PlasmoDB: a functional genomic database for malaria parasites

Cristina Aurrecoechea1, John Brestelli2, Brian P. Brunk2, Jennifer Dommer2, Steve Fischer2, Bindu Gajria2, Xin Gao2, Alan Gingle3, Greg Grant4, Omar S. Harb2,*, Mark Heiges1, Frank Innamorato2, John Iodice2, Jessica C. Kissinger1,5, Eileen Kraemer6, Wei Li2, John A. Miller6, Vishal Nayak2, Cary Pennington1, Deborah F. Pinney2, David S. Roos7, Chris Ross1, Christian J. Stoeckert Jr2, Charles Treatman2 and Haiming Wang1

1Center for Tropical & Emerging Global Diseases, University of Georgia, Athens, GA 30602, 2Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, 3Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, 4School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, 5Department of Genetics, 6Department of Computer Science, University of Georgia, Athens, GA 30602 and 7Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

Received September 15, 2008; Accepted October 3, 2008

ABSTRACT
PlasmoDB (http://PlasmoDB.org) is a functional genomic database for Plasmodium spp. that provides a resource for data analysis and visualization in a gene-by-gene or genome-wide scale. PlasmoDB belongs to a family of genomic resources that are housed under the EuPathDB (http://EuPathDB.org) Bioinformatics Resource Center (BRC) umbrella. The latest release, PlasmoDB 5.5, contains numerous new data types from several broad categories—annotated genomes, evidence of transcription, proteomics evidence, protein function evidence, population biology and evolution. Data in PlasmoDB can be queried by selecting the data of interest from a query grid or drop down menus. Various results can then be combined with each other on the query history page. Search results can be downloaded with associated functional data and registered users can store their query history for future retrieval or analysis.

INTRODUCTION
Plasmodium spp. are obligate intracellular protozoan parasites of humans and animals, and are the causative agents of malaria. Transmission of these parasites to humans occurs via the Anopheles mosquito vector and the geographic distribution of endemic regions puts almost half of the world’s population at risk to contracting malaria. This disease is a major source of morbidity and mortality worldwide, which results in 300–500 million clinical cases and 1–2 million deaths annually (1,2). While several species of Plasmodium cause disease in humans (including P. vivax, P. malariae, P. ovale and P. knowlesi), P. falciparum is by far the deadliest (1,3). The life cycle of the Plasmodium parasite takes it through multiple cell types (in the vertebrate host and arthropod vector) during which the parasite undergoes multiple developmental changes (both sexual and asexual). The different life-cycle stages are marked by specific genomic, transcriptomic, proteomic and metabolomic states. Understanding how these changes are triggered and orchestrated requires mechanisms to view and interrogate genomic and functional genomic data in a powerful and intuitive manner. Over the past 10 years, PlasmoDB has evolved into a venue that integrates such data and allows the user to perform complex queries tailored to their specific needs and interests.

UPDATED DATA CONTENT
The data available in PlasmoDB has expanded to include genomic and functional data from eight Plasmodium species and is summarized in Table 1 (4). The current release (PlasmoDB 5.5) contains fully sequenced and annotated genomes of P. falciparum, P. vivax, P. yoelii, P. berghei,
Table 1. Types of data available in PlasmoDB and example queries

