# Chapter 1

# The Eukaryotic Pathogen Databases: A Functional Genomic Resource Integrating Data from Human and Veterinary Parasites

**Omar S. Harb and David S. Roos**

## Abstract

Over the past 20 years, advances in high-throughput biological techniques and the availability of computational resources including fast Internet access have resulted in an explosion of large genome-scale data sets "big data." While such data are readily available for download and personal use and analysis from a variety of repositories, often such analysis requires access to seldom-available computational skills. As a result a number of databases have emerged to provide scientists with online tools enabling the interrogation of data without the need for sophisticated computational skills beyond basic knowledge of Internet browser utility. This chapter focuses on the Eukaryotic Pathogen Databases (EuPathDB: http://eupathdb.org) Bioinformatic Resource Center (BRC) and illustrates some of the available tools and methods.

**Key words** Eukaryotic, Pathogen, Parasite, EuPathDB, Genomic, Database, Search strategy, Bioinformatics

## 1 Introduction

The EuPathDB BRC [1] is mainly funded by the National Institutes of Health with additional funding for the kinetoplastid component (TriTrypDB) [2] coming from the Bill and Melinda Gates Foundation and the Wellcome Trust and in collaboration with GeneDB [3]. The overarching goal of EuPathDB is to incorporate genomic and postgenomic data from the global research community and making it possible to interrogate the data in an integrative manner.

While EuPathDB includes a family of databases supporting various eukaryotic pathogens (Table 1), the look and feel of these databases have been streamlined to facilitate mobility between databases without the need for reeducation. Hence, protocols described herein can be used universally on any EuPathDB website. In addition, a number of collaborative efforts with groups using the EuPathDB infrastructure [4] extend this usability to

Christopher Peacock (ed.), *Parasite Genomics Protocols*, Methods in Molecular Biology, vol. 1201,
DOI 10.1007/978-1-4939-1438-8_1, 

**Table 1**
**This table lists EuPathDB resources, their web addresses, and the included organisms**

| Database | Web address | Supported organisms |
|---|---|---|
| EuPathDB | http://eupathdb.org | All EuPathDB organisms listed below |
| AmoebaDB | http://amoebadb.org | *Acanthamoeba castellanii, Entamoeba histolytica, E. dispar, E. invadens, E. moshkovskii, E. nuttalli* |
| CryptoDB | http://cryptodb.org | *Cryptosporidium parvum,C. parvum, C. muris* |
| GiardiaDB | http://giardiadb.org | *Giardia lamblia* assemblages A, B, and E |
| MicrosporidiaDB | http://microsporidiadb.org | *Anncaliia algerae, Edhazardia aedis, Encephalitozoon cuniculi, E. hellem, E. intestinalis, E. romaleae, Enterocytozoon bieneusi, Hamiltosporidium tvaerminnensis, Nematocida parisii, Nosema ceranae, Vavraia culicis, Vittaforma corneae* |
| PiroplasmaDB | http://piroplasmadb.org | *Babesia bovis, B. microti, Theileria annulata, T. parva* |
| PlasmoDB | http://plasmodb.org | *Plasmodium berghei, P. chabaudi, P. cynomolgi, P. falciparum, P. gallinaceum, P. knowlesi, P. reichenowi, P. vivax, P. yoelii* |
| ToxoDB | http://toxodb.org | *Toxoplasma gondii, Eimeria tenella, Gregarina niphandrodes, Neospora caninum* |
| TrichDB | http://trichdb.org | *Trichomonas vaginalis* |
| TriTrypDB | http://tritrypdb.org | *Crithidia fasciculata, Trypanosoma brucei, T. congolense, T. cruzi, T. vivax, Leishmania major, L. infantum, L. braziliensis, L. donovani, L. Mexicana, L. panamensis, L. tarentolae, Endotrypanum monterogeii* |
| OrthoMCL | http://orthomcl.org | Includes proteins from over 150 organisms across bacteria, archaea, and eukarya |

other genomic resources including FungiDB (http://fungidb.org) [5], SchistoDB (http://schistodb.net) [6], TBDB (http://www.tbdb.org/wdk/) [7], and BetaCell (http://www.betacell.org) [8].

