Published in final edited form as:

Methods Mol Biol. 2018 January 01; 1757: 69–113. doi:10.1007/978-1-4939-7737-6_5.

EuPathDB: the eukaryotic pathogen genomics database resource

Susanne Warrenfeltz¹, Evelina Y Basenko², Kathryn Crouch³, Omar S. Harb⁴, Jessica C. Kissinger^{1,5,6}, David S. Roos⁴, Achchuthan Shanmugasundram², Fatima Silva-Franco²

¹Center for Tropical & Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA

²Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

³Wellcome Trust Centre for Molecular Parasitology, University of Glasgow, Glasgow G12 8TA, UK

⁴Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

⁵Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

⁶Department of Genetics, University of Georgia, Athens, GA 30602, USA

Abstract

Fighting infections and developing novel drugs and vaccines requires advanced knowledge of pathogen's biology. Readily accessible genomic, functional genomic, and population data aids biological and translational discovery. The Eukaryotic Pathogen Database Resources (<http://eupathdb.org>) are data mining resources that support hypothesis driven research by facilitating the discovery of meaningful biological relationships from large volumes of data. The resource encompasses 13 sites that support over 170 species including pathogenic protists, oomycetes and fungi as well as evolutionarily related non-pathogenic species. EuPathDB integrates pre-analyzed data with advanced search capabilities, data visualization, analysis tools and a comprehensive record system in a graphical interface that does not require prior computational skills. This chapter describes guiding concepts common across EuPathDB sites and illustrates the powerful data mining capabilities of some of the available tools and features.

Keywords

Bioinformatics; parasite; pathogen; genomics; transcriptomics; orthology; fungi; proteomics; sequence analysis

1 Introduction

The Eukaryotic Pathogen Database [1,2] (EuPathDB, <http://eupathdb.org>) brings the power of bioinformatics to the scientific community by integrating pre-analyzed omics data with advanced search capabilities, data visualization and analysis tools that facilitate the discovery of meaningful biological relationships from large volumes of data. EuPathDB provides a sophisticated data mining platform for biologists with no prior computational training to explore omics data in support of their hypothesis driven research.

The resource is organized into 13 sites (Table 1) and supports over 170 eukaryotic parasites, relevant free-living non-parasitic organisms and selected pathogen hosts. EuPathDB resources provide customized sites for accessing genomes and functional data based on taxonomic groupings, host data collected during infection, evolutionary relationships based on OrthoMCL clustering, and a portal site that enables queries across all data in EuPathDB. The sites are built on the same web architecture and use the same common vocabulary and organizing logic, which eases the transfer between the search, visualization and analysis sections and allows users to move between sites without re-education.

EuPathDB integrates a wide range of data from many sources and repositories (Table 2). The breadth of data broadens the data mining capabilities by providing multiple forms of experimental evidence to search, visualize and analyze. As the data are integrated, they are analyzed with standard workflows, ensuring that data from different sources can be compared. An in-house analysis pipeline also creates orthology profiles across all genomes so that comparisons can be made across organisms.

Data mining in EuPathDB sites can take four general paths. First, record pages compile all available data concerning a feature (e.g. gene, SNP, pathway, compound, EST, genomic sequence etc.) and offer rich data mining opportunities. Second, the search strategy system's unique infrastructure facilitates the exploration of relationships across data sets, data types and organisms to produce a refined set of features that share biological characteristics of interest. Third, visualization tools such as the Genome Browser (GBrowse) [3] coupled to EuPathDB's breadth of sequence-based data offer the ability to view different data types in your genomic area of interest. And fourth, tools such as enrichment analyses and a private Galaxy workspace for primary data analyses enhance data mining.

EuPathDB makes it easy to interrogate biological questions relating to issues such as stage-specific expression, gene model integrity or alternative splice variants, and to compile lists of genes that share multiple biological characteristics (e.g. kinases secreted at a particular time, where they may affect host responses). This chapter describes the structure and utility of EuPathDB and illustrates some of the available tools and methods. Since EuPathDB sites are built using the same infrastructure and user interface, the steps described herein can be applied to any EuPathDB site.

2 Using EuPathDB Sites

The exercises below offer example data mining strategies. Because new versions of EuPathDB resources new are released about every two months and may contain new annotation and functional data, reader results (gene numbers, etc) may vary slightly from that published here.

2.1 Home Pages

EuPathDB home pages are organized using the same web architecture and provide users with easy access to all the searches, tools, educational material and helpful database and community information.

2.1.1 Anatomy of EuPathDB home pages

1. Visit the EuPathDB home page (<http://eupathdb.org>) and explore the four sections: the header (Fig. 1A), component link-out section (Fig. 1B), the searches and tools section (Fig. 1C), and the side bar (Fig. 1D).
2. The header (Fig. 1A) is available from all pages and includes a gray menu bar that offers dropdown menus or direct links for accessing most searches, tools, data set information, bulk downloads, and the Galaxy workspace. Above the gray menu bar are two search boxes (Fig. 1A arrow) for quick access to gene record pages or to a text search that returns genes whose records contain the term(s) of interest. Directly below the search boxes are links to 'Login', 'Register' or 'Contact Us'. Although not required for accessing data records and building search strategies, registering provides access to additional tools and functionality such as the ability to save and share search strategies, to add genes to the 'My Basket' and 'My Favorites, to add comments on gene records and the Galaxy workspace. The My Strategies page is an important part of the sites, serving as a workspace for creating strategies and viewing search and strategy results. The 'Contact Us' link opens a form for sending questions, comments and suggestions to our email support line.
3. The component link-out section (Fig. 1B) is available on the EuPathDB homepage and at the bottom of all pages of all sites. This section offers direct links to the taxon-specific sites as well as OrthoMCL DB. Click the icons to navigate to the site of your choice.
4. The searches and tools section contains three panels for accessing searches and tools (Fig. 1C). Searches listed under 'Search for Genes' (Fig. 1C, green arrow) return only genes while searches that return non-gene entities such as SNPs, isolates, or ESTs are available from 'Search for Other Data Types' (Fig. 1C, blue arrow). The searches are organized into categories that can be expanded to reveal individual searches. Alternatively, searches can be filtered using the 'Find a Search' tool. For example, typing 'signal' in the 'Find a Search' box of the 'Search for Genes' panel filters the searches and reveals the 'Predicted Signal Peptide' search within the category 'Protein targeting and localization' (Fig. 1E).

Also available from the home page are tools for BLAST, Results Analysis, Sequence Retrieval, Genome Browser, Companion annotation pipeline [4] and EuPaGDT (Eukaryotic Pathogen CRISPR guide RNA Design Tool) [5] (Fig. 1C, red arrow). The Results Analysis tool enables functional enrichment of output gene lists from the search strategies (See the Data Analysis section for further details).
5. The side bar (Fig. 1D) contains expandable sections for data summary, release notes, Twitter feed, community resources, links to workshop material, tutorials and help. Newly added items for these sections are highlighted in yellow.

2.1.2 Gene ID and Gene Text search access from home pages—There are two ways to access the Gene ID and Gene Text searches: through the search boxes in the header

(Fig. 1A, arrow) and through the dedicated search pages categorized in the ‘Search for Genes’ panel on the home page (Fig. 1C green arrow). Entering a gene ID in the header ‘Gene ID’ box navigates directly to the record page for that gene. Entering a text term or phrase (within quotation marks) in the header ‘Gene Text Search’ box initiates a pre-configured search for genes whose records contain the text term or phrase. The dedicated search pages offer additional options. The Gene ID search page, accessed under the ‘Annotation, curation and identifiers’ category (Fig. 1C orange arrow), allows a user to search for gene IDs in bulk. A list of gene IDs can be pasted into the text box, uploaded from a file, or converted from a user’s basket. The gene text search page can be found in the ‘Text’ category (first category in the list) and allows a user to configure the sections of the gene record that they want to search. For example, the text search can be limited to search only the product description of genes. Both searches support a wild card to perform partial text or ID searches. For example, a test search of the term “phospho*” (the asterisk * is the wild card) will return any gene whose record contains any word with the prefix “phospho”.

1. *Use the ‘Gene Text Search’ in the EuPathDB home page header to find genes that are likely proteases.* Enter the term ‘protease’ (without the quotes) in the search box (Fig. 1A) and click on the search icon to the right of the box to initiate a query against all annotated genomes for genes whose records include the term protease. The results (>20,000 genes, results may vary in subsequent database releases) appear in the ‘My Strategies’ section (Fig. 2) and consist of the strategy panel with a graphic representation of the strategy (Fig. 2A), a component website filter which displays the distribution of genes across the taxon-specific sites (Fig. 2B upper table), an organism table which displays the distribution of genes for the genomes that were queried (ranging from 0 to >400 genes per species) (Fig. 2B, lower table), and the Gene Results consisting of two tabs. The ‘Gene Results’ tab lists gene IDs and associated data for genes returned by the search (Fig. 2C, showing). The ‘Genome View’ tab (Fig. 2C, black arrow) presents a graphic representation of the genomic sequences ‘painted’ with the gene results when there are less than 10,000 genes in the result. The Analyze Results button (Fig. 2C, blue arrow) opens a tool offering enrichment and other analyses of the gene result.
2. *Explore your result.* The ‘Gene Results’ table contains columns of data associated with the genes that were returned by your search. You can add columns to the table using the ‘Add Columns’ button (Fig. 2C, green arrow.) Look at the product description column. Cathepsin B precursor, GL50803_10217, is returned by the search but does not have the term protease in the product description. In this case, the term protease was found in an InterPro domain and a user comment associated with the gene.
3. *Find several genes using the Gene ID search.* Navigate to the EuPathDB home page by clicking the Home button, the first tab in the header’s gray menu bar. Open the Gene ID search page by first clicking on the category ‘Annotation, curation, identifiers’ then clicking on ‘Gene ID(s)’ in the ‘Search for Genes’ panel. On the next page, paste the following list of IDs in the search box and click on the ‘Get Answer’ button:

TGME49_049180, TA08775, PFD0830w, PCHAS_072830, PBANKA_071930, NCU10053T0, NCLIV_065390, LmxM.06.0860, ECU01_1430, CMU_010300,

The results are displayed as a search strategy including all the genes from the above list in one step.

