Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw1105

EuPathDB: the eukaryotic pathogen genomics database resource

Cristina Aurrecoechea[1], Ana Barreto[2,3], Evelina Y. Basenko[1], John Brestelli[2,3], Brian P. Brunk[2,4], Shon Cade[4], Kathryn Crouch[5], Ryan Doherty[2,4], Dave Falke[1], Steve Fischer[2,3], Bindu Gajria[2,4], Omar S. Harb[2,4,*], Mark Heiges[1], Christiane Hertz-Fowler[6], Sufen Hu[2,4], John Iodice[2,3], Jessica C. Kissinger[1,7,8], Cris Lawrence[2,4], Wei Li[2,4], Deborah F. Pinney[2,3], Jane A. Pulman[9], David S. Roos[4], Achchuthan Shanmugasundram[9], Fatima Silva-Franco[9], Sascha Steinbiss[10], Christian J. Stoeckert, Jr[2,3], Drew Spruill[1], Haiming Wang[1], Susanne Warrenfeltz[1] and Jie Zheng[2,3]

[1]Center for Tropical & Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA, [2]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [3]Department of Genetics, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [4]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA, [5]Wellcome Trust Centre for Molecular Parasitology, University of Glasgow, Glasgow G12 8TA, UK, [6]Centre for Genomic Research, University of Liverpool, Liverpool L69 7ZB, UK, [7]Department of Genetics, University of Georgia, Athens, GA 30602, USA, [8]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, [9]Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK and [10]Wellcome Trust Sanger Institute, Parasite Genomics, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

Received October 13, 2016; Editorial Decision October 25, 2016; Accepted October 28, 2016

ABSTRACT

The Eukaryotic Pathogen Genomics Database Resource (EuPathDB, http://eupathdb.org) is a collection of databases covering 170+ eukaryotic pathogens (protists & fungi), along with relevant free-living and non-pathogenic species, and select pathogen hosts. To facilitate the discovery of meaningful biological relationships, the databases couple preconfigured searches with visualization and analysis tools for comprehensive data mining via intuitive graphical interfaces and APIs. All data are analyzed with the same workflows, including creation of gene orthology profiles, so data are easily compared across data sets, data types and organisms. EuPathDB is updated with numerous new analysis tools, features, data sets and data types. New tools include GO, metabolic pathway and word enrichment analyses plus an online workspace for analysis of personal, non-public, large-scale data. Expanded data content is mostly genomic and functional genomic data while new data types include protein microarray, metabolic pathways, compounds, quantitative proteomics, copy number variation, and polysomal transcriptomics. New features include consistent categorization of searches, data sets and genome browser tracks; redesigned gene pages; effective integration of alternative transcripts; and a EuPathDB Galaxy instance for private analyses of a user's data. Forthcoming upgrades include user workspaces for private integration of data with existing EuPathDB data and improved integration and presentation of host–pathogen interactions.

INTRODUCTION

A unique infrastructure and search strategy system distinguish the Eukaryotic Pathogen Database Resource (EuPathDB, http://eupathdb.org) from other organism databases. The power of EuPathDB lies in the ability to query across hundreds of data sets while refining a set of genes, proteins, pathways or organisms of interest. The interface is designed for easy mastery by biological researchers, enabling *in silico* experiments that interrogate diverse and complex data sets. Despite the sophisticated strategy system, browsing gene pages and genomic spans or regions remains a simple and informative task in this innovative and valuable resource.

EuPathDB facilitates the discovery of meaningful biological relationships between genomic features such as genes

*To whom correspondence should be addressed. Tel: +1 215 746 7019; Fax: +1 215 573 3111; Email: oharb@sas.upenn.edu

Table 1. EuPathDB resources and organisms supported

Database	Web address	Link to access list of organisms supported
EuPathDB	http://eupathdb.org	EuPathDB organisms
AmoebaDB	http://amoebadb.org	AmoebaDB organisms
CryptoDB	http://cryptodb.org	CryptoDB organisms
FungiDB	http://fungidb.org	FungiDB organisms
GiardiaDB	http://giardiadb.org	GiardiaDB organisms
HostDB	http://hostdb.org	HostDB organisms
MicrosporidiaDB	http://microsporidiadb.org	MicrosporidiaDB organisms
PiroplasmaDB	http://piroplasmadb.org	PiroplasmaDB organisms
PlasmoDB	http://plasmodb.org	PlasmoDB organisms
ToxoDB	http://toxodb.org	ToxoDB organisms
TrichDB	http://trichdb.org	TrichDB organisms
TriTrypDB	http://tritrypdb.org	TriTrypDB organisms
OrthoMCL	http://orthomcl.org	Includes proteins from over 150 organisms across bacteria, archaea and eukarya

or SNPs by integrating pre-analyzed data with sophisticated data mining, visualization and analysis tools that are designed to be used by wet-bench researchers. Organized into 13 free, online databases EuPathDB supports over 170 eukaryotic pathogens with genomic sequence and annotation, functional genomics data, host-response data, isolate and population data and comparative genomics. Table 1 provides a web address and a link to a list of organisms supported for each database. All databases are built with the same infrastructure and use the Strategies Web Development Kit (1), which provides a graphical interface for building complex search strategies and exploring relationships across data sets and data types (Figure 1; strategy http://plasmodb.org/plasmo/im.do?s=7b88206dd42007c8).

As one of four National Institute of Allergy and Infectious Disease (NIAID/NIH) funded Bioinformatics Resource Centers (2–6) EuPathDB provides data, tools and services to scientific communities researching pathogens in the NIAID list of emerging and re-emerging infectious diseases which includes NIAID category A–C priority pathogens and many fungi. Additional EuPathDB support for the kinetoplastid and fungal research communities is funded by The Wellcome Trust in collaboration with GeneDB (7), including support for focused curated annotation. This manuscript describes expanded content, features and tools added since 2013 that increase the data mining and discovery power of EuPathDB.

NEW IN EuPathDB

Over the past 4 years, EuPathDB has routinely updated existing databases and added two new databases. We added new data, expanded the range of supported data types, enhanced infrastructure and added new analysis tools.