| Type of Data          | Species for which this data is available                                                                 | Example query                                                                 |
|-----------------------|-----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Genomic data          |                                                                                                           |                                                                               |
| Full sequence and annotation | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi*                     | Search annotations for specific keyword (see Figure 1C).                        |
| Sequence only         | *P. reichenowi*, *P. gallinaceum*                                                                         | Find sequence similarity using BLAST.                                          |
| Transcript expression data | *P. falciparum*, *P. berghei*, *P. yoelii*                                                               | Identify genes expressed at specific life-cycle stages.                        |
| Microarray            |                                                                                                           | Confirm gene models and alternative gene models.                               |
| EST                   | *P. falciparum*, *P. vivax*, *P. berghei*, *P. yoelii*                                                   | Identify genes with transcript evidence.                                       |
| SAGE                  | *P. falciparum*                                                                                            | Identify genes with protein expression evidence at specific life-cycle stages.|
| Protein expression data | *P. falciparum*, *P. berghei*, *P. yoelii*                                                               |                                                                               |
| Population biology    | *P. falciparum*                                                                                            | Find highly polymorphic genes or distinguish isolates based on their SNP profile.|
| SNP                   |                                                                                                           |                                                                               |
| Microsatellite        |                                                                                                           |                                                                               |
| Isolate data          |                                                                                                           |                                                                               |
| Protein interaction   | *P. falciparum*                                                                                            | Identify possible interaction partners of a gene of interest.                   |
| Yeast two hybrid      |                                                                                                           |                                                                               |
| Interactome map       |                                                                                                           |                                                                               |
| Putative function     | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi*                      | Identify genes that have GO annotations.                                       |
| GO annotation         | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. knowlesi*                                    | Identify genes with enzymatic annotations.                                    |
| EC numbers            | *P. falciparum*, *P. yoelii*, *P. knowlesi*                                                               | Identify parasite-specific or missing metabolic pathways.                       |
| Metabolic pathways    | *P. falciparum*                                                                                            |                                                                               |
| Evolutionary          |                                                                                                           |                                                                               |
| Orthology based       | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi*                     | Identify genes specific to apicomplexa.                                        |
| Homology based        | *P. falciparum* and *P. yoelii*                                                                           | Identify homologs of a gene or list of genes of interest.                      |
| Protein features      |                                                                                                           |                                                                               |
| Protein motifs        | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi*                      | Identify genes with specific protein attributes.                               |
| Interpro/pfam domains |                                                                                                           |                                                                               |
| Molecular weight      |                                                                                                           |                                                                               |
| Isoelectric point     |                                                                                                           |                                                                               |
| Protein structure     |                                                                                                           |                                                                               |
| Immune epitopes       |                                                                                                           |                                                                               |
| Protein localization  |                                                                                                           |                                                                               |
| Signal peptide        | *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi*                     | Identify genes targeted to the host cell.                                      |
| Transmembrane domains |                                                                                                           |                                                                               |
| Targeting to the RBC  |                                                                                                           |                                                                               |
| Apicoplast targeting  | *P. falciparum*                                                                                            | Identify genes targeted to the apicoplast.                                     |

*P. chabaudi* and *P. knowlesi*. Importantly, PlasmoDB 5.5 contains results of annotation efforts from multiple sources including the recent systematic effort to update the *P. falciparum* genome that is an ongoing project started at a workshop in late 2007 co-organized by the Wellcome Trust Sanger Institute (WTSI) and EuPathDB (formerly ApiDB) teams. Reannotation data have been released in incremental steps (snapshots) in order to provide timely information to users of PlasmoDB and to solicit user comments regarding the reannotations.

Transcript expression data [microarray, expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE)] available through PlasmoDB has expanded dramatically over the past few releases to include microarray data from multiple life-cycle stages, gene knock-out mutants of *P. falciparum* and *P. berghei* (5–12) and multiple stages of *P. yoelii* (mosquito, erythrocytic and liver stages) (13). Also included are EST data from over 130 libraries (*P. falciparum*, *P. vivax*, *P. berghei* and *P. yoelii*) (14,15) [dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) and SAGE data (*P. falciparum* only) (16–18). Protein expression evidence includes data from various life-cycle stages (*P. falciparum*, *P. berghei* and *P. yoelii*) (11,13,19–21; Leiden Malaria Group, unpublished data).

Population biology evidence (*P. falciparum* only) includes mapping of microsatellite data (22) onto the genome (available as a genome browser track), single nucleotide polymorphism (SNP) data from resequencing efforts of more than 20 *P. falciparum* strains (*P. reichenowi* is included as an out-group for comparison purposes) and data from nearly 100 *P. falciparum* isolates (23–25).