4. *Notice the gene results.* The filter table contains hits from several different species. Examine the Product description column. These genes are orthologs of dihydrofolate reductase-thymidylate synthase. Try running the same search in PlasmoDB.org (<http://plasmodb.org>). Since PlasmoDB accesses a reduced taxonomic group of genomes while EuPathDB access all genomes and data, only the plasmodium orthologs of dihydrofolate reductase-thymidylate synthase are returned by the PlasmoDB search.

2.2 Exploring Record Pages

Record pages compile all available data for an entity, including genes, SNPs, ESTs, isolates, pathways, compounds, genomic sequences, genomic segments and ORFs. The following 2 examples describe the features, navigation and data content of gene and metabolic pathway record pages.

2.2.1 Gene Record pages

1. *Anatomy of the gene page:* Visit PlasmoDB (<http://plasmodb.org>) and enter the gene ID for apical membrane antigen 1, PF3D7_1133400, in the 'Gene ID' box in the header. Designed for easy navigation and access to data of interest, gene record pages contain three major areas, the summary (Fig. 3A and B), the data section (Fig. 3C) and the content navigation (Fig. 3D). The summary provides basic information about the gene (Fig. 3A). The 'Shortcuts' that appear in the summary (Fig. 3B) serve two functions: clicking on the magnifying glass icon at the bottom right corner of the thumbnail provides a graphic summary of that data type (Fig. 4A green arrow); clicking on the image itself, or the title above it, will navigate to that section of the page (Fig. 4A, blue ovals and dashed lines). Several gene page sections contain a 'View in genome browser' link which opens the genome browser with the pertinent data tracks open (Fig. 4B, blue arrow) (see Subheading 2.4 on data visualization). The 'Add to basket' and 'Add to favorites' links (Fig. 3A, arrow) will save or bookmark the gene for later use. The 'Download Gene' link opens the download tool where the FASTA formatted sequence or all information on the gene page can be downloaded (Fig. 5). The content navigation section on the left side of the gene page (Fig. 3D) serves as a configurable table of contents of all information found in the data section and remains available on the left side of the page as you scroll down. The data section contains all information available for the gene of interest (Fig. 3C). Table 3 describes the data available on the gene page. The data are presented both in graphs and in searchable tables, and sections can be collapsed or expanded using the triangle present in the title of each section (Fig. 3C, black arrow).

2. *Transcriptomics section:* Use the 'Contents' navigation menu (Fig. 3D) to navigate to the transcriptomics section of the PF3D7_1133400 gene page by clicking on the section title 'Transcriptomics'. Alternatively, use the 'Search Section Names' tool at the top of the Contents navigation menu to search for the Transcriptomics section. The transcriptomics table (Fig. 6) appears in the data section of the gene page and contains collapsible rows for each dataset. Scroll down and click the triangle present in the header of experiments 'Polysomal and steady-state asexual stage transcriptomes' (Fig. 6, blue circle) [6] and 'Transcriptomes of 7 sexual and asexual life stages' [7]. The rows expand to reveal expression graphs, data tables and coverage plots relative to the data set. Explore the graphs and data tables for these two experiments. At what life cycle stage is the expression highest for Pf3D7_1133400? (answer = schizont stage)
3. *Proteomics section:* To navigate to the proteomics section you can easily use the contents navigation menu on the left side of the page and click on 'Proteomics' or use the 'Back to top' arrow to return to the summary section and click on the Proteomics shortcut image. The 'Mass Spec.-based Expression Evidence Graphic' (Fig. 7A) contains a summary table with a row for each transcript that includes an image of all mapped peptides from each proteomics data set. The mouse can be used to hover over the glyphs representing the mapped peptides to obtain details about the peptide (experiment and sample names, sequence etc.) (Fig. 7B). While there is abundant proteomics data, three experiments in particular support expression at the schizont stage– 'Schizont Phosphoproteome (3D7)(2012)' [8], 'Schizont Phosphoproteome (3D7)(2011)' [9], and 'Cytoplasmic and nuclear fractions from rings, trophozoites and schizonts (3D7)' [10]. Each of these has mapped peptides from schizont samples. (Fig. 7A blue arrows)
4. *Annotation, curation and identifiers section:* EuPathDB encourages the community to enhance annotations by providing a platform to add comments to the record pages (Fig. 8). The comment system links knowledge from community experts to gene and other records. Once a user comment is added, it appears immediately on the gene page and becomes searchable through the text search. Some genomes are professionally curated by EuPathDB staff. When appropriate, user comments are integrated into the official annotation for these genomes.

Navigate to this section using the contents navigation menu on the left. This section contains useful information including previous identifiers, gene synonyms, annotation notes and user comments. Note that PF3D7_1133400 has two user comments (Fig. 8A) that are summarized in a table (Fig. 8B). Each comment can be explored further by clicking on the comment ID (Fig. 8B, green arrow). To add a new comment, click the 'add a comment' link (Fig. 8B, blue highlighting) and complete the form (Fig. 8C). You must be registered and logged in to add a comment. Table 4 gives examples of information to include in a comment.

5. *Orthology and Synteny section:* Explore orthology for the Cyclin-like F box protein 1A in *Trypanosoma brucei* strain TREU 927 (Tb927) in TriTrypDB. Navigate to TriTrypDB (<http://tritrypdb.org>) and enter Tb927.1.4540 into the Gene ID search box in the header (Fig. 9A, arrow). Use the Contents navigation menu to navigate to the Orthology and Synteny section (Fig. 9B, blue box). The ortholog group ID, OG5_132982 (Fig. 9B, green arrow), to which this gene has been assigned appears as a link to the OrthoMCL database where one can explore the group's features and distribution across a wider range of taxa.

Several interesting things about this gene can be discovered from this table. Based on the gene IDs of the table's entries for Tb927 genes, four paralogs of the Tb927.1.4540 gene can be found on chromosome 1 of Tb927 (i.e. Tb927.1.4560 = organism.chromosome.gene number) (Fig. 9B, green box). The close proximity of these genes (clustered gene numbers in the IDs) suggests that these paralogs may have arisen as a result of tandem duplication. Another paralog is found on chromosome 11 of Tb927 (Tb11.v5.0705) (i.e. Tb11.v5.0705 = organism & chromosome.genome version.gene number) which may have arisen separately.

Close the Orthologs and Paralogs table by clicking the triangle next to the title (Fig. 9B, black arrow). With the table closed, we can see the 'Retrieve multiple sequence alignment or multi-FASTA' tool that can be used to conduct a multi-sequence alignment (MSA) using ClustalW across up to 15 organisms from the current database, with outputs in either ClustalW or multi-FASTA format. To use the alignment tool, choose 15 or fewer organisms from the tree in the center of the tool or search for your organisms of choice with the search box. Select an output format and click Submit Query. The results will appear in a separate browser window.

Close the Alignment tool by clicking the triangle next to the title or scroll down to the Synteny graphic that is centered on the Tb927.1.4540. This graph displays output from a Mercator [11] analysis that maps larger regions of orthology across all loaded genomes. The structure of orthologous genomic segments is often conserved, containing similar sets of genes in a similar order. In the graph, synteny is indicated with gray shadowing.

Notice the structure and order of genes in the Synteny graph (Fig. 9C) and hover over the gene glyphs to reveal gene details (Fig. 9D). The parent gene (Tb927.1.4540) (Fig. 9C, red box) is followed downstream by several paralogs (Fig. 9C, red arrows) that we considered while looking at the orthology table. Notice the presence of multiple paralogs in *T. brucei*, *T. evansi* and *T. congolense*, and the absence of orthologs of the gene in the wider orthologous (syntenic) region of the genome in *Leishmania*, *Endotrypanum* and *Crithidia*. Interestingly, only some of the genes shown in the table for *T. vivax* and none of the genes shown in the table for *T. rangeli* or *T. grayi* are shown in this graphic. The reason for this is that while these genes are orthologous with Tb927.1.4540, they do not sit on a region of the genome that shares wider orthology (synteny)

with the region around this gene in Tb927. This property is also described for each gene in the 'Orthologs and Paralogs within TriTrypDB' table with the column headed 'Is Syntenic'. Links are provided to open this image in the Genome Browser where one can customize the organisms shown, zoom in and out and add other tracks.

2.2.2 Metabolic Pathway Record pages—Metabolic pathways from KEGG [12–14], MetaCyc [15], TrypanoCyc [16] or LeishCyc [17,18] are loaded in EuPathDB sites and mapped to genes that are annotated with appropriate enzyme commission (EC) numbers. Pathway record pages integrate these networks with annotations, gene expression profiles and orthology data via the Cytoscape [19–21] platform. Metabolic pathways can be retrieved based on several criteria including compounds (substrates or reactants), gene lists, pathway identifiers or names (Fig. 10A).