Databases

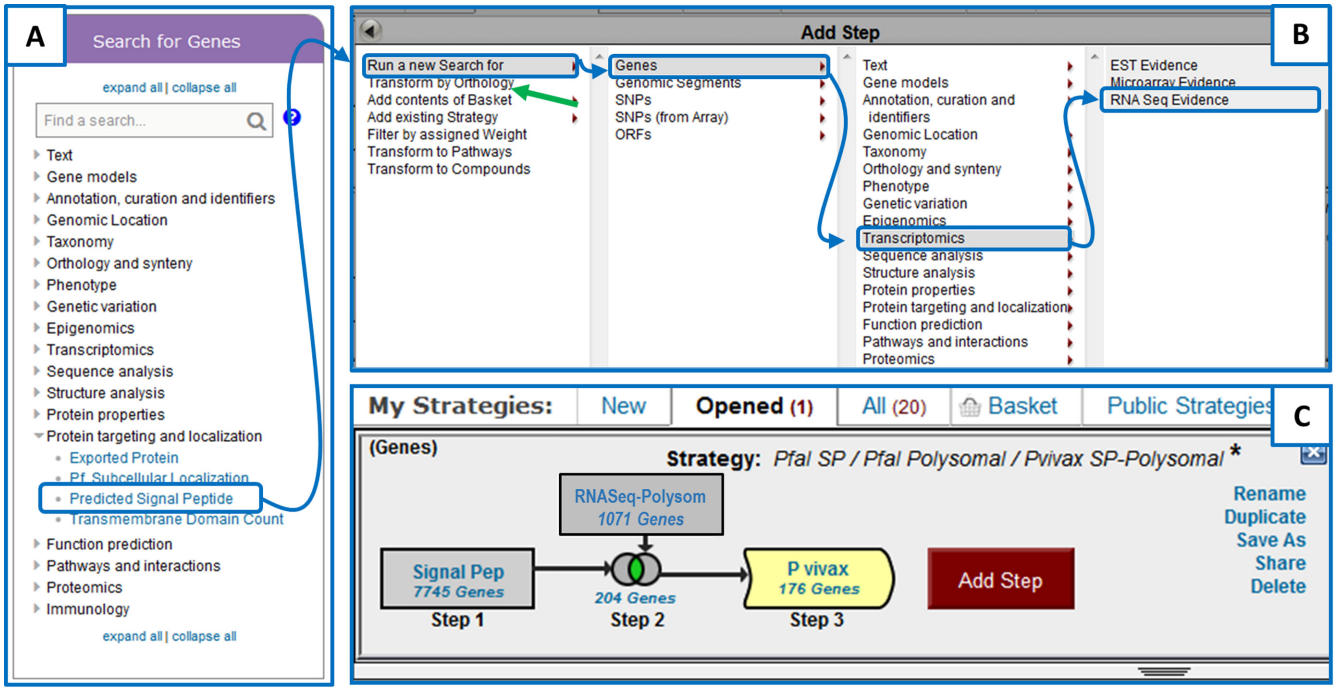
EuPathDB resources have been expanded to include FungiDB (http://fungidb.org) (8), which supports fungi and oomycetes, and HostDB (http://hostdb.org), for interrogation of host responses to infection. HostDB supports host data obtained during infections by organisms supported by EuPathDB's 10 parasite lineage-specific databases. Minot *et al*. (9), for example, infected murine macrophages with 29 *Toxoplasma gondii* strains and collected mixed parasite–host samples for RNA sequencing. Reads that align to the *T. gondii* genome are integrated into ToxoDB whereas HostDB houses those sequencing reads that align to the *M. musculus* genome. Because all EuPathDB databases employ the same data analysis pipelines, search strategy system, visualization and analysis tools, the *T. gondii* and *M. musculus* data can be compared. For example, one can easily identify parasite genes that are differentially expressed between two *T. gondii* strains from ToxoDB as well as host genes that are differentially expressed during infection with the same two strains from HostDB. Enrichment analyses and comparison of these lists offers insights into host–pathogen interactions and responses.

Tools

EuPathDB tools are conceived and designed to reduce analysis barriers, enhance data mining and improve communication within and between the scientific communities we serve. The near-seamless integration of strategy results with tools for functional enrichment analyses and transcript interpretation as well as our new Galaxy workspace and the availability of publicly shared strategies augment the data mining experience in EuPathDB.

Galaxy workspace. EuPathDB sites now include a Galaxy-based (10) workspace for large-scale data analyses, e.g. RNA-seq read mapping to a reference genome. Developed in partnership with Globus Genomics (11), workspaces offer a private analysis platform with published workflows and pre-loaded annotated genomes for the organisms we support. The workspace is accessed through the 'Analyze My Experiment' (Figure 2A) tab on the home page of any EuPathDB resource and can be used to upload your own data *e.g.* RNA-seq reads, compose and run preconfigured or custom workflows (Figure 2B and C), retrieve your results, visualize them in EuPathDB (Figure 2D), and share workflows and data analysis results with colleagues.

Explore transcript subsets. Transcript subsets occur when a multi-transcript gene has at least one transcript that does not meet the search criteria. For example, signal peptides are short sequences at the N-terminus of secretory proteins



D

	Search Returns	Combined result is	Data Type	Data Set
Step 1	Genes that encode proteins with predicted signal peptides	N/A	Genomic analysis	Results of SignalP program run on all annotated proteins in PlasmoDB, 7745 genes identified.
Step 2	Genes expressed in the 80-100th percentile of merozoite sample analyzed by polysomal trancriptomics	Intersection: Signal peptide containing genes that are likely translated in merozoites	Transciptomics	PlasmoDB workflow results from the analysis of RNA sequence data of a ribosome profiling experiment, 204 genes identified.
Step 3	Pvivax orthologs of SP containing genes that are likely translated in merozoites	N/A	Orthology Tranform	Ortholog determination from OrthoMCL analysis of proteomes from 150 organisms across the tree of life, 176 orthologous genes identified.