Searches in EuPathDB are categorized based on the type of returned results. Data in EuPathDB is obtained from publications (or directly from researchers), and from sequence and data repositories such as GenBank, sequencing centers (i.e., the Sanger Institute, the Broad Institute, the J. Craig Venter Institute). Information regarding the source of the data is available on multiple pages within EuPathDB resources and in the extensive data set section (*see* Subheading 4, below).

## 2 Materials

1. Computer (desktop, laptop tablet, or smartphone).
2. Internet browser such as Firefox, Safari, Internet Explorer, or Chrome.
3. Internet access with sufficient bandwidth for web surfing.

## 3 Methods

### *3.1 Building a Search Strategy (In Silico Experiment)*

Searches in EuPathDB resources start by executing an initial query from any of over 80 different available searches. Searches can be used to define sets of genes, isolates, SNPs, genomic segments (i.e., DNA motifs), expressed sequence tags (ESTs), open reading frames (ORFs), or SAGE tags (Fig. 1a). Searches are organized in expandable categories (click on the plus symbol to expand a category) (Fig. 1b). Results of a query are placed into a search strategy, which may be expanded by combining these results with those of additional searches. Results can be combined with each other using intersect, union, or minus operations. To build a search strategy, follow these steps:

1. Define the question you are interested in asking. For example, one may be interested in finding all genes in all apicomplexan parasites available in EuPathDB that are secreted, contain at least four transmembrane domains, have evidence of expression in any parasitic stage based on RNA-sequence evidence, and do not have orthologs in mammals. An answer to such a question is attainable using the integrated search strategy developed by EuPathDB.
2. Identify the searches that will allow you to answer your question. Searches in EuPathDB are triggered against the underlying data such as finding genes with defined characteristics (i.e., genes that have a predicted signal peptide or a specified number of transmembrane domains). An initial search starts by selecting the appropriate link on the home page, clicking on the plus symbol next to a search category, and then selecting the search of interest (Fig. 1b).
3. Define the search parameters and run your first search (Fig. 1c). Species of interest may be selected from the taxonomically organized checklist. Once you are satisfied with your parameters, click on the "Get Answer" button. This will initiate a search strategy that includes a step with the results of the signal peptide search (Fig. 1d).
4. Grow your search strategy by adding additional steps (Fig. 2a). Adding steps is done by clicking on the add step button, selecting a search from the pop-up window, and choosing how to