1. Navigate to the TriTrypDB (<http://tritrypdb.org>) home page and find the Glycolysis 1 (TrypanoCyc) metabolic pathway: click on the 'Metabolic Pathways' category to expand its contents, then select the 'Pathway Name/ID' search (Fig. 10A). Begin typing the pathway name (Glycolysis I: GLYCOLYSIS-1 TrypanoCyc) in the 'Pathway Name or ID' parameter and then choose the correct pathway from the list that appears (Fig. 10B). Notice that pathway names may appear more than once since they are obtained from multiple sources. Fig. 10C-E show portions of the Glycolysis 1 pathway cycle in *Trypanosoma brucei* as annotated in TrypanoCyc and represented in TriTrypDB.
2. *Explore the pathway.* The organization of the record page is similar to the gene page with a summary at the top, a Contents section for navigation and a data section with tables and images. The interactive pathway image depicts the series of enzymatic reactions as enzyme and compound nodes (Fig. 10C, 1 and 2, respectively) with byproducts shown in gray. Small adjustments to the pathway layout can be made by dragging nodes and byproducts to new locations. Panning and zooming the view can be achieved with the tools in the top left corner (Fig. 10C, 3), or by clicking and dragging a node or side product to pan or scrolling to zoom. Enzymes that catalyze reactions are displayed in rectangular boxes (Fig. 10C, 1), labelled with an enzyme commission (EC) number if known, and a name or reaction identifier if the EC number is not known. A red outline denotes that at least one gene encoding an enzyme with this EC number is present in the current component database. Compounds are identified using ChEBI identifiers. Where available, the compound structure is shown for primary metabolites (Fig. 10C, 2), whereas side compounds are represented as text. Clicking on any node displays a 'Node Details' pop-up (Fig. 10D) which includes links to genes annotated with the EC number for enzymes and a link to the compound record page for compounds.
3. *Annotate the pathway.* Annotate the pathway with expression data that explores differential expression between procyclic (insect) and bloodstream forms of *T. brucei* (Fig. 10E). Choose 'Paint Enzymes' and 'By Experiment' to pull up a list of all experimental data that can be 'painted' on the enzyme nodes. Choose '*T.*

brucei *brucei* TREU927 Bloodstream and Procyclic Form Transcriptomes (Siegel et al.) [22] and then ‘Paint’ to display the expression data in the enzyme nodes. Examine the experimental data in the nodes and notice that several of the enzymes in this pathway are downregulated in procyclic forms compared to bloodstream forms. This can be interpreted as an indication of differential sugar metabolism between the two lifecycle stages, which makes sense given the very different environments they inhabit.

4. Navigate to the 5-aminoimidazole ribonucleotide biosynthesis I (PWY-6121) pathway as above (Metabolic Pathways section 1) and explore the tool for annotating the pathway with the distribution of genes across phylogeny. Choose ‘Paint Enzymes’ and ‘By Genera’. From the ‘Genera Selector’ choose Kinetoplastida and Mammalia and then click ‘Paint’. Each enzyme is replaced with a chart showing whether a gene encoding the enzyme is present in three genera from the Kinetoplastida (*Crithidia*, *Leishmania*, *Trypanosoma*), and two genera from the Mammalia (*Homo*, *Mus*, blue). Click on a node to see a larger image of the distribution. Notice that all the enzymes from this pathway are encoded in Mammalia, but none of these enzymes are encoded in any of the represented Kinetoplastida. This pathway is a part of the super pathway involved in *de novo* purine biosynthesis and this representation agrees with the observation that *Trypanosoma* cannot synthesize purines *de novo* but instead rely on scavenging from the host. It can be inferred that this is also true of other Kinetoplastida.

2.3 Data mining with searches and strategies

EuPathDB offers over 100 pre-configured searches in a unique and powerful strategy system that allows you to explore relationships across data sets, data types and organisms. Searches query individual data sets that provide evidence for a specific biological property and return a list of records that meet the search criteria and therefore have the biological characteristic defined by the data set. Strategies (Fig. 11A) can be created by adding, subtracting, joining, intersecting or collocating (Fig. 11B) the results of subsequent searches. The colocation tool is used to explore relationships based on relative genomic location, such as interrogating SNPs located 500nt upstream of genes. A nesting tool allows you to control the logic when combining search results. Results from any step in a strategy can be analyzed using gene ontology (GO) [23,24] enrichment, pathway enrichment or genome visualization tools. The following two examples illustrate how to create strategies and leverage orthology in the EuPathDB strategy system.

2.3.1 Strategy Example 1—This example creates a strategy (Fig. 11A) in PlasmoDB (<http://PlasmoDB.org>) that finds *Plasmodium vivax* proteases that are likely expressed during the gametocyte stage. The strategy employs three searches and uses the Transform by Orthology tool to convert *P. falciparum* genes into their *P. vivax* orthologs. Steps 1 and 2 return proteases using two different lines of evidence – a text search in step 1 and a GO term search in step 2. These searches are combined with a union to obtain a more comprehensive list of possible proteases. Step 3 returns genes with evidence for expression during the

gametocyte stages based on *P. falciparum* RNA sequencing data [25]. Steps 2 and 3 are combined using the intersect operator to produce a list of genes that have both biological properties: these genes are suspected proteases with evidence for expression during gametocyte stages. The *P. falciparum* genes from step 3 are transformed into their *P. vivax* orthologs with the Transform by Orthology tool to produce a list of *P. vivax* genes that are likely proteases expressed in the gametocyte stage. This transformation exploits orthologous clustering of EuPathDB organisms to infer functional characteristics determined in *P. falciparum* to *P. vivax*. The following offers detailed instructions for building the strategy. The completed strategy is also available here: <http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf>.

1. *Find genes that are possible proteases using the text search to query gene records for the term 'protease'* (Fig. 12). To reach the search, click on the 'Text' category link on the home page 'Search for Genes' menu (Fig. 12A). Next click on the 'Text (product name, notes, etc.)' link to open the 'Text Search: Identify Genes by Text (product name, notes, etc.)' page (Fig. 12B). Each search is loaded with default parameters that can be configured before running the search. The default setting for the 'Organism' parameter is set to search all organisms in the database while the default setting for the 'Fields' parameter will query every field but 'Similar proteins (BLAST hits v. NRDB/PDB)'. Type the word 'protease' (without the quotes) in the 'Text term (use * as wildcard)' box (Fig. 12B, arrow) and click 'Get Answer' to initiate the search. The search results (Fig. 12C) are displayed in the 'My Strategies' section which consists of a strategy panel with an interactive image of the strategy, followed by an organism filter showing the distribution of hits across the genomes queried, and a result table with the list of genes returned by the search. The first column in the result table is the gene ID and serves as a link to the gene record. Searches and strategies can be saved (Fig. 12C, blue bordered inset) and are given a unique URL that can be used to share the strategy with colleagues.
2. *Expand the list of proteases with a second line of evidence for proteolytic activity.* There may be some proteases that do not have the term 'protease' in their record but do have an assigned GO annotation associated with proteolysis. The ontologies are a controlled vocabulary for describing the molecular function, biological process or subcellular location of a gene product. GO annotations in PlasmoDB were either provided by the sequencing and annotation centers or inferred based on a gene product's similarity to protein domains from the InterPro databases [26].

To add a GO term step to the search strategy, click on the red 'Add Step' button in the strategy panel (Fig. 13A) that opens the 'Add Step' popup (Fig. 13B). Next, navigate to the GO Term search page, by clicking on 'Run a new Search for', 'Genes', 'Function Prediction', and 'GO Term'. Specify the GO Term or GO ID by typing either the GO Term (proteolysis) or ID (GO:0006508) and then choosing the correct term from the list that appears (Fig. 13C, black arrow). Since this is not the first search in the strategy, running this search requires defining how to combine the results of this search with the previous one. Choose

to union the two ID lists to add genes discovered in the GO term search that are not already in the list of possible proteases (Fig. 13C, blue arrow). See Fig. 11B for more information about combining searches. Click 'Run Step' to initiate the search. The resulting strategy (Fig. 13D) contains two steps and returns over 2,500 genes whose products are likely to have proteolytic activity based on two lines of evidence, the word protease found in their gene records and/or a GO term assignment of GO:0006508 proteolysis.

3. *Filter by genes highly expressed at the gametocyte stage.* To ensure that our list of proteases is highly expressed at the gametocyte stage, we can intersect the Step 2 results with a search for genes based on transcript expression in the gametocyte stage. Click 'Add Step' from the strategy panel and navigate the 'Add Step' panel through 'Run a new Search for', 'Genes', 'Transcriptomics', 'RNA Seq Evidence'. A list of available RNA sequencing data sets and their associated searches will appear (Fig. 14A). Notice that there are no gametocyte RNA seq datasets associated with *P. vivax*, so we will search the *P. falciparum* dataset in this step and then transform the results into their *P. vivax* orthologs.
4. Choose the Percentile search (P) for 'Female and Male Gametocyte Transcriptomes (Lasonder et al.)' (Fig. 14A, blue borders) to open the search page. The data set contains an RNA sequencing analysis of *P. falciparum* male and female gametocyte samples from a study published in 2016 [25]. To create the Percentile search, EuPathDB obtained the raw sequencing reads, applied a standard RNA-seq mapping workflow, ranked expression values from highest to lowest, and then grouped genes into percentile groups in each sample. Running the percentile search using the default 'max/min expression percentile' parameters will return the genes whose expression levels are in the top 20% for the samples chosen in the 'Samples' parameter.

Choose both samples, male gametocyte and female gametocyte, from the search page (Fig. 14B, blue arrow). Since the goal is to create a list of genes that are proteases and expressed in gametocytes, choose to intersect the Percentile search with the Step 2 results. Clicking Run Step will initiate the RNA Seq Evidence search, intersect the new results with the Step 2 results and return a Step 3 result that includes genes that are in both result sets. The genes returned in Step 3 result (Fig. 14C) will therefore possess biological properties of all data sets searched, possible proteolytic activity and high gametocyte expression.

Notice the genes in the Step 3 result are only *P. falciparum* genes. This is evident in the organism filter table which shows over 60 genes under *P. falciparum* 3D7 but none in other organisms (Fig. 14D). This is because the RNA sequencing experiment was performed in *P. falciparum*.

5. *Use the 'Transform by Orthology' tool to transform the P. falciparum gametocyte proteases to their P. vivax orthologs.* Since gametocyte expression data is unavailable for *P. vivax*, this step of the strategy takes advantage of data obtained in *P. falciparum* to generate a list of *P. vivax* genes that are likely expressed in *P. vivax*. Click the red 'Add Step' button following Step 3 and then choose

‘Transform by Orthology’ in the first column of the popup (Fig. 15A). Arrange the ‘Organism’ parameter to include only *P. vivax* Sal1 and leave the ‘Syntenic Orthologs Only’ parameter in the default ‘No’ setting (Fig. 15B). The *P. vivax* Sal1 genes returned by the search will be orthologs of the *P. falciparum* input genes. The resulting four-step strategy returns *P. vivax* genes that are likely proteases expressed in gametocytes (Fig. 15C).

6. *Explore your results!* It is important to perform a critical review of strategy results to determine whether your final gene list truly possess the biological characteristics intended by your search strategy. For example, one quick check can be performed by reviewing the ‘Product Description’ for terms associated with proteolysis.