Figure 1. PlasmoDB strategy showing graphical interface for exploring relationships across data sets, data types and organisms. (The strategy can be found here: http://plasmodb.org/plasmo/im.do?s=7b88206dd42007c8) (**A**) Home page bubble for choosing the first search of a strategy, showing the 'Predicted Signal Peptide' search categorized under 'Protein targeting and localization'. Clicking on the search title opens a form where users are prompted to choose required parameter values (if any) and initiate the search. The results of this search are displayed in Step 1 of panel C. (**B**) Interface for choosing subsequent searches. To add the Ribosomal profiling search that is based on RNA Seq data, users navigate the interface through 'Run a new search for', 'Genes', 'Transcriptomics', 'RNA Seq Evidence'. Alternatively, to transform a result in to orthologs of another species as in step 3 of the strategy, users choose 'Transform by Orthology' (green arrow) instead of the navigation indicated above. (**C**) Three-step strategy that returns *P. vivax* orthologs (Step 3) of *P. falciparum* genes that are likely translated in merozoites (step 2) and that are predicted to encode proteins with signal peptides. (**D**) Table detailing the data sets and data types interrogated in this strategy.

and EuPathDB predicts signal peptides for all annotated genomes using SignalP (12). The Predicted Signal Peptide search returns genes and transcripts with predicted signal peptides. If one transcript of a multi-transcript gene excludes the exon containing the signal peptide, the search returns the gene but not the signal peptide-deficient transcript. Searches and strategies that query transcript-specific data (Figure 3A; strategy http://plasmodb.org/plasmo/im.do?s=859df329f857438e) are equipped with an Explore tool for interrogating or filtering transcript subsets. The explore tool appears in the Gene Results tab above the table of IDs (Figure 3C) and offers filters for transcripts based on their inclusion in the result set. Filters are applied to the strategy result and update the gene result list. For two-step strategies where both steps query transcript specific data, the explore tool offers further filters for viewing transcripts that were returned by both searches, either search or neither search.

Enrichment analyses. Gene Ontology, Metabolic Pathway and Word enrichment analyses are available for gene strategy results to aid with their interpretation (Figure 3F). These functional analyses apply the Fisher's Exact test to determine over-represented pathways, ontology terms and product description terms. Clicking the Analyze Results tab of any gene strategy result (Figure 3E) and selecting an enrichment analysis will open an analysis tab where users are prompted for parameter values. The results of an enrichment analysis are presented in tabular form and include a list of enriched GO terms, pathways or product description words and associated data.

