S.E., Searle,S.M.J., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglir,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
28. Beisson,J., Bétermier,M., Bré,M., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. *et al.* (2010) *Paramecium tetraurelia*: the renaissance of an early unicellular model. *Cold Spring Harb. Protoc.*, **2010**, pdb.emo140.
29. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
30. 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.
31. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
32. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
33. 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.
34. Laligné,C., Klotz,C., de Loubresse,N.G., Lemullois,M., Hori,M., Laurent,F.X., Papon,J.F., Louis,B., Cohen,J. and Koll,F. (2010) Bug22p, a conserved centrosomal/ciliary protein also present in higher plants, is required for an effective ciliary stroke in *Paramecium*. *Eukaryot. Cell*, **9**, 645–655.