2.3.2 Example 2: Searching by Orthology and Phyletic Profile—A common use for orthology data in data mining is to find genes that are restricted to particular taxa. For example, vaccine candidates or druggable targets ideally would be proteins that are unique to the pathogen and are not found in the host. Orthology might also be useful for finding genes that are related to a process or organelle. Apicomplexans have a unique four-membraned organelle, the apicoplast, which is thought to have arisen through two endosymbiotic events. This organelle, and the products of the genes that act there, are tempting drug targets, making it important to identify genes that act in the apicoplast.

In this example, we will begin the strategy with a search for *P. falciparum* genes that are likely expressed in the apicoplast. *Plasmodium* harbors a motif that targets proteins to the apicoplast and there is a search in PlasmoDB that returns genes encoding this motif. The search is also available in EuPathDB where the *P. falciparum* results can be transformed to orthologs in other species. In Step 2, the *P. falciparum* apicoplast genes will be transformed to their orthologs in *Toxoplasma gondii* strains ME49 and GT1 and the closely related *Neospora caninum*. In Step 3, we will use the Orthology and Phylogenetic Profile search to restrict the list of *T. gondii* and *N. caninum* ‘apicoplast’ genes to those that do not have orthologs in human or *Cryptosporidium*. Since *Cryptosporidium* has lost its apicoplast, any genes in the list of *T. gondii* and *N. caninum* ‘apicoplast’ genes that have orthologs in *Cryptosporidium* are less likely to be apicoplast-specific. We will also remove genes that have orthologs in human since the best parasite druggable targets or vaccine candidates would be genes and proteins that are not present in humans to avoid interactions and side effects. The completed strategy is also available here: <http://eupathdb.org/eupathdb/im.do?s=3353bf3401d62d48>.

1. Navigate to EuPathDB (<http://eupathdb.org>) to begin the strategy with a search for *P. falciparum* genes containing a motif that targets proteins to the apicoplast. Note that this search must be carried out in EuPathDB in order to perform orthology transforms between organisms that are hosted in different component sites. Use the ‘Search for Genes’ panel (Fig. 16A) or the header drop down menu to view the category ‘Protein targeting and localization’ and open the ‘P.f. Subcellular Localization’ search page (Fig. 16B). Choose ‘Apicoplast’ for the

‘Localization’ parameter and click ‘Get Answer’. The results appear as Step 1 in the strategy panel (Fig. 16C).

2. Transform the *P. falciparum* apicoplast genes to their *T. gondii* ME49, *T. gondii* GT1 and *N. caninum* Liverpool orthologs. Click ‘Add Step’ from the strategy panel and choose ‘Transform by Orthology’. Arrange the ‘Organism’ parameter of the transform tool to include only the three organisms of interest (Fig. 16D). The results of the ortholog transform appear as Step 2 in the strategy (Fig. 16E). While the majority of these genes will act in the apicoplast, some may have additional functions. This gene list can be refined using information from *Cryptosporidium*, a species of apicomplexan that is closely related to *Toxoplasma* and *Neospora* but which has lost its apicoplast. The results can be narrowed to include only genes that are likely to be truly apicoplast-specific by removing genes that have orthologs in *Cryptosporidium*. If the interest is in drug targets, the list can be further refined to exclude genes that have orthologs in vertebrates.
3. Click ‘Add Step’ and navigate the ‘Add Step’ panel through ‘Run a new search for’, ‘Genes’, ‘Orthology and synteny’, and choose ‘Orthology Phylogenetic Profile’ (Fig. 16F). The search opens displaying two parameters. The ‘Find genes in these organisms’ parameter allows selection of the organisms from which genes will be returned. Choose ‘clear all’ and then choose *T. gondii* ME49, *T. gondii* GT1 and *N. caninum* Liverpool from the Apicomplexa category. Use the ‘Select orthology profile’ (Fig. 17A) parameter to define the orthology profile of the genes returned by the search. Arrange green check marks for organisms in which orthologs must be present, and red for organisms in which orthologs cannot be present (red crosses). In this example, all *Cryptosporidium* and all Mammalia should be excluded from the ortholog profile of the genes returned by the search (Fig. 17A) while nothing is required to be included (no green check marks). Then choose to intersect the results of the ‘Orthology Phylogenetic Profile’ with the previous search results (Fig. 17A, arrow). The strategy produces a list of possible apicoplast genes in *T. gondii* ME49, *T. gondii* GT1 and *N. caninum* Liverpool based on *P. falciparum* data from an algorithm that predicts apicoplast targeting based on the presence of a motif.

2.4 Data mining with visualization - Visualization of Genomic Data with GBrowse

GBrowse is a highly configurable tool for visualization of sequence feature data at the genome-wide scale and is embedded into all EuPathDB database sites. In this section, we will examine a single gene using GBrowse to visualize data aligned to the genome in the region of the gene – TGME49_200320 hypoxanthine-xanthine-guanine phosphoribosyl transferase, HXGPRT. We will be able to interpret alternative splicing and gene model accuracy.

1. Navigate to ToxoDB (<http://toxodb.org>) and go to the HXGPRT gene page by entering the gene ID, TGME49_200320, in the Gene ID box in the header (Fig. 18A, blue arrow). Access GBrowse by clicking the ‘View in genome browser’ button from the ‘Gene models’ section (Fig. 18A, green arrow). The initial view

on the GBrowse page defaults to the gene region with a track displaying annotated transcripts colored by the direction of transcription as well as tracks for splice site junctions which provide evidence for intron/exon boundaries (Fig. 18B). Hover over the glyphs in the tracks to reveal details. The 'Landmark or Region' box shows the coordinates of the displayed region (Fig. 18B, 1). Entering alternative coordinates, a gene ID or a transcript ID into this box will bring the specified region into view. The 'Overview', 'Region' and 'Details' scales (Fig. 18B, 2) show the entire chromosome or contig, a zoomed view of the chromosome, and the selected region of the chromosome or contig, respectively. Displayed regions are highlighted in yellow along the scale.

Each track has a toolbar in the track header that can be used to hide (-), remove (x), share (radiowaves), configure (wrench) or access a track description (?) (Fig. 18B, 3). With the configure tool one can change the track dimensions, axes, glyph types and colors, etc. GBrowse layouts can be saved using the 'Save Snapshot' utility (Fig. 18B, blue box, login required), or a URL can be generated to share the track using the 'File' menu at the top of the page. Downloads of track images can also be obtained from this menu. Finally, personal data tracks can be made or uploaded in the 'Custom Tracks' tab.

2. Expand the region to 10 kbp using the dropdown menu in the panning and zooming tool (Fig. 18B, 4). Zooming and relocating can also be achieved through the Landmark tool or by using the mouse to highlight the region of interest in any of these three layers.
3. To display additional data tracks, click on the Select Tracks tab (Fig. 18B, arrow). A wide variety of data types are available to display, including gene models, splice site junctions, synteny, sequence variations, epigenetic datasets from ChIP-on-ChIP or ChIPseq, transcriptomics, proteomics and others. Multiple tracks can be selected, but data from different organisms cannot be displayed at the same time. Tracks are organized by data type according to the same common logic as searches on the home page and a search box (Fig. 18C, blue box) can be used to quickly find tracks of interest. Type 'Craig' in the search box and then choose the two tracks labeled 'Annotated Transcripts with CRAIG UTR Prediction' and 'CRAIG denovo Gene Model Prediction'. The tracks are automatically added to the display in the Browser tab. These two tracks are output from the CRAIG algorithm and provide alternative gene models.
4. Return to the Browser tab and compare the gene models between tracks. Note that the 3' UTR from the CRAIG model is longer than that in the official annotation.
5. Figure 19 shows a GBrowse view displaying 9 data tracks that provide evidence for interrogating alternative splicing in HXGPRT. Return to the Select Tracks tab and turn on the other tracks (Table 5) to create the display in Figure 19. The GBrowse view is also available at this URL: <http://tinyurl.com/m8d4qtp>.

In this display, tracks have been rearranged for convenience. This can be achieved by clicking the title bar of any track and dragging it up or down as required. The tracks labelled A in Figure 19 show the gene model from the official annotation (upper) and two splice site junction tracks that open by default when we accessed GBrowse from the gene page. Tracks labeled B are the gene models from the CRAIG *de novo* prediction tool, one of which is highlighted in yellow. Highlighting can be customized in the ‘Preferences’ tab. Tracks C-E show data from a subset of the RNA sequence datasets available in ToxoDB. The y-axis represents the number of reads aligned. Note that each of these tracks is only displaying a subset of the available subtracks. In track D, the displayed subtracks are overlaid rather than stacked as in track E. Subtracks can be selected, rearranged and overlaid in the track-specific subtracks menu by clicking on ‘Showing x of y subtracks’ (Fig. 19D, arrow). Tracks C and D both show reads aligned to the 3’ end of HGGPRT corresponding to the longer UTR predicted by CRAIG. Track D shows some evidence for transcription from within first intron, and it can be seen from track E that transcription from exon 3 is lower than other exons, suggesting that exon 3 is sometimes skipped. This agrees with reports of alternative splice forms in this gene. The splice junction track in A shows splice-crossing reads unified from all available RNA-seq datasets. The presence of reads that span exon 2 to exon 4 support the presence of the alternative splice form. Track F shows expressed sequence tag (EST) alignments. These also show evidence of transcripts in which exon 3 is skipped, and additionally lend support to the read-through of intron 1 observed in the RNA-seq data.

2.5 Data analysis

2.5.1 Result Analysis Tool: enrichment analysis of a strategy result—While EuPathDB’s sophisticated search strategy system creates biologically meaningful gene lists, the enrichment analyses aid interpretation by identifying over-represented biologically relevant labels such as GO Terms and metabolic pathways in a result set. EuPathDB offers enrichment analyses for GO Terms, metabolic pathways and words in the gene product description. The enrichment analyses perform a Fischer’s Exact Test comparing functional annotations assigned to genes in the result list with the all genes in the genome. In this method, the results of Example 1 will be analyzed for enriched GO Terms and metabolic pathways.