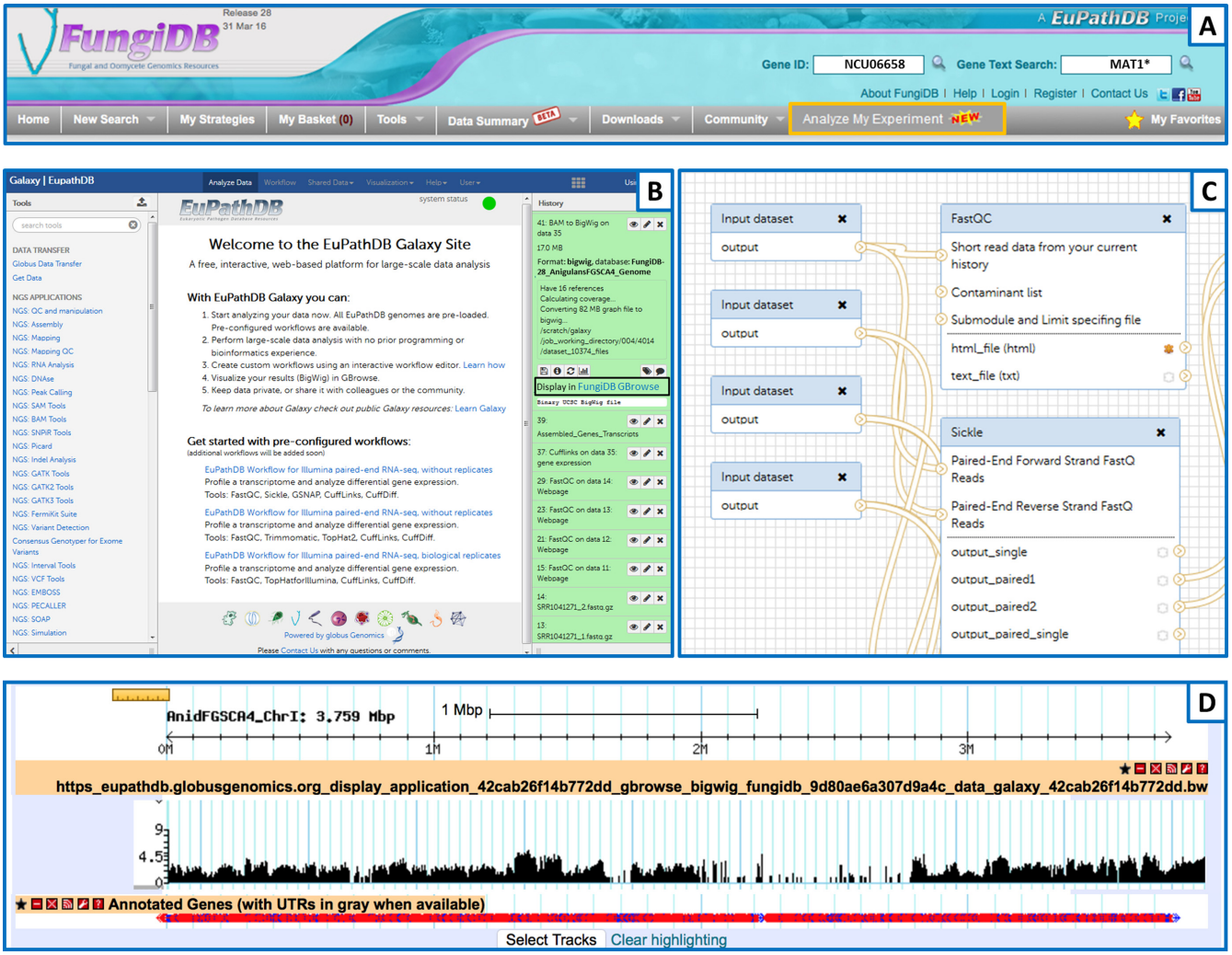


Figure 2. Galaxy Workspace. (**A**) FungiDB header showing the Analyze My Experiment (orange box) link for navigating to the EuPathDB Galaxy Workspace. (**B**) The EuPathDB Galaxy Workspace home page with preconfigured workflows available in the center section. Available tools are located in the left panel and the History panel showing result and data files on the right in green. The 'Display in FungiDB' link (black box) navigates to GBrowse with the Galaxy data file open as a data track in the user's current GBrowse session. (**C**) Partial workflow showing the 'drag and drop' function of building workflow. (**D**) Bigwig file displayed in FungiDB Gbrowse directly from EuPathDB Galaxy using the 'Display in FungiDB' (black box) link in panel B.

Public strategies. Strategies marked as Public when saved to a user's profile will also be shared with the community in the 'Public Strategies' tab of the 'My Strategies' interface. Users control the availability of the strategy and can remove it at any time. The panel also includes example strategies provided by EuPathDB.

Data sets search tool. Each data set integrated into EuPathDB is documented with a data set record which contains information about the data including a description, contact information for the investigator that generated the data, literature references, and when available, example graphs and links to searches and genome browser tracks. Links to data set records appear on gene pages and on search pages beneath the parameters. A searchable table of all data sets is available from the Data Summary tab in the gray drop-down menu bar.

Data content and data types

EuPathDB's philosophy is to provide a data mining platform that allows users to ask their own questions in support of hypothesis driven research. The extensive range of data types (genomic, transcriptomic, proteomic, metabolomic, etc.) maintained by EuPathDB broadens the user's ability to mine extensively by providing multiple forms of experimental evidence to interrogate. As the omics world expands, EuPathDB endeavors to support meaningful data types and has expanded its coverage over the past few years.

Genome sequence, annotation and functional genomics data. EuPathDB resources now support over 170 organisms with 255 genome sequences, 199 of which include genome-wide annotation. The addition of FungiDB as a EuPathDB resource brought many genomes from this large and diverse research community. Updates to EuPathDB's Reflow workflow system (2) make it possible to quickly and reliably an-