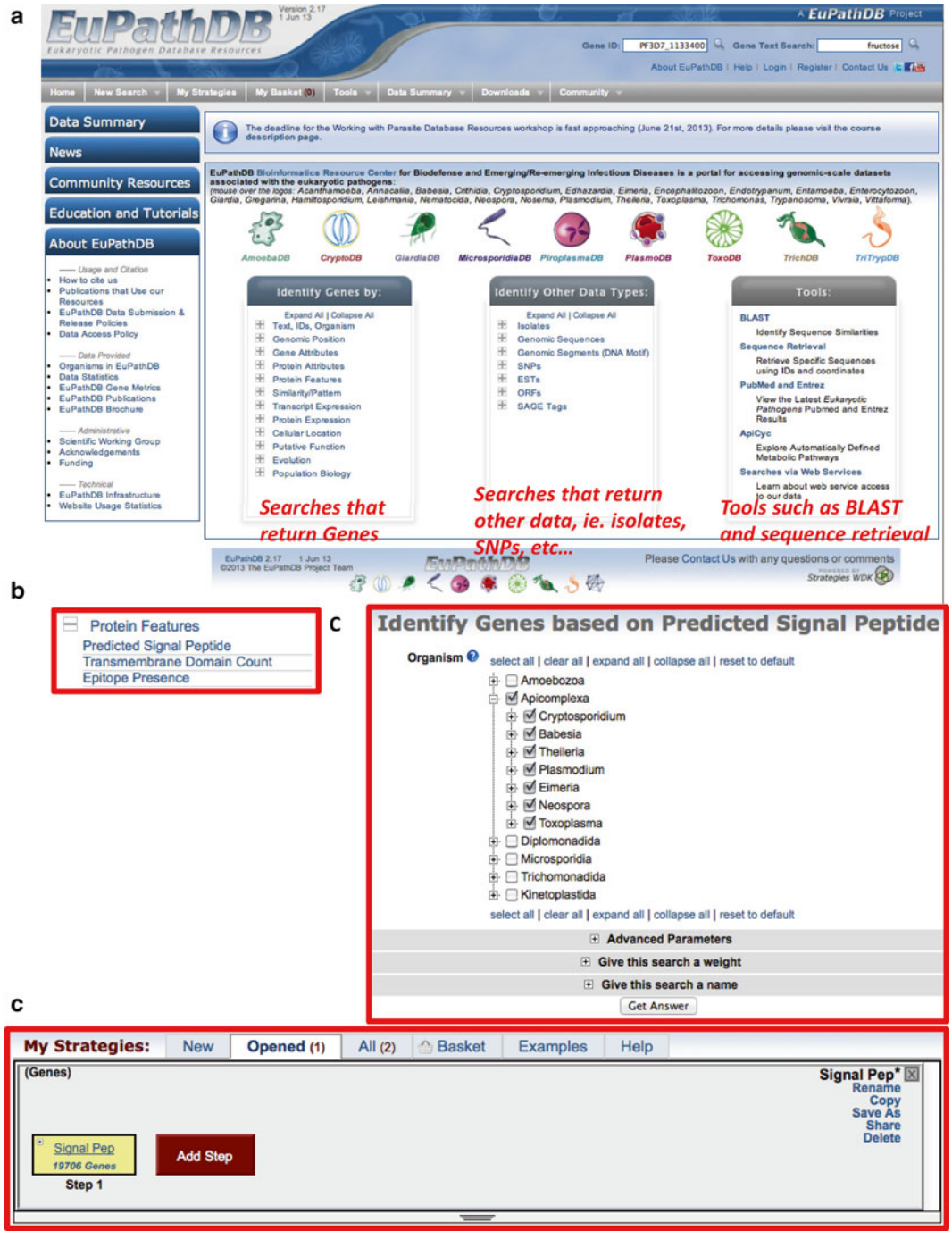


**Fig. 1** Screenshots from EuPathDB depicting the home page and example first search. (**a**) The EuPathDB home page, searches are organized based on the data type they return. (**b**) Categories can be expanded by clicking on the plus symbol to reveal specific searches. (**c**) Once a search is selected the next web page provides search options. In this example, the search page for genes with predicted signal peptides is displayed. Organisms are taxonomically organized and species of interest may be selected. (**c**) Once a search is engaged a search strategy is revealed. This example shows the results of running a signal peptide search on all Apicomplexan organisms in EuPathDB