1. *Retrieve the result of the strategy in Example 1: P. vivax* genes that are likely proteases expressed in gametocytes. Access the strategy from your My Strategies section or use this dedicated URL to retrieve our saved strategy: <http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf>. Focus the strategy on the last result by clicking on the Step 4 (orthology) transform box (Fig. 20A, arrow). The active result is highlighted yellow in the strategy panel and its results are displayed in the ‘Gene Results’ tab.
2. *Run a Gene Ontology Enrichment on the Step 4 strategy result.* Click ‘Analyze Results’ (Fig. 20B, arrow) to create a tab for the new analysis. Choose the GO

Ontology Enrichment tool (Fig. 20C) to open the tool (Fig. 20D). The Organism parameter reflects the genome from the result list genes, the genome that is ‘background’. The Ontology parameter allows you to choose one of the three GO ontologies to enrich against. GO ontologies are structured, controlled vocabularies that describe gene products in terms of their related biological processes, cellular components and molecular functions. For statistical reasons, only one ontology may be analyzed at once. If you are interested in more than one, run separate GO enrichment analyses. Choose the ‘Cellular Component’ ontology to look for common cellular location assignments for the result list and click ‘Submit’. The results are displayed as a table of GO IDs and associated data, including p-values and adjusted significance parameters (Fig. 20E).

3. Explore your results while paying attention to the significance metrics. There are several enriched GO terms that indicate the product is located in the proteasome complex. Since the strategy finds proteases that are expressed in gametocytes, it follows logically that this set of genes will be enriched in proteasome complex genes. Start a new analysis to find enriched GO terms from the Biological Process ontology and determine if the enriched biological processes are expected based on the strategy.

2.5.2 EuPathDB Galaxy—EuPathDB includes a Galaxy workspace, a web-based bioinformatics analysis platform that houses a large variety of bioinformatics tools to facilitate large-scale data analysis where no programming experience is required [27]. The EuPathDB Galaxy workspace is managed and maintained in partnership with Globus Genomics (<https://www.globus.org/genomics>), a cloud-based platform for large-scale sequencing analyses [28]. The EuPathDB Galaxy workspace offers pre-loaded EuPathDB reference genomes, several RNA-seq and SNP calling workflows, private data analysis, data and result sharing with individual EuPathDB Galaxy users or the EuPathDB Galaxy community, and data export.

1. Visit FungiDB (<http://fungidb.org>) and click on the ‘Analyze My Experiment’ tab located within the main menu in grey (Fig. 21A) to access the Galaxy workspace. To use Galaxy services, one must have a free account with any of the EuPathDB sites and then complete a short Galaxy sign-up process. Once in the Galaxy workspace, a user is directed to the Welcome page (Fig. 21B), which contains a short introduction to the instance and links to several workflows. The left panel (Fig. 21, B1) offers a number of NGS Applications, microarray, data manipulation, statistical, FASTA, and data managements tools. The center panel (Fig. 21, B2) is controlled by the main Galaxy menu at the top of the page and is linked to the workflow history panel (Fig. 21, B3) on the right. Its interactive interface allows you to create, run, and save custom workflows, visualize histories and result analysis.
2. Upload a dataset from a local computer, the EBI website, or set up an end-point for large data transfers (Fig. 22). The data import tools can be accessed from the left panel under the Get Data option or by clicking on the Upload symbol as shown (Fig. 22A).

3. To begin a pre-configured workflow select the desired type of analysis from the list of workflows in the center panel (Fig. 21, B2). The workflow will be opened in the center panel with a series of prompts to select filenames and parameter values for each analysis tool. A standard RNA-seq workflow includes steps for assessing raw reads quality, trimming and alignment of reads, differential expression calculation, genome mapping, and visualization of results graphically in GBrowse. BigWig files generated during the analysis will be automatically linked to the local EuPathDB GBrowse session, which is not visible to other users (Fig. 21C).
4. Customize a pre-configured workflow by importing a workflow and changing or adding workflow steps within the editor interface (Fig. 22B). Alternatively, new workflows can be created by accessing the Workflow menu at the top of the page and selecting the Create new workflow option (Fig. 21B).

Acknowledgments

EuPathDB would like to acknowledge their current funders, the National Institutes of Health (US), the Wellcome Trust (UK), as well as past funders and The Bill and Melinda Gates Foundation (US), The Burroughs Wellcome Fund (US).

References

1. Aurrecochea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, Crouch K, Doherty R, Falke D, Fischer S, Gajria B, et al. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* 2017; 45(D1):D581–D591. DOI: 10.1093/nar/gkw1105 [PubMed: 27903906]
2. Aurrecochea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, Grant G, et al. EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res.* 2013; 41(Database issue):D684–691. DOI: 10.1093/nar/gks1113 [PubMed: 23175615]
3. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res.* 2002; 12(10):1599–1610. [PubMed: 12368253]
4. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD. Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* 2016; 44(W1):W29–34. DOI: 10.1093/nar/gkw292 [PubMed: 27105845]
5. Peng D, Tarleton R. EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens. *Microb Genom.* 2015; 1(4):e000033.doi: 10.1099/mgen.0.000033 [PubMed: 28348817]
6. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, Prudhomme J, Florens L, Le Roch KG. Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biol.* 2013; 14(11):R128.doi: 10.1186/gb-2013-14-11-r128 [PubMed: 24267660]
7. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, Cui K, Barillas-Mury C, Zhao K, Su XZ. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics.* 2011; 12:587.doi: 10.1186/1471-2164-12-587 [PubMed: 22129310]
8. Lasonder E, Green JL, Camarda G, Talabani H, Holder AA, Langsley G, Alano P. The *Plasmodium falciparum* schizont phosphoproteome reveals extensive phosphatidylinositol and cAMP-protein kinase A signaling. *J Proteome Res.* 2012; 11(11):5323–5337. DOI: 10.1021/pr300557m [PubMed: 23025827]
9. Solyakov L, Halbert J, Alam MM, Semblat JP, Dorin-Semblat D, Reininger L, Bottrill AR, Mistry S, Abdi A, Fennell C, Holland Z, et al. Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. *Nat Commun.* 2011; 2doi: 10.1038/ncomms1558

10. Oehring SC, Woodcroft BJ, Moes S, Wetzel J, Dietz O, Pulfer A, Dekiwadia C, Maeser P, Flueck C, Witmer K, Brancucci NM, et al. Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite *Plasmodium falciparum*. *Genome Biol.* 2012; 13(11):R108.doi: 10.1186/gb-2012-13-11-r108 [PubMed: 23181666]
11. Dewey CN. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol.* 2007; 395:221–236. [PubMed: 17993677]
12. Kanehisa M. The KEGG database. *Novartis Found Symp.* 2002; 247:91–101. [PubMed: 12539951]
13. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006; 34(Database issue):D354–357. [PubMed: 16381885]
14. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016; 44(D1):D457–462. DOI: 10.1093/nar/gkv1070 [PubMed: 26476454]
15. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2016; 44(D1):D471–480. DOI: 10.1093/nar/gkv1164 [PubMed: 26527732]
16. Shameer S, Logan-Klumpler FJ, Vinson F, Cottret L, Merlet B, Achcar F, Boshart M, Berriman M, Breitling R, Bringaud F, Butikofer P, et al. TrypanoCyc: a community-led biochemical pathways database for *Trypanosoma brucei*. *Nucleic Acids Res.* 2015; 43(Database issue):D637–644. DOI: 10.1093/nar/gku944 [PubMed: 25300491]
17. Saunders EC, MacRae JI, Naderer T, Ng M, McConville MJ, Likic VA. LeishCyc: a guide to building a metabolic pathway database and visualization of metabolomic data. *Methods Mol Biol.* 2012; 881:505–529. DOI: 10.1007/978-1-61779-827-6_17 [PubMed: 22639224]
18. Doyle MA, MacRae JI, De Souza DP, Saunders EC, McConville MJ, Likic VA. LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC Syst Biol.* 2009; 3:57.doi: 10.1186/1752-0509-3-57 [PubMed: 19497128]
19. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics.* 2016; 32(2):309–311. DOI: 10.1093/bioinformatics/btv557 [PubMed: 26415722]
20. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. A travel guide to Cytoscape plugins. *Nat Methods.* 2012; 9(11):1069–1076. DOI: 10.1038/nmeth.2212 [PubMed: 23132118]
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13(11):2498–2504. DOI: 10.1101/gr.1239303 [PubMed: 14597658]
22. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* 2010; 38(15):4946–4957. DOI: 10.1093/nar/gkq237 [PubMed: 20385579]
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25(1):25–29. [PubMed: 10802651]
24. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015; 43(Database issue):D1049–1056. DOI: 10.1093/nar/gku1179 [PubMed: 25428369]
25. Lasonder E, Rijpmma SR, van Schaijk BC, Hoeijmakers WA, Kensche PR, Gresnigt MS, Italiaander A, Vos MW, Woestenenk R, Bousema T, Mair GR, et al. Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res.* 2016; 44(13):6087–6101. DOI: 10.1093/nar/gkw536 [PubMed: 27298255]
26. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 2017; 45(D1):D190–D199. DOI: 10.1093/nar/gkw1107 [PubMed: 27899635]

27. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C, Gruning B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016; 44(W1):W3–W10. DOI: 10.1093/nar/gkw343 [PubMed: 27137889]
28. Liu B, Madduri RK, Sotomayor B, Chard K, Lacinski L, Dave UJ, Li J, Liu C, Foster IT. Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *J Biomed Inform.* 2014; 49:119–133. DOI: 10.1016/j.jbi.2014.01.005 [PubMed: 24462600]

A The interactive header is visible from any EuPathDB page. The tabs and dropdown menus in the gray menu bar provide access to all EuPathDB searches and tools.

B The component site link outs section provides direct links to the taxon-specific sites.

C The core section consisting of three panels: 'Search for Genes', 'Search for Other Data Types' and 'Tools'.

D The side bar contains useful links and information including news releases, community resources and a summary of integrated data.

E Find a Search Tool. This text search finds available searches within the Search for Genes bubble.