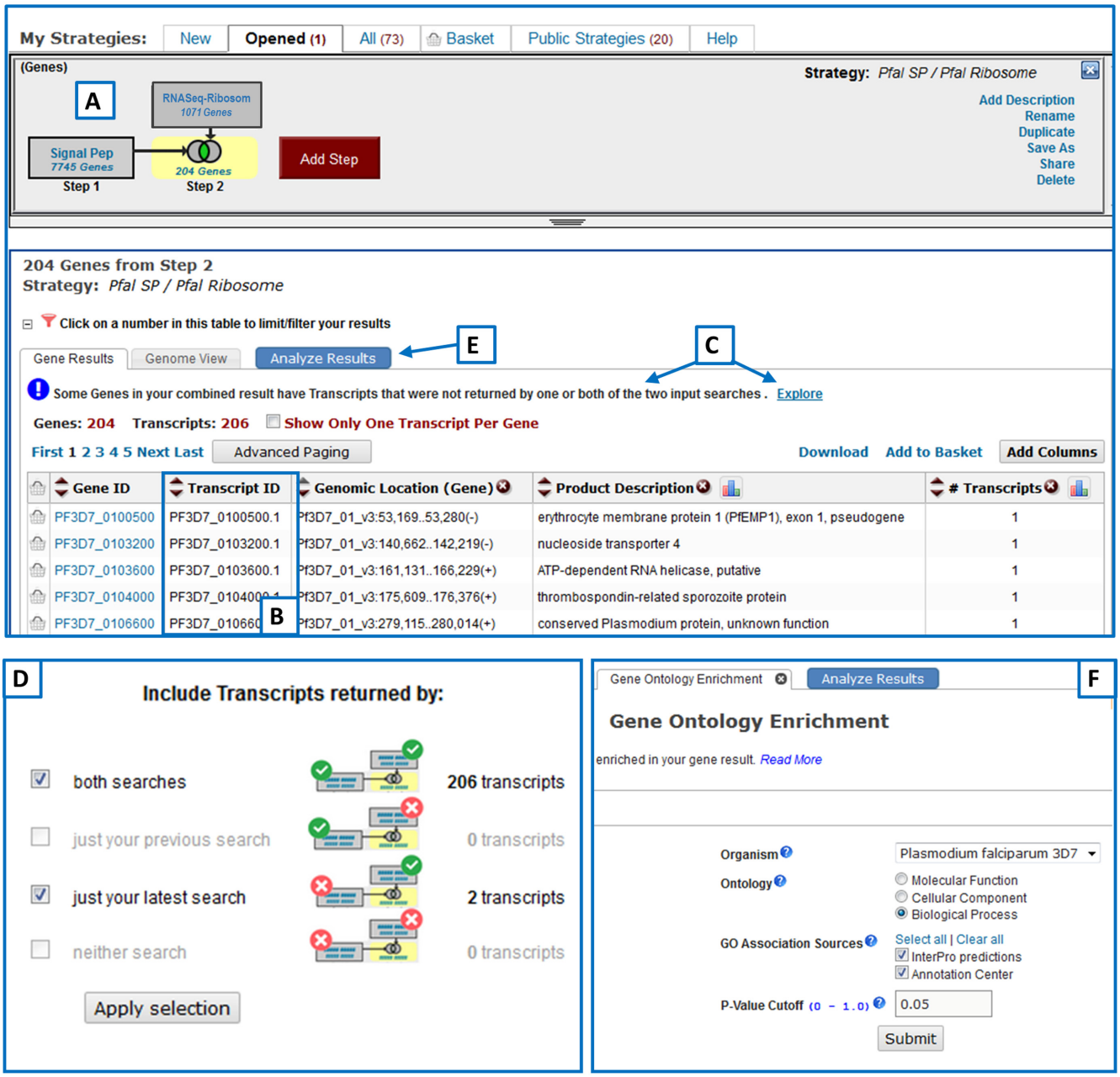


Figure 3. Explore transcripts and enrichment analyses. (**A**) PlasmoDB 2-step strategy that returns genes with signal peptides that are likely translated based on ribosomal transcriptomics data. This strategy can be found at http://plasmodb.org/plasmo/im.do?s=859df329f857438e (**B**) The result table contains a column of Transcript IDs. (**C**) When a search returns transcript subsets, the Gene Result tab will contain a statement inviting users to explore the transcript results. Clicking 'Explore' opens the Explore Transcripts tool. (**D**) The Explore Transcripts tool for viewing transcripts that did or did not meet the search criteria for the current or previous searches. Choosing an option and clicking Apply Selection will filter the strategy result and display your chosen transcripts in the Gene Result tab. (**E**) The Analyze Results Tab opens a new tab for your chosen enrichment analysis. (**F**) Gene Ontology Enrichment Analysis Tool. Analysis results appear below the parameters and include enriched terms plus *P*-values.

alyze and load data. Thus, over the past 4 years, numerous functional data sets have been loaded. Data sets of interest can be located with the data set search tool described above.

Protein microarray. This new data type offers a measure of host response to infection by revealing pathogen-specific antibodies in host serum or plasma samples. A typical data set includes data from serum samples collected from patients during an infection (or from healthy controls) that were hybridized to arrays spotted with possible pathogen antigens (peptides representing gene products) (13–16). Searches that query this data type are classified under Immunology and graphs of a pathogen gene's antigenicity for each sample appear on gene pages. The searches employ the filter parameter for selecting samples based on clin-

ical characteristics of patients when configuring the search (17).

Metabolic pathways. Pathways are integrated from MetaCyc, KEGG, TrypanoCyc and LeishCyc (18–22) as networks of enzymatic reactions and substrate/product compounds. Genes are mapped to Pathways based on EC numbers. Pathway record pages feature a Cytoscape image which can be 'painted' with experimental data, *e.g.* gene expression values or ortholog profiles. For easy transition to functional analysis, gene search results can be converted to pathways using the Transform to Pathways function in the Add Step popup or users can run a pathways enrichment analysis of their gene result to identify pathways that are statistically enriched.

Compounds. Compound records are integrated from the Chemical Entities of Biological Interest (ChEBI) database (23) and associated to genes through metabolic pathway mappings. Lists of compounds are returned based on molecular weight or formula, compound ID, enzyme EC number, Compound ID and text. Lists of genes and metabolic pathways can be transformed into their associated compounds using the Transform function.

Phenotypes of fitness from genome-wide CRISPR screen. A genome-wide loss of function screen using CRISPR technology is available in ToxoDB and provides a measure of a gene's contribution to parasite fitness (24). Phenotypes represent the fitness of CRISPR gene knockout organisms based on comparing the frequency of guide RNA sequences remaining in culture after three lytic cycles to the original guide RNA library. A search categorized under Phenotypes returns genes based on phenotype score. A GBrowse track showing guide RNAs mapped against the *T. gondii* GT1 genome is available.

Curated phenotypes. Phenotypes curated from the literature for several *Aspergillus* and *Cryptococcus* strains are now integrated into FungiDB. Phenotypes curated from the literature by the Sanger Institute's Pathogen Genomics Group based on siRNA data are available in TriTrypDB for *T. brucei brucei* strain 927. A phenotype table appears on the record pages of genes that have curated phenotypes. A search returning genes based on the phenotype is available and categorized under Phenotype in both FungiDB and TriTrypDB.

Quantitative proteomics. This new data type provides evidence for differential protein expression from experimental methods such as SILAC (25,26). The searches appear under the Proteomics, Quantitative Mass-Spec Evidence and return genes based on the fold change in protein expression between samples. Gene pages include graphs of these data when available.

Copy number variation. Whole genome resequencing data are used to estimate chromosome and gene copy number in re-sequenced strains (27). The median read depth is set to the organism's ploidy and each chromosome's median read depth is normalized to this value. Contigs that are not assigned to chromosomes are excluded from this analysis. Gene copy number is similarly calculated using a normalized read depth for each gene. To compare the number of genes in the re-sequenced genome to the reference genome, genes are grouped into clusters that are inferred to have originated by duplication. Searches are categorized under Genetic Variation and either returns genes with a certain copy number, or genes with different copy numbers between strains.

Polysomal transcriptomics. RNA-sequencing of polysome or ribosome associated transcripts reveals potential translation events. Data sets of this data type are available in PlasmoDB (28,29) and TryTripDB (30). Categorized under Transcriptomics, RNA Seq Evidence, the searches against this new data type return genes with differential translation potential (Fold Change search) or genes within a certain percentile rank within a sample. Gene pages contain expression graphs and RNA sequencing coverage plots are available statically in gene pages and dynamically in GBrowse. These coverage plots provide evidence for the CDS and translational start site usage.

Metadata. Biological sample characteristics such as host clinical parameters for pathogen isolates or blood samples offer valuable information for stratifying samples while configuring searches. EuPathDB integrates metadata when available and presents it in the filter parameter interface to take advantage of the rich data type when selecting samples for data mining (see below).

New features and infrastructure upgrades

The most recent EuPathDB release represents significant updates to the underlying data and infrastructure. In addition to refreshing all data to the latest versions, we added workspaces, redesigned our gene pages, incorporated alternative transcripts into gene pages and searches, updated search categories and contemporized the RNA sequence analysis workflow.

Categories. Searches, the experimental data sets they query, and genome browser tracks for visualization are now displayed with a common logic across the Web sites. The categories are based on the EMBRACE Data & Methods Ontology (EDAM) (31), which relates biological concepts with bioinformatic analyses. The result is a logical, consistent menu structure from home page to gene page to genome browser. For example, the category names and order in the home page 'Search for Genes' (Figure 1B) is the same as the 'Contents' section of the gene page (Figure 4C).