OrthoMCL analyses provide ortholog determinations between the different species facilitating discovery of shared genes between lineages (26). Protein function assignments are aided by a number of additional functional data types available through PlasmoDB 5.5 including...
Figure 1. Screenshots from PlasmoDB 5.5 and query workflow. (A) The top of the screenshot shows the PlasmoDB logo. On the left side are links to various sections of PlasmoDB and a point for logging in or registering as a user (not required for using the site but useful for storing search histories. The query grid is in the center and provides an access point to all searchable data in PlasmoDB. (B) This is a scheme of a workflow that a user may follow when building a set of queries. Beginning at the left, queries can be performed starting from the query grid and the results can be joined using operations available through the query history page. (C) Screen shots of a ‘key word’ search page, an example gene query history and a gene results page. Note the add column feature in the results page that allows the addition of columns with additional data and the ability to sort results.
Finally, examination of the list reveals several genes have proteomic evidence in gametocyte yields 27 genes). Further, one may be interested in the genes that have protein expression evidence in a particular stage in the parasite’s life cycle (the results of an intersection with genes that have proteomic evidence in gametocyte yields 27 genes). Finally, examination of the list reveals several genes encoding for rifins (a family of clonally variant proteins expressed on the surface of infected red blood cells) (38), and a user may wish to investigate genes other than rifins—this can be accomplished by excluding (‘not’ operation) results of a keyword query using the term ‘rifin’ (Figure 1B, Step 5 and Figure 1C, left most panel). A user may examine the specific gene pages for more gene-specific details, download results with their associated data or log in (if they have not done so already) to ensure that their search strategy is saved for future examination.

HOW TO USE PLASMODB

A visitor to PlasmoDB can use the database in two general ways: (i) To retrieve all available information associated with a particular gene of interest using a search for an exact gene ID, gene name or gene product name. (ii) To ask single questions (Table 1) and/or conduct a series of searches followed by refining the results by combining them or subtracting them from one another. Starting with the PlasmoDB home page (Figure 1A), a user can perform a quick search by entering an identifier or test term, or select a specific query from a number of drop-down menus (data not shown). Alternatively, queries may be accessed by visiting the ‘Queries and Tools’ section of PlasmoDB (Figure 1A), which includes a grid displaying all available queries/searches. By using the queries and tools, a user can interrogate data in PlasmoDB—the third column of Table 1 includes example data-specific questions that are available.

When conducting queries with the purpose of combining results it may be useful to visualize the searches in a workflow environment where nodes are connected using different criteria (‘and’, ‘or’, ‘not’) (Figure 1B). In PlasmoDB this would be accomplished by performing a number of queries and subsequently combining the results in the ‘query history’ section (Figure 1C, middle screen shot). For example, one may be interested in identifying a short list of possible vaccine candidates. One possible way of accomplishing this would be by identifying all proteins predicted to be exported to the host cell in *P. falciparum*. There are three exported protein datasets in PlasmoDB and a union (‘or’ function) of all three results retrieves 405 genes (Figure 1B, steps 1 and 2). To restrict this list further, intersecting (‘and’ function) these results with genes that have no orthologs in mammals reduces the results to 321 genes (Figure 1B, Step 3). Next a user may further prune this list by intersecting the results with other queries, such as genes that are nonpolymorphic between a chloroquine sensitive (3D7) and resistant strain (Dd2). This cuts the number of candidates to 32 genes (Figure 1B, Step 4 and Figure 1C, right screen shot). Alternatively, one may be interested in the genes that have protein expression evidence in a particular stage in the parasite’s life cycle (the results of an intersection with genes that have proteomic evidence in gametocyte yields 27 genes). Further, one may be interested in the genes that have protein expression evidence in a particular stage in the parasite’s life cycle (the results of an intersection with genes that have proteomic evidence in gametocyte yields 27 genes). Finally, examination of the list reveals several genes encoding for rifins (a family of clonally variant proteins expressed on the surface of infected red blood cells) (38), and a user may wish to investigate genes other than rifins—this can be accomplished by excluding (‘not’ operation) results of a keyword query using the term ‘rifin’ (Figure 1B, Step 5 and Figure 1C, left most panel). A user may examine the specific gene pages for more gene-specific details, download results with their associated data or log in (if they have not done so already) to ensure that their search strategy is saved for future examination.