The EuPathDB Bioinformatics Resource Center provides a portal for accessing genomic-scale datasets associated with the diverse eukaryotic microbes (mouse-over the following logos for information on component websites):

AmoebaDB CryptoDB FungiDB GiardiaDB MicrosporidiaDB PiroplasmaDB PlasmoDB ToxoDB TrichDB TriTrypDB OrthoMCL

Search for Genes

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Transcriptomics
- Sequence analysis
- Structure analysis
- Protein features and properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

expand all | collapse all

Searches that return genes

Search for Other Data Types

expand all | collapse all

Find a search...

- Popset Isolate Sequences
- H-LP Genotype Isolates
- Genomic Sequences
- Genomic Segments
- ESTs
- ORFs

expand all | collapse all

Searches that return other data, ie. Isolates, metabolic pathways, etc..

Tools

BLAST

Identify Sequence Similarities

Results Analysis

Analyze Your Strategy Results

Sequence Retrieval

Retrieve Specific Sequences using IDs and coordinates

Companion

Annotate your sequence and determine orthology, phylogeny & synteny

EuPaGDT

Eukaryotic Pathogen CRISPR guide

RNA/DNA Design Tool

PubMed and Entrez

View the Latest PubMed and Entrez Results

Searches via Web Services

Learn about web service access to our data

Tools such as BLAST and sequence retrieval

EuPathDB 31 9 Mar 17

©2017 The EuPathDB Project Team

POWERED BY Strategies WDK

Please Contact Us with any questions or comments

Search for Genes

signal

Protein targeting and localization

- Predicted Signal Peptide

Figure 1. EuPathDB home page and its main features.

A. The interactive header is visible from any EuPathDB page. The tabs and dropdown menus in the gray menu bar provide access to all EuPathDB searches and tools. **B.** The component site link outs section provides direct links to the taxon-specific sites. **C.** The core section consisting of three panels: 'Search for Genes', 'Search for Other Data Types' and 'Tools'. **D.** The side bar contains useful links and information including news releases, community resources and a summary of integrated data. **E.** Find a Search Tool. This text search finds available searches within the Search for Genes bubble.

Figure 2. Result of a text search in EuPathDB.

Search results are presented in the My Strategies section and consist of three parts. **A.** The Strategy panel provides a graphic representation of the search or strategy result. The search result highlighted in yellow is the 'active' result and further displayed in the Filter tables (B) and the Gene Result (C). **B.** The Component site and organism filter tables shows the distribution of hits from the result across the taxon-specific sites and the organisms queried, respectively. **C.** The Result tables currently showing the Gene Result tab which lists all hits for the active search result. The first column, Gene ID, is a link to the record page for that gene.

Figure 3. Gene record page: Main sections.

A. Record pages include an overview section at the top, with basic information including gene ID, product description or genome location. **B.** Shortcuts are available on the right side of the overview, and provide quick navigation links, but also quick views of the images that appear in the data section of the gene record. **C.** The data section is displayed below the overview. Organized in consistent, site-wide categories, the data section contains all available information about the gene. **D.** The searchable, and collapsible 'Contents' menu gives easy access to all the data sections (C). The contents section will remain visible while

scrolling the record page and clicking on the double arrow icon will collapse the menu, giving full screen width to the record entry.

Figure 4. Gene record page: Shortcuts.

A. Shortcuts can be found at the top of the gene page, on the right side of the overview section. Clicking on the magnifying glass icon (blue circle), will open a graphical display summarizing the data. Clicking on a shortcut image, or on the title above it (blue oval), navigates to the corresponding section of the record page (**B**).

PlasmoDB Plasmodium Genomics Resource
Release 31
9 Mar 2017

Gene ID: PF3D7_1133400 Gene Text Search: synth*

About PlasmoDB | Help | Login | Register | Contact Us

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community | Analyze My Experiment | My Favorites

Add to basket | Add to favorites | **Download Gene**

PF3D7_1133400 apical membrane antigen 1

Name: AMA1
Type: protein coding
Chromosome: 11
Location: PF3D7_11_v3:1,293,856..1,295,724(+)
Species: Plasmodium falciparum
Strain: 3D7

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.

Shortcuts

Synteny | BLAT Alignments | SNPs | Transcriptomics

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community | Analyze My Experiment | My Favorites

Download Gene: PF3D7_1133400

Choose a Report: * Text: choose from columns and/or tables ?
FASTA (sequence retrieval, configurable) ?

Choose Attributes

select all | clear all | expand all | collapse all

Search Attributes...

- ☒ Organism
- ☒ Gene models
- ☒ Annotation, curation and identifiers
- ☒ Genomic Location
- ☒ Orthology and synteny
- ☒ Genetic variation

select all | clear all | expand all | collapse all

Choose Tables

select all | clear all | expand all | collapse all

Search Tables...

- ☒ Gene models
- ☒ Annotation, curation and identifiers
- ☒ Link outs
- ☒ Genomic Location
- ☒ Literature
- ☒ Orthology and synteny
- ☒ Phenotype
- ☒ Sequence analysis
- ☒ Structure analysis
- ☒ Protein features and properties
- ☒ Protein targeting and localization
- ☒ Function prediction
- ☒ Pathways and interactions
- ☒ Proteomics
- ☒ Immunology

select all | clear all | expand all | collapse all

Download Type

- ☐ Text File
- ☒ Show in Browser

Additional Options

- ☒ Include empty tables

Submit

Figure 5. Gene record page: The ‘Download Gene’ link.

Information available in the gene record, including sequences, can be easily exported using the ‘Download gene’ link, located at the top of the overview section. Users can create their own tables choosing gene attributes of interest.

10 Transcriptomics

Figure 6. Transcriptomics table. Transcript expression datasets are organized in searchable data tables, with expandable rows that reveal detailed data. Each dataset includes expression data in tabular and graphical format, as well as coverage plots for RNA sequence data sets.

Figure 7. Proteomics data on gene page.
A. The Mass Spec.-based Expression table displays peptides mapped to the gene’s protein product. **B.** Hover over the glyphs to reveal details concerning the mapped peptides.

PF3D7_1133400 apical membrane antigen 1 **A**

Name: AMA1
Type: protein coding
Chromosome: 11
Location: Pf3D7_11_v3:1,293,856..1,295,724(+)

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

[View this gene at GeneDB](#)
[View 2 user comments, or add a comment](#)

User Comments Data sets **B**

[Add a comment](#)

Search this table. Showing 2 rows

Comment ID	Headline	PubMed ID(s)	# Related Genes	Reviewed	# Uploaded Files	Made by	Date
1126	Predicted function		1	Pending		SRINIVASAN RAMACHANDRAN, INSTITUTE OF GENOMICS AND INTEGRATIVE	2007-10-17 00:00:00.0

Add a comment to gene PF3D7_1133400 **C**

Please add only scientific comments to be displayed on the gene page for PF3D7_1133400. If you want to report a problem, use the [support page](#). Your comments are appreciated. If this is a **new gene**, please also add a comment in the corresponding [Genome Sequence](#).

Part I: Comment

Headline *

Category (check all that apply) ☐ Gene Model ☐ Name/Product ☐ Function ☐ Expression ☐ Sequence ☐ Phenotype

Comment *

Location

Strand: ☒ Forward ☐ Reverse

Genome Coordinates: 1293856-1295724 [help](#)

Part II: Evidence for This Comment (Optional)

Upload File

Select a file: No file chosen

Brief Description: (4000 max characters)

PubMed ID(s) [help](#)

the article details of the PubMed ID(s) above

<enter valid gene ID>

Digital Object Identifier (DOI) [help](#)

Name(s) [help](#)

Genbank Accession(s) [help](#)

Part III: Other Genes to which you want to apply this comment (Optional)

Gene Identifiers [help](#)

Figure 8. Submitting user comments.

A. Summary section of PF3D7_1133400 gene record page showing “add a comment” link.
B. User Comments table listing comments and associated information. **C.** Form for adding a comment to a gene.

Figure 9. Orthology and Synteny data on gene pages.

A. Header section of TriTrypDB. Enter the gene ID, Tb927.1.4540 to reach the gene page.

B. Contents navigation panel with section 7 chosen will direct the data section to the Orthology and Synteny section. **C.** The gene page Synteny graph showing tracks for *T. brucei* TREU927 and *T. brucei* Lister 427. **D.** Hovering over the glyphs in the Synteny graph reveals details concerning the gene.

A Search for Other Data Type

expand all | collapse all

Find a search...

- ▶ Popset Isolate Sequences
- ▶ Genomic Sequences
- ▶ Genomic Segments
- ▶ SNPs
- ▶ ESTs
- ▶ ORFs
- ▼ Metabolic Pathways
 - Compounds
 - Genes
 - Identifier (pathway, gene, compound, etc.)
 - Pathway Name/ID
- ▶ Compounds

B Identify Metabolic Pathways based on Pathway Name/ID

Pathway Source: Any

Pathway Name or ID: Begin typing to see suggestions...

Pathway wildcard search

glycolys

- Glycolysis / Gluconeogenesis (ec00010) (KEGG)
- glycolysis I (GLYCOLYSIS) (LeishCyc)
- glycolysis I (GLYCOLYSIS-1) (TrypanoCyc)
- glycolysis I (from glucose 6-phosphate) (GLYCOLYSIS) (MetaCyc)
- glycolysis I (from glucose 6-phosphate) (GLYCOLYSIS) (TrypanoCyc)
- glycolysis II (from fructose 6-phosphate) (PWY-5484) (MetaCyc)
- glycolysis II (from fructose 6-phosphate) (PWY-5484) (TrypanoCyc)
- glycolysis III (from glucose) (ANAGLYCOLYSIS-PWY) (MetaCyc)
- glycolysis IV (plant cytosol) (PWY-1042) (MetaCyc)

C

5.4.2.12

2-phosphonate-3-glycerate(3-)

4.2.1.11

water

phosphonate-2-olpyruvate

hydron

ADP(3-)

ATP(4-)

pyruvate

D Node Details

EC Number or Reaction: 4.2.1.11

Enzyme Name: Phosphopyruvate hydratase

Show 89 gene(s) which match this EC Number

Search on OrthoMCL for groups with this EC Number

E

File Layout Paint Enzymes

Clear all

By Experiment

By General

NOTE Click on nodes

hydron

ADP(3-)

ATP(4-)

pyruvate

Figure 10. Metabolic Pathways represented in TriTrypDB.