EuPathDB's extensive record system documents integrated data and analysis results for entities such as Genes, Genomic Sequences, SNPs, Isolates, Compounds and Metabolic Pathways. Record pages have a new streamlined look, contain improved navigation tools, and are reorganized to reflect EDAM-based categories (Figure 4). To view the gene page for PF3D7_0905700, autophagy-related protein 3, putative that is highlighted in Figure 4, go to http://plasmodb.org/plasmo/app/record/gene/

A

PF3D7_0905700 autophagy-related protein 3, putative

Name: ATG3
Type: protein coding
Chromosome: 09
Location: Pf3D7_09_v3:284,200..286,640(-)

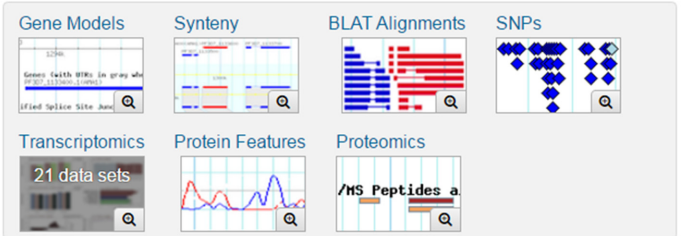
Species: Plasmodium falciparum
Strain: 3D7
Status: Curated Reference Strain

View updated annotation at GeneDB
View 5 user comments, or add a comment

GeneDB curates, researches and improves this genome, and will incorporate appropriate User Comments into the official annotation. If you wish to publish whole genome or large-scale analyses, please contact the primary investigator or use the published version in the PlasmoDB version 5.3 download folder.

B

Shortcuts



Also see PF3D7_0905700 in the Genome Browser or Protein Browser

C

Contents

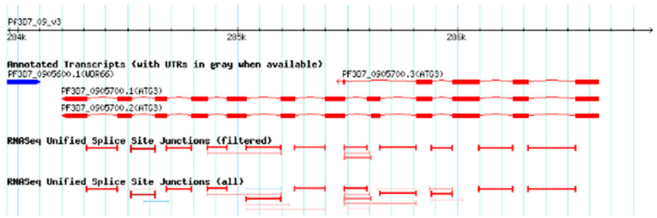
Search section names...

- 1 Gene models
- 2 Annotation, curation and identifiers
- 3 Genomic Location
- 4 Literature
- 5 Taxonomy
- 6 Orthology and synteny
- 7 Phenotype
- 8 Genetic variation
- 9 Transcriptomics
- 10 Sequences
- 11 Sequence analysis
- 12 Structure analysis
- 13 Protein properties
- 14 Function prediction
- 15 Pathways and interactions
- 16 Proteomics
- 17 Immunology

D

▼ Gene Models

View in genome browser



View in genome browser

▼ Transcripts Data sets

Search this table... Showing 3 rows

Transcript	# exons	Transcript length	Protein length
PF3D7_0905700.1	12	960	319
PF3D7_0905700.2	12	942	313
PF3D7_0905700.3	5	372	123

2 Annotation, curation and identifiers

▶ Alternate Product Descriptions Data sets

External DB Version 2015-06-18
Gene Name or Symbol ATG3

▶ Names, Previous Identifiers, and Aliases Data sets

▼ Notes from Annotator Data sets

Search this table... Showing 2 rows

Transcript ID(s)	Date	Note
PF3D7_0905700.1		also confirmed by cDNA amplification (comment from jbosch@u.washington.edu) alternate spliced form revealed by splice bridging read pair transcriptome sequence. Confirmed by 59 reads (PMID:20141604).
PF3D7_0905700.3		alternate spliced form revealed by splice bridging read pair transcriptome sequence. Confirmed by 3 reads (PMID:20141604).

Figure 4. Redesigned Gene Page. URL for this gene page- http://plasmodb.org/plasmo/app/record/gene/PF3D7_0905700 (**A**) Gene IDs and product descriptions are displayed in the upper left corner with other information and links directly below. (**B**) 'Shortcuts' serve two functions. Clicking on the Shortcut's magnifying glass icon offers a larger view of the data, while clicking on the image (or its title) navigates to the data within the gene page. (**C**) The collapsible, interactive and searchable 'Contents' section reflects EDAM-based categories and remains visible /stationary while scrolling the data (D). A blue section indicator (circle) points to the currently displayed data category. The check boxes to the right of the category names can be used to hide data. (**D**) Data is presented in collapsible, interactive, searchable, and sortable tables that contain transcript-specific information when data can be unambiguously assigned to a transcript. (**E**) The 'Transcriptomics' table featuring expandable rows with detailed information and graphs for each data set and coverage plots for RNA sequence data sets (showing one of eight tracks to conserve space in this figure). (**F**) Protein features table with the same expandable structure as the Transcriptomics table and showing protein domains, BLASTP Hits, Low Complexity Regions and Secondary Structure predictions.

Figure 5. Filter Parameter for composing sample groups based on metadata. (**A**) Samples are chosen from participants age 0 to 10. The left panel displays categories of sample characteristics while the right shows details of the data for that category. A summary of the sample group characteristics appears above the panel—333 out of 421 samples are below age 10.9 (blue arrow). (**B**) Adding a characteristic to refine the sample group. A second characteristic is chosen from the left panel (Health Status) and the Malaria group is chosen. The summary now shows the group characteristics—263 out of 421 samples have age <10.9 and malaria health status (blue arrow).

PF3D7_0905700. For example, in gene record pages, gene IDs and product descriptions are prominently displayed in the upper left corner of the page with other pertinent gene information and links directly below (Figure 4A). Also at the top of the page are 'Shortcuts' (Figure 4B) which serve two functions—clicking on the Shortcut's magnifying glass icon offers a larger view of the data, while clicking on the image (or its title) navigates to the data within the gene page. 'View in Genome Browser' links (e.g. above and below the Gene Models image in Figure 4D) accompany data that are also available for dynamic viewing in the Genome Browser. These links open the Genome Browser (GBrowse) (32) with the pertinent data track added to the user's current browser session.