FUTURE DIRECTIONS

It is expected that PlasmoDB will continue its data content and tool expansion as user needs require. We anticipate the incorporation of multiple new data sets including microarray, proteomic and specific parasite isolate data. Additionally, over the next few years we look forward to incorporating sequence data from a dramatically expanded *Plasmodium spp.* sequencing effort (http://www.genome.gov/26525388). In the coming year, we will also release a new user interface that will include a workflow-based search strategy page, similar to what is shown in Figure 1B, which we anticipate will provide a more biologically intuitive and dynamic experience for scientists accessing PlasmoDB and other EuPathDB sites.

ACKNOWLEDGEMENTS

The authors wish to thank members of the *Plasmodium* research community for their willingness to share genomic-scale data sets, often prior to publication, and for numerous comments and suggestions that have helped to improve the functionality of PlasmoDB. We also wish to thank Dr Akhil Vaidya for his valuable advice and continued support to PlasmoDB. We also thank past and present staff associated with the ApiDB-BRC project, and our research laboratory colleagues whose contributions have facilitated the creation and maintenance of this database resource.

FUNDING

Federal funds from the National Institute of Allergy and Infectious Diseases; National Institutes of Health; Department of Health and Human Services, under Contract No. HHSN266200400037C. Funding to pay the Open Access publication charges for this article was provided by this contract.

Conflict of interest statement. None declared.

REFERENCES

1. Phillips, R.S. (2001) Current status of malaria and potential for control. *Clin. Microbiol. Rev.*, 14, 208–226.
2. Snow,R.W., Guerra,C.A., Noor,A.M., Myint,H.Y. and Hay,S.I. (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, 434, 214–217.
3. Singh,B., Kim Sung,L., Matusop,A., Radhakrishnan,A., Shamsul,S.S., Cox-Singh,J., Thomas,A. and Conway,D.J. (2004)
A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *Lancet*, **363**, 1017–1024.

4. Bahl,A., Brunck,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al*. (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.

5. Bozdech,Z., Linas,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS* Biol., **1**, E5.

6. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De La Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. *et al*. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.

7. Linas,M., Bozdech,Z., Wong,E.D., Adai,A.T. and DeRisi,J.L. (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.*, **34**, 1166–1173.

8. Baum,J., Maier,G., Good,R.T., Simpson,K.M. and Cowman,A.F. (2005) Invasion by *P. falciparum* merozoites suggests a hierarchy of molecular interactions. *PLoS* Pathog., **1**, e37.

9. DeRisi,J.L., Voss,T.S., Marty,A.J., Duffy,M.F., Good,R.T., Duraisingh,M.T., Maier,A.G., Winzeler,E.A. and Cowman,A.F. (2003) Drug-induced alterations in gene expression of the asexual parasite. *Mol. Biochem. Parasitol.*, **136**, 1–11.

10. Capra,A.R., DeRisi,J.L. (2005) The transcriptome of the intraerythrocytic parasitic stage. *Science*, **307**, 305–310.

11. Haldar,K., Miller,M., Bogart,L., Schaffner,S.F., Milner,D.A. Jr, Daily,J.P., Szr,S.O., Ndiaye,D., Ndir,O. *et al*. (2007) A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.*, **39**, 113–119.

12. 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. *Nat. Genet.*, **25**, 25–29.

13. Blank,A., Small,J., Cutlip,M., Jordan,J.L., Wang,C., Johnson,M. *et al*. (2008) A database for comparative study of transmembrane protein topology with a hidden Markov model: application to complete genomes. *Nucleic Acids Res.*, **36**, D363–D368.

14. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Boeckmann,B., Bryant,S.H., Cerutti,L. *et al*. (2004) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.

15. Navalon,G., Sadowski,P., Zhu,M., Levesque,J.F., Boucrot,E., Hellewell,J. *et al*. (2006) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, **438**, 103–107.

16. Tabb,D.L., et al. (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, **419**, 520–526.