A. The Search for Other Data Types panel with the Metabolic Pathways category open to reveal the types of searches that return Metabolic Pathway records. **B.** The Pathway Name ID search page depicting the 'typeahead' function for entering pathway names in the Pathway Name/ID parameter. **C.** Partial view of the Glycolytic 1 pathway showing the zoom function (1) product, an enzyme node (2) and a compound node (3). **D.** Node details popup that appears when an enzyme or compound node is clicked. **E.** Enzyme node painted with expression graph from integrated experimental data.

Figure 11. Creating strategies by combining search results.

A. PlasmoDB Strategy returning a list of 74 genes that are likely *P. vivax* proteases and expressed in gametocytes. The strategy is also available here: <http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf>. Creating this strategy in the current database may produce a different result since genome annotations may be updated with new database releases. **B.** Table showing the 5 options for combining searches into a strategy. When two searches are combined, the two result sets (list of IDs) are merged according to the operator that you specify. If the searches return the same type of genomic feature they can be combined using any of the 5 operators (i.e. search 1 returns genes, search 2 returns genes). However, searches that return different genomic features (i.e. search 1 returns genes, search 2 returns SNPs) will yield no results when combined with intersect, union or minus operators because there are no IDs in the list of genes (search 1 result) that are present in the list of SNPs (search 2 results). To combine a search that returns genes with a search that returns SNPs, you must use the collocation option (1 relative to 2) to find, for example, genes with SNPs in their upstream regions.

Europe PMC Funders Author Manuscripts

Figure 13. Creating Step 2 of the Strategy (Example 1).

A. The Add Step button for initiating subsequent strategy steps. **B.** The Add Step popup for choosing searches the next search in the strategy. All searches are available from this popup. **C.** The GO Term search depicting the choice of GO Terms using the 'GO Term or GO ID' parameter type ahead. **D.** The strategy result after running the second search in the strategy – the GO Term search.

Add Step A

Add Step 3 : RNA Seq Evidence

Filter Data Sets: Legend: DE Diff. Ch... FC 1-fold Ch... P Percentile

Type keyword(s) to filter

Organism	Data Set	Choose a search
<i>P. berghei</i> ANKA	5 asexual and sexual stage transcriptomes (Hoeijmakers et al.)	FC P
<i>P. chabaudi</i> chabaudi	Trophozoite transcriptomes after mosquito transmission or direct injection into mice (Spence et al.)	DE FC P
<i>P. falciparum</i> 3D7	NSR-seq Transcript Profiling of malaria-infected pregnant women and children (Vignali et al.)	FC P
<i>P. falciparum</i> 3D7	Polysomal and steady-state asexual stage transcriptomes (Bunnik et al.)	FC P
<i>P. falciparum</i> 3D7	Blood stage transcriptome (3D7) (Otto et al.)	FC P
<i>P. falciparum</i> 3D7	Female and Male Gametocyte Transcriptomes (Lasonder et al.)	FC P
<i>P. vivax</i> Sal-1	Transcription profile of intraerythrocytic cycle (Zhu et al.)	FC P
<i>P. yoelii</i> yoelii 17X	Salivary gland sporozoite transcriptomes: WT vs. Puf2-KO (Lindner et al.)	FC P

Add Step B

Add Step 3 : *P. falciparum* 3D7 Female and Male Gametocyte Transcriptomes RNASeq (percentile)

Experiment: Female and Male Gametocyte Transcriptomes - Sense

Samples: ☒ male gametocyte ☒ female gametocyte Choose both samples

Minimum expression percentile:

Maximum expression percentile:

Matches Any or All Selected Samples?: any

Protein Coding Only: protein coding

Combine Genes in Step 2 with Genes in Step 3:

Choose to intersect step 2 results with these results

☒ 2 Intersect 3
 ☐ 2 Minus 3
☐ 2 Union 3
 ☐ 3 Minus 2
☐ 2 Relative to 3, using genomic colocation

Run Step

C

GO: proteolysis 2160 Genes → PINF54 Gametocyte 1699 Genes → protease 1631 Genes

Step 1: protease 1631 Genes → Step 2: 2803 Genes → Step 3: 65 Genes → Add Step

D

All Results	Ortholog Groups	<i>P. berghei</i> ANKA	<i>P. chabaudi</i> chabaudi	<i>P. coatneyi</i> I lackeri	<i>P. cynomolgi</i> strain 01	<i>P. falciparum</i> 3D7	<i>P. falciparum</i> (nr Genes: 65)
65	64	0	0	0	0	65	0

Figure 14. Creating Step 3 of the PlasmoDB strategy.

A. The Add step popup showing the available searches against RNA sequencing data sets. **B.** The search form for the chosen gametocyte RNA sequencing data set. **C.** The strategy results after adding Step 3. **D.** The filter table for Step 3 results. Only *P. falciparum* genes are returned in step three because the RNA sequencing experiment was performed with *P. falciparum* parasites.

Figure 15. Transform by Orthology tool.

A. The Add Step popup for accessing the tool. **B.** The transform by Orthology tool configured to transform to *P. vivax* Sal1. **C.** The final four step strategy returning 74 *P. vivax* genes that likely have protease activity and expressed in gametocytes.

Figure 16. Find *T gondii* and *N caninum* genes that are predicted to be localized to the apicoplast. **A.** The EuPathDB Search for Genes panel with the Protein targeting and localization category opened. The P.f. Subcellular Localization search is accessible here. **B.** The P.f. Subcellular Localization search page containing only one parameter. **C.** The strategy panel showing the result of Step 1. **D.** The Transform by Orthology tool arranged to transform genes from the previous step into *T. gondii* ME49, *T. gondii* GT1 and *N. caninum* Liverpool. **E.** The strategy panel after the transformation. **F.** The Add Step panel configured to access the Orthology Phylogenetic Profile search.

Figure 17. The Orthology Phylogenetic Profile search

A. Parameter for defining the orthology-based phylogenetic profile of the genes returned by the search. The phylogenetic profile of a gene is a series of "present" or "absent" calls, reflecting the inclusion of a gene in ortholog groups determined by the OrthoMCL algorithm. As shown, the parameter is configured to return genes that do not have orthologs in *Cryptosporidium* or Mammalia. **B.** A three-step strategy that returns a refined set of *T. gondii* ME49, *T. gondii* GT1 and *N. caninum* Liv genes that are likely targeted to the

apicoplast. The completed strategy is available here: <http://eupathdb.org/eupathdb/im.do?s=3353bf3401d62d48>

Figure 18. The Genome Browser main features.

A. The ‘View in genome browser’ link from all gene pages, open the browser in the region of the gene. **B.** The browser’s main features: the landmark region (1), the Overview, Region and Details scales (2), track controls (3), zoom and scrolling controls. **C.** The Select Tracks tab for choosing tracks to display in the browser.

Figure 19. The Genome Browser for data visualization and mining.

A. TGME49 genome in the region of the HXGPRT gene as displayed in the Genome Browser. Data tracks showing the current gene model and supporting splice junctions (introns) determined from RNA sequencing data. **B.** Tracks created from CRAIG gene prediction analysis output. These tracks show an alternative to the official gene model. **C.** RNA Sequencing reads from a single tachyzoite sample aligned to the genome. **D.** RNA sequencing reads aligned to the genome and displayed with three subtracks overlaid for easy viewing. **E.** Three subtracks representing time points of an RNA sequencing experiment

measuring transcriptomes of cat enteroepithelial stages. **F.** Expressed sequence tag alignments.

Figure 20. The Result Analysis Tool.

A. PlasmoDB strategy focuses on Step 4. Use this URL to access the strategy <http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf>. **B.** The strategy's gene result showing the Analyze Results button. **C.** The Gene Ontology Enrichment tool button. **D.** The Gene Ontology Enrichment tool showing parameters **E.** Results of a GO enrichment analysis, displaying enriched GO IDs and associated data.

Figure 21. EuPathDB Galaxy access and main features.

A. Shown in FungiDB. The Galaxy instance can be accessed via the Analyze My Experiment tab, which is conveniently located within the main menu (in grey). **B.** From left to right. The workspace has four major components: the left panel (1) lists available large-scale data analysis tools, the center panel (2) which is the main interactive interface and also contains pre-configured workflows for the RNA-seq analysis, and the job history (3) panel on the right. The main panel is controlled via the Galaxy menu at the top. **C.** BigWig file displaying RNA-seq peaks for a gene in the filamentous fungus *Aspergillus nidulans*. Files

are automatically directed to FungiDB via Display in FungiDB GBrowse links available in the job history panel.

A

globus Genomics

Tools

Download from web or upload from disk

Regular Composite

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed

Name	Size	Type	Genome	Settings	Status
New File	67 b	Auto-detect	FungiDB-29_Aoryza...		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR643/SRR643876/SRR643876.fastq

New File 67 b Auto-detect FungiDB-29_Aoryza...