The collapsible and interactive 'Contents' section reflects the new EDAM-based categories and features a search function for quickly locating a category (Figure 4C). The contents section remains stationary and visible while

scrolling the gene page data (Figure 4D). A section indicator (small blue circle) appears to the left of the category name of the data currently in view. Clicking a category name directs the page to that data section. The check boxes to the right of the category names can be used to customize the data display. Data from categories with empty check boxes will be hidden from view.

Data tables (4E, 4F and within Figure 4D) are collapsible, interactive, contain sortable columns and present transcript-specific information when data can be unambiguously assigned to a transcript. Tables with two or more rows include a search function. The Transcriptomics (Figure 4E), Protein Properties and Features (Figure 4F), Mass Spec -based Expression Evidence and Sequences tables contain expandable rows for retrieving detailed information. Each row of the Transcriptomics table represents a data set and expanding a row reveals graphs, data tables, and a data set description, as well as coverage plots for RNA sequencing data. Expansion of the rows in the Protein Properties and Features table reveals the domains, BLASTP hits and other analysis results pertinent to the transcript's protein product. The Mass Spec-based Expression Evidence Graphic table shows proteomic evidence associated with each transcript. The Sequences table offers genomic, coding, predicted mRNA and predicted protein sequences for each transcript.

Transcripts represented on gene pages and in search results. Human and mouse genes (HostDB) have extensive alternative transcripts and there is increasing evidence that many eukaryotic pathogen genes have more than one transcript. EuPathDB infrastructure was updated to better represent transcript information. Transcripts are graphically represented on gene pages and listed in gene page tables when data can be unambiguously assigned to a transcript (Figure 4D). All gene search results now include a Transcript ID column (Figure 3C). The results of searches that query transcript-specific data (e.g. Predicted Signal Peptide) contain an Explore Tool (see Tools section of this manuscript) for investigating transcript subsets (Figure 3B).

Filtering samples based on metadata. Sequences from pathogen isolates and data from host clinical blood samples are often accompanied by rich metadata-sample characteristics including host, age, geographic location, disease status and parasitemia. EuPathDB's new filter parameter (Figure 5) increases the user's power to mine data via display of sample characteristics (metadata) on the interface for selection of samples while configuring a search or multiple sequence alignment. For example, the filter parameter makes it possible to compare the antigenicity of parasite genes between infected children and uninfected children within the same dataset. The filter parameter is available for searches and sequence alignments that access SNP, ChIP-seq and host-response data.

RNA-sequence analysis workflow updated. Our pipeline for analyzing and loading RNA-sequence data was updated to use standard tools and to accommodate data sets with biological replicates. The new workflow aligns reads with GSNAP and calculates FPKM/RPKM with HT-Seq (33,34). DESeq2 is used to determine differential expression for experiments that have appropriate biological replicates (35).

Future directions

Future development efforts at EuPathDB will concentrate on expanding private analysis workspaces and better integration and support for host response to pathogen infection. The Galaxy toolshed contains many tools for data analysis. We expect to enhance our existing Galaxy workspace with new workflows such as alignment of resequencing reads and SNP calls or production of multiple sequence alignments and phylogenetic analyses. Critical to our expanded workspace will be the ability for users to fully integrate the results of their analyses into EuPathDB so that they can query, view, and share their results in the context of the publicly available data in EuPathDB.

A high priority for EuPathDB in the coming year is to better represent host responses to pathogen infection and enable users to mine these data to identify genes (or other entities) and relationships of interest. Currently, only a few omics data sets are available for host response, but we expect this situation to change rapidly. We will be expanding not only the amount of host data that we load, but also the types of host response data so that we can include high-throughput metabolic and immune profiling and rich descriptions of all study, experiment and sample metadata. We will be loading these rich multi-dimensional studies and we will be implementing a variety of tools and analyses to mine these data at a systems level.

ACKNOWLEDGEMENTS

The authors wish to thank members of the EuPathDB research communities for their willingness to share genomic-scale data sets, often prior to publication and for numerous comments and suggestions from our scientific advisors and the scientific community at large, which have helped to improve the functionality of EuPathDB resources. We also thank past and present staff associated with the EuPathDB BRC project, and our research laboratory colleagues whose contributions have facilitated the creation and maintenance of this database resource.

FUNDING

National Institute of Allergy and Infectious Diseases, National Institutes of Health [HHSN272201400030C to D.S.R. and J.C.K.]; The Wellcome Trust [108443/Z/15/Z, WT085822MA to C.H.F]. Funding for open access charge: JCK internal funds provided by University of Georgia.
Conflict of interest statement. None declared.

REFERENCES

1. Fischer,S., Aurrecoechea,C., Brunk,B.P., Gao,X., Harb,O.S., Kraemer,E.T., Pennington,C., Treatman,C., Kissinger,J.C., Roos,D.S. *et al.* (2011) The Strategies WDK: a graphical search interface and web development kit for functional genomics databases. *Database (Oxford)*, **2011**, bar027.