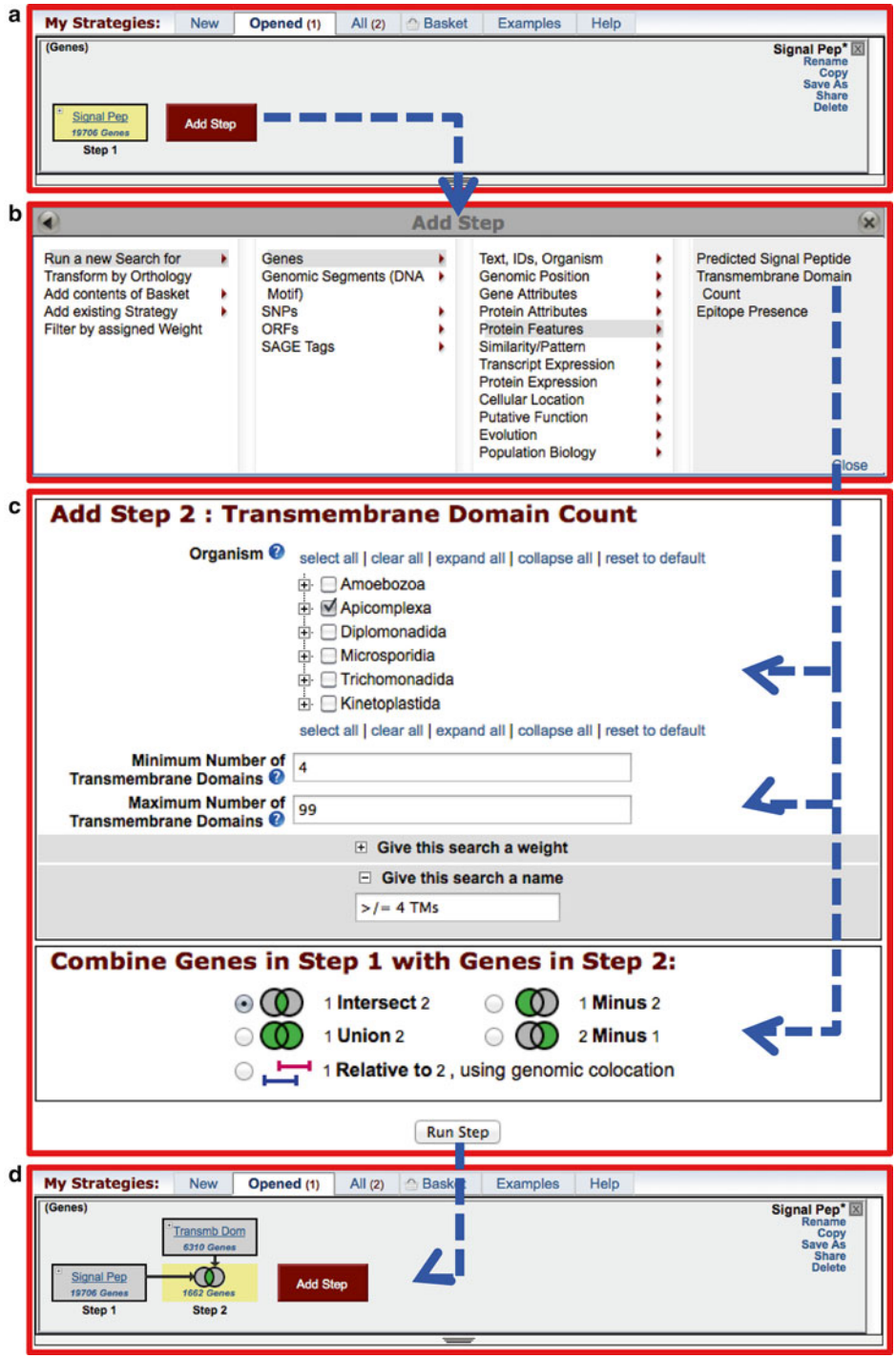


**Fig. 2** Screenshots from EuPathDB depicting the process of adding a step to a search strategy. (**a**) Click on the "Add Step" button to reveal a pop-up window with all available searches in (**b**). Navigate and select the search of interest. (**c**) Once a search is selected a second pop-up window with search parameters becomes available. In addition to selecting search parameters, the method of combining step results needs to be selected (*see* Fig. 3). (**d**) Clicking on the "Run Step" button adds the results of the new search to those of the first search resulting in a two-step strategy

| Name of operation | Symbol of operation in EuPathDB | Definition |
|---|---|---|
| Intersect |  | If A = {1,2,3} and B = {1,2,4,5} then A intersect B = {1,2} |
| Union |  | If A = {1,2,3} and B = {1,2,4,5} then A union B = {1,2,3,4,5} |
| Difference |  | If A = {1,2,3} and B = {1,2,4,5} then A -B = {3} |
| Colocation |  | A is defined by its genomic location relative to the genomic location B. |