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR643/SRR643876/SRR643876.fastq

Type (set all): Auto-detect Genome (set all): FungiDB-29_AoryzaRIP40_Genome

Choose local file Paste/fetch data Pause Reset Start Close

B

Input dataset x

output

Input dataset x

output

Input dataset x

output

Input dataset x

output

FastQC

Queue read data from your current history

Contaminant list

Submodule and limit specifying file

html_file (html)

text_file (txt)

Sickle

Paired-End Forward Strand FastQ Reads

Paired-End Reverse Strand FastQ Reads

output_single

output_paired1

output_paired2

output_paired_single

output_log (txt)

FastQC

Short read data from your current history

Contaminant list

Submodule and limit specifying file

html_file (html)

text_file (txt)

Sickle

Paired-End Forward Strand FastQ Reads

Paired-End Reverse Strand FastQ Reads

GSNAP

Select a fastq dataset

Select the paired reads reverse dataset

gsnap_stderr (txt)

gsnap_out (txt, sam, gsnapi)

gsnap_fq (fastq)

unpaired_mult (txt, sam, gsnapi)

unpaired_uniq (txt, sam, gsnapi)

unpaired_transloc (txt, sam, gsnapi)

halfmapping_mult (txt, sam, gsnapi)

halfmapping_uniq (txt, sam, gsnapi)

halfmapping_transloc (txt, sam, gsnapi)

paired_mult (txt, sam, gsnapi)

paired_uniq (txt, sam, gsnapi)

paired_transloc (txt, sam, gsnapi)

concordant_mult (txt, sam, gsnapi)

concordant_uniq (txt, sam, gsnapi)

concordant_transloc (txt, sam, gsnapi)

nomapping (txt, sam, gsnapi)

nomapping_fq (fastq)

nomapping_1_fq (fastq)

nomapping_2_fq (fastq)

GSNAP

Select a fastq dataset

Select the paired reads reverse dataset

gsnap_stderr (txt)

gsnap_out (txt, sam, gsnapi)

gsnap_fq (fastq)

unpaired_mult (txt, sam, gsnapi)

Sort Sam

SAM/BAM dataset to be sorted

outfile (bam)

Sort Sam

SAM/BAM dataset to be sorted

outfile (bam)

Figure 22. File Transfer to Galaxy and Workflows.

A. To upload raw read files to Galaxy, the Paste/Fetch data button can be used to specify ftp addresses of the raw reads files at EBI. Genomes can be selected from the Genome drop-down menu. **B.** Create workflows in the EuPathDB Galaxy workspace. A portion of the sample RNA-seq workflow is shown. This workflow can be modified and saved for later use.

Table 1
EuPathDB Resources

Resource	Address (http://)	Organisms Supported
EuPathDB	eupathdb.org	All organisms below
AmoebaDB	amoebadb.org	<i>Acanthamoeba</i> (11), <i>Entamoeba</i> (5), <i>Naegleria</i>
CryptoDB	cryptodb.org	<i>Chromera</i> , <i>Cryptosporidium</i> (6), <i>Gregarina</i> , <i>Vitrella</i>
FungiDB	fungidb.org	<i>Agaricomycetes</i> (2), <i>Chytridiomycetes</i> (2), <i>Eurotiomycetes</i> (18), <i>Pucciniomycetes</i> (2), <i>Saccharomycetes</i> (4), <i>Schizosaccharomycetes</i> (3), <i>Sordariomycetes</i> (10), <i>Tremellomycetes</i> (4), <i>Ustilaginomycetes</i> (3), <i>Zygomycetes</i> (3) + oomycetes
HostDB	hostdb.org	<i>Homo sapiens</i> , <i>Mus musculus</i>
GiardiaDB	giardiadb.org	<i>Giardia assemblages</i> (3), <i>Spironucleus</i>
MicrosporidiaDB	microsporidiadb.org	<i>Anncalia</i> , <i>Edhazaea</i> , <i>Encephalitozoon</i> , <i>Enterocytozoon</i> (4), <i>Hamiltosporidium</i> , <i>Mitosporidium</i> , <i>Nematocida</i> (2), <i>Nosema</i> (2), <i>Ordospora</i> , <i>Pseudoloma</i> , <i>Spraguea</i> , <i>Trachipleistophora</i> , <i>Vavraia</i> , <i>Vittaforma</i>
PiroplasmaDB	piroplasmadb.org	<i>Babesia</i> (3), <i>Cytauzoon</i> , <i>Theileria</i> (4)
PlasmoDB	plasmodb.org	<i>Plasmodium</i> (8)
ToxoDB	toxodb.org	<i>Cyclospora</i> , <i>Eimeria</i> (8), <i>Hammondia</i> , <i>Neospora</i> , <i>Sarcocystis</i> (2), <i>Toxoplasma</i> (18 strains)
TriTrypDB	tritrypdb.org	<i>Blechnomonas</i> , <i>Crithidia</i> , <i>Endotrypanum</i> , <i>Leishmania</i> (15), <i>Leptomonas</i> (2), <i>Trypanosoma</i> (7)
TrichDB	trichdb.org	<i>Trichomonas</i>
OrthoMCL	orthomcl.org	Proteins from 150 organisms across the tree of life

Number in parentheses indicates the number of organisms from that genus.

Table 2
List of major data types and example techniques

Data Type	Example Technique	Example Source
Genome Sequence and Annotation	Illumina, 454, PacBio	NCBI GenBank,
Orthology Profiles	Orthology group assignments via in-house OrthoMCL analysis	In-house analysis
Genome analyses	Splign, Tandem repeat finder, Low complexity finders	In-house analysis
Domain predictions	SignalP, HMMProfam, TMHMM	In-house analysis
Transcriptomics	RNA Sequencing, microarray, ESTs	NCBI SRA, GEO
Proteomics	Mass spec evidence Quantitative MS evidence	Individual labs
Epigenomics	ChIP-chip, ChIP Seq	NCBI SRA, GEO
Metabolomics	Mass Spec evidence, Metabolite	Individual labs, literature
Isolate data	Population resequencing PopSet sequences	NCBI SRA, PopSet
Host Pathogen Interactions	Protein array (serum)	Individual labs, literature
Metabolic Pathways	N/A	MetaCyc, KEGG, TrypanoCyc, LeishCyc
Compounds	N/A	EBI ChEBI
Phenotypes	CRISPR Screens, curation.	Broad, AspGD, Individual labs, literature

Table 3
Gene page sections and content descriptions

Section	Section Contents
Gene models	Gene structure, introns, exons, UTRs, alternative transcripts). Includes a gene model graphic and summary of supporting transcriptomic data.
Annotation, curation and identifiers	User comments, notes from curators, community annotation projects, alternative product descriptions, gene names, synonyms and previous identifiers.
Link outs	Links to other databases and resources that serve as alternative or specialized sources for additional information about our gene (ex: Entrez Gene, UniProtKB, PDB, GeneDB, Ensembl...)
Genomic Location	Coordinates of the gene at the sequence level (chromosome, scaffold or contig, nucleotide). Links to the Genome Browser centered in the gene of interest.
Literature	Publications containing useful information about the gene. Either automatically retrieved from GenBank records or manually curated.
Taxonomy	Classification of the organism following the NCBI taxonomy.
Orthology and synten	Ortholog group assignments as predicted by OrthoMCL This section also contains a tool for aligning the gene sequence against up to 15 of the genomes included in the database.
Phenotype	Collection of mutant phenotypes, manually curated from publications or inferred from high-throughput phenotyping experiments.
Genetic Variation	Graphic summary of the SNPs detected in this region, with links to our genomic variation GBrowser tracks. Alignment tool for explore differences between isolates
Transcriptomics	Transcript expression datasets are arranged in searchable data tables, with expandable rows. Each dataset includes expression data in tabular and graphical format, as well as coverage plots for RNA sequence data sets.
Sequences	Data table containing genomic, mRNA and protein sequences for each transcript.
Sequence analysis	An interactive graphic summary of BLAT hits against the GenBank non-redundant protein sequence database (NRDB).
Structure analysis	3D structure predictions and similar Protein Data Bank (PDB) chains.
Protein features and properties	Protein domains predicted for this gene, displayed both in a graphical representation and in data tables
Function prediction	Complete Gene Ontology annotations plus enzyme classification numbers, with links to EC numbers and GO term descriptions and publications.
Pathways and interactions	Collection of manually curated and computationally predicted metabolic pathways and protein interactions.
Proteomics	Data tables and graphic summaries of proteomic datasets (Mass Spec-based expression evidence and post translation modification datasets).
Immunology	Predicted epitopes from The Immune Epitope Database (IEDB), and host response datasets, that are organized in expandable data tables, similar to the proteomics and transcriptomics tables.

Table 4
Suggestions for User Comment Content

Comment Type	Example comment
Gene name, including synonym	Purine Phosphoribosyl Transferase, is also known as HPRT, HGPRT, Hypoxanthine Phosphoribosyltransferase, Ppt1, Ppt-1, <i>etc.</i>
Reference	See PMID ##### for functional characterization of this gene." Same reference can be linked to multiple genes, if more than one gene is characterized in the manuscript.
Functional Characterization	This 'hypothetical protein' has been shown to be a copper transporter based on heterologous expression in <i>Xenopus</i> oocytes... Contact <xxxx> for further details.
Subcellular localization	GFP tagging demonstrates that this protein localizes to the mitochondrion, as shown in the attached images. See attached image
Phenotype	Gene knockout has resulted in decreased growth... Contact <xxxx> for further details
Structural information on annotated gene models	The predominant transcript initiation site for this gene has been mapped to ~561 nt upstream of the annotated ATG by 5'RACE and RNase protection. This conclusion is consistent with available RNA-seq data, but differs from the reference annotation. See attached experimental evidence.

Table 5
Tracks used in the ToxoDB Data Visualization in GBrowse Example

Track Title	Category
Annotated Transcripts (with UTRs in gray when available)	Gene Models, Transcripts
RNA-Seq Unified Splice Site Junctions (filtered)	Gene Models, Introns
RNA-Seq Unified Splice Site Junctions (inclusive)	Gene Models, Introns
Annotated Transcripts w/ CRAIG UTR Prediction	Gene Models, Splice Sites
CRAIG <i>de novo</i> Gene Model Prediction	Gene Models, Splice Sites
Tachyzoite Transcriptome 3 and 4 days post-infection (VEG NcLIV) mRNAseq Coverage aligned to <i>T. gondii</i> ME49 (Reid et al.)	Transcriptomics, RNA-Seq, <i>T. gondii</i> ME49, Linear Scale
Tachyzoite Transcriptome Time Series (ME49) Strand Specific mRNA seq Coverage aligned to <i>T. gondii</i> ME49 (Gregory) (linear scale)	Transcriptomics, RNA-Seq, <i>T. gondii</i> ME49, Linear Scale
Transcriptomes of Cat Enteroepithelial Stages (CZ-H3) Strand Specific mRNAseq Coverage aligned to <i>T. gondii</i> ME49 (Hehl Lab)	Transcriptomics, RNA-Seq, <i>T. gondii</i> ME49, Linear Scale
EST Alignments	Sequence analysis, BLAT and Blast Alignments