2. Aurrecoechea,C., Barreto,A., Brestelli,J., Brunk,B.P., Cade,S., Doherty,R., Fischer,S., Gajria,B., Gao,X., Gingle,A. *et al.* (2013) EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res.*, **41**, D684–D691.
3. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
4. Giraldo-Calderón,G.I., Emrich,S.J., MacCallum,R.M., Maslen,G., Dialynas,E., Topalis,P., Ho,N., Gesing,S., Madey,G., VectorBase Consortium *et al.* (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.*, **43**, D707–D713.
5. Pickett,B.E., Greer,D.S., Zhang,Y., Stewart,L., Zhou,L., Sun,G., Gu,Z., Kumar,S., Zaremba,S., Larsen,C.N. *et al.* (2012) Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses*, **4**, 3209–3226.
6. Squires,R.B., Noronha,J., Hunt,V., García-Sastre,A., Macken,C., Baumgarth,N., Suarez,D., Pickett,B.E., Zhang,Y., Larsen,C.N. *et al.* (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respir Viruses*, **6**, 404–416.
7. Logan-Klumpler,F.J., De Silva,N., Boehme,U., Rogers,M.B., Velarde,G., McQuillan,J.A., Carver,T., Aslett,M., Olsen,C., Subramanian,S. *et al.* (2012) GeneDB–an annotation database for pathogens. *Nucleic Acids Res.*, **40**, D98–D108.
8. Stajich,J.E., Harris,T., Brunk,B.P., Brestelli,J., Fischer,S., Harb,O.S., Kissinger,J.C., Li,W., Nayak,V., Pinney,D.F. *et al.* (2012) FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Res.*, **40**, D675–D681.
9. Minot,S., Melo,M.B., Li,F., Lu,D., Niedelman,W., Levine,S.S. and Saeij,J.P. (2012) Admixture and recombination among *Toxoplasma gondii* lineages explain global genome diversity. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 13458–13463.
10. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.
11. Liu,B., Madduri,R.K., Sotomayor,B., Chard,K., Lacinski,L., Dave,U.J., Li,J., Liu,C. and Foster,I.T. (2014) Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *J. Biomed. Inform.*, **49**, 119–133.
12. Petersen,T.N., Brunak,S., von Heijne,G. and Nielsen,H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
13. Baum,E., Sattabongkot,J., Sirichaisinthop,J., Kiattibutr,K., Davies,D.H., Jain,A., Lo,E., Lee,M.C., Randall,A.Z., Molina,D.M. *et al.* (2015) Submicroscopic and asymptomatic *Plasmodium falciparum* and *Plasmodium vivax* infections are common in western Thailand—molecular and serological evidence. *Malar. J.*, **14**, 95.
14. Crompton,P.D., Kayala,M.A., Traore,B., Kayentao,K., Ongoiba,A., Weiss,G.E., Molina,D.M., Burk,C.R., Waisberg,M., Jasinskas,A. *et al.* (2010) A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6958–6963.
15. Dent,A.E., Nakajima,R., Liang,L., Baum,E., Moormann,A.M., Sumba,P.O., Vulule,J., Babineau,D., Randall,A., Davies,D.H. *et al.* (2015) *Plasmodium falciparum* protein microarray antibody profiles correlate with protection from symptomatic malaria in Kenya. *J. Infect. Dis.*, **212**, 1429–1438.
16. Kamya,M.R., Arinaitwe,E., Wanzira,H., Katureebe,A., Barusya,C., Kigozi,S.P., Kilama,M., Tatem,A.J., Rosenthal,P.J., Drakeley,C. *et al.* (2015) Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. *Am. J. Trop. Med. Hyg.*, **92**, 903–912.
17. Gutierrez,J.B., Harb,O.S., Zheng,J., Tisch,D.J., Charlebois,E.D., Stoeckert,C.J. Jr and Sullivan,S.A. (2015) A framework for global collaborative data management for malaria research. *Am. J. Trop. Med. Hyg.*, **93**, 124–132.
18. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
19. Doyle,M.A., MacRae,J.I., De Souza,D.P., Saunders,E.C., McConville,M.J. and Likic,V.A. (2009) LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC Syst. Biol.*, **3**, 57.
20. Saunders,E.C., MacRae,J.I., Naderer,T., Ng,M., McConville,M.J. and Likic,V.A. (2012) LeishCyc: a guide to building a metabolic pathway database and visualization of metabolomic data. *Methods Mol. Biol.*, **881**, 505–529.
21. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
22. Shameer,S., Logan-Klumpler,F.J., Vinson,F., Cottret,L., Merlet,B., Achcar,F., Boshart,M., Berriman,M., Breitling,R., Bringaud,F. *et al.* (2015) TrypanoCyc: a community-led biochemical pathways database for *Trypanosoma brucei*. *Nucleic Acids Res.*, **43**, D637–D644.
23. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
24. Sidik,S.M., Huet,D., Ganesan,S.M., Huynh,M.H., Wang,T., Nasamu,A.S., Thiru,P., Saeij,J.P., Carruthers,V.B., Niles,J.C. *et al.* (2016) A GENOME-wide CRISPR screen in *toxoplasma* identifies essential apicomplexan genes. *Cell*, **166**, 1423–1435.
25. Chen,X., Wei,S., Ji,Y., Guo,X. and Yang,F. (2015) Quantitative proteomics using SILAC: principles, applications, and developments. *Proteomics*, **15**, 3175–3192.
26. Gunasekera,K., Wuthrich,D., Braga-Lagache,S., Heller,M. and Ochsenreiter,T. (2012) Proteome remodelling during development from blood to insect-form *Trypanosoma brucei* quantified by SILAC and mass spectrometry. *BMC Genomics*, **13**, 556.
27. Rogers,M.B., Hilley,J.D., Dickens,N.J., Wilkes,J., Bates,P.A., Depledge,D.P., Harris,D., Her,Y., Herzyk,P., Imamura,H. *et al.* (2011) Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res.*, **21**, 2129–2142.
28. Caro,F., Ahyong,V., Betegon,M. and DeRisi,J.L. (2014) Genome-wide regulatory dynamics of translation in the *Plasmodium falciparum* asexual blood stages. *Elife*, **3**, doi:10.7554/eLife.04106.
29. Bunnik,E.M., Chung,D.W., Hamilton,M., Ponts,N., Saraf,A., Prudhomme,J., Florens,L. and Le Roch,K.G. (2013) Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biol.*, **14**, R128.
30. Jensen,B.C., Ramasamy,G., Vasconcelos,E.J., Ingolia,N.T., Myler,P.J. and Parsons,M. (2014) Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*. *BMC Genomics*, **15**, 911.
31. Ison,J., Kalas,M., Jonassen,I., Bolser,D., Uludag,M., McWilliam,H., Malone,J., Lopez,R., Pettifer,S. and Rice,P. (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325–1332.
32. 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.
33. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
34. Wu,T.D., Reeder,J., Lawrence,M., Becker,G. and Brauer,M.J. (2016) GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods Mol. Biol.*, **1418**, 283–334.
35. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.