**Fig. 3** A graphical representation of the available operations for combining results in a search strategy. Note that the colocation option requires additional parameter selections (described elsewhere in this chapter)

combine the results of this search with those of the previous one. Figure 3 illustrates the type of available operations and their definitions. These include union, intersect, difference, and colocation.

5. Results from all searches are displayed below a search strategy and are dynamically updated as additional steps are added, revised, or deleted. As with any experiment determine if the results are sound: What are the false positives or negatives and are the results plausible?

### *3.2 Using the Orthology Transform Tool*

Genes may be identified based on their characteristics defined by experimental data. Typically, experimental data (i.e., microarray, mass spectrometry, RNA-seq) are collected from a single species of a parasite due to the interest of a lab or experimental accessibility. Orthology may be used to leverage data collected from other species to define genes in your organism of interest. For example, the orthology transform tool enables you to define *Plasmodium falciparum* and *P. vivax* orthologs of genes expressed in liver stages from a microarray experiment performed on *P. yoelii* [9]:

1. Navigate to the Microarray section of PlasmoDB (Fig. 4a) and then select the microarray experiment you wish to query (for this example select "P.y. Liver Stages (fold change))" (Fig. 4b).
2. Define the search parameters. For this example, select up-regulated genes by at least twofold in the blood stage (BS) vs. liver stage 40-h (LS40) comparison (Fig. 4c).
3. Click on the "Get Answer" button. This will start a search strategy with a result of 70 *P. yoelii* genes that are up-regulated in liver stages compared to blood stages (Fig. 4d).

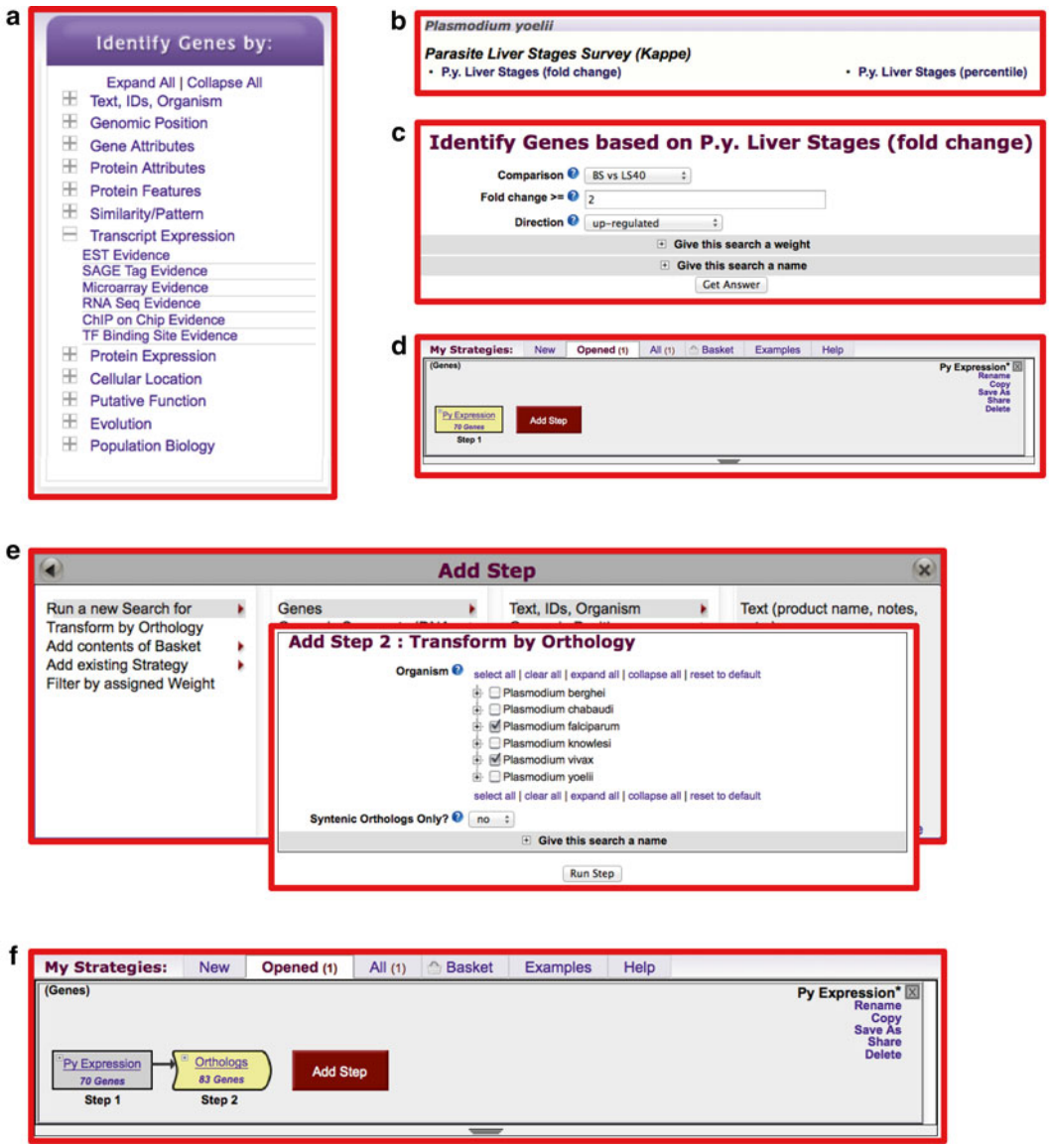


**Fig. 4** Screenshots depicting the utility of the orthology transform tool. (**a**) The "Identify Genes by" portion of the PlasmoDB home page with the "Transcript Expression" category expanded. (**b**) A portion of the microarray expression page depicting the experiment chosen for the search. (**c**) Once an experiment and analysis are selected, search parameters are revealed. (**d**) A search strategy depicting results from a microarray experiment specific to *P. yoelii*. (**e**) Transforming results from one species to another requires adding a step and then selecting the "Transform by orthology" option. The "Transform by Orthology" pop-up window allows the selection of species to transform to. (**f**) A search strategy with the *P. yoelii* results transformed to orthologs in *P. vivax* and *P. falciparum*

4. To define the orthologs of these genes in *P. falciparum* and *P. vivax*, click on add step, in the pop-up select the "Transform by Orthology" option (Fig. 4d), then select the species you wish to transform your results to (Fig. 4e), and click on

"Get Answer." The results are any *P. vivax* and *P. falciparum* genes that are orthologs of the *P. yoelii* genes (Fig. 4f).

Note that orthology in EuPathDB databases is determined using OrthoMCL [10–12].

### 3.3 Building a Search Strategy to Define Secreted Kinases

1. Finding genes using keywords:
   There are a variety of ways to reach a specific gene record in EuPathDB databases. The most straightforward approach is to use the text search option using a specific keyword to identify a gene of interest. This type of approach relies on text available form the annotation, community user comments, genome ontology, InterPro domains, BLAST similarity, etc. The following protocol describes how to identify kinases in PlasmoDB (http://plasmodb.org):
   1. Enter the keyword "kinase" (without quotations) in the search box using either option (a) or (b). Click on the search icon if using option (a), or on the get answer button if using option (b).
      (a) In the "Gene Text Search" box at the top right of any webpage (Fig. 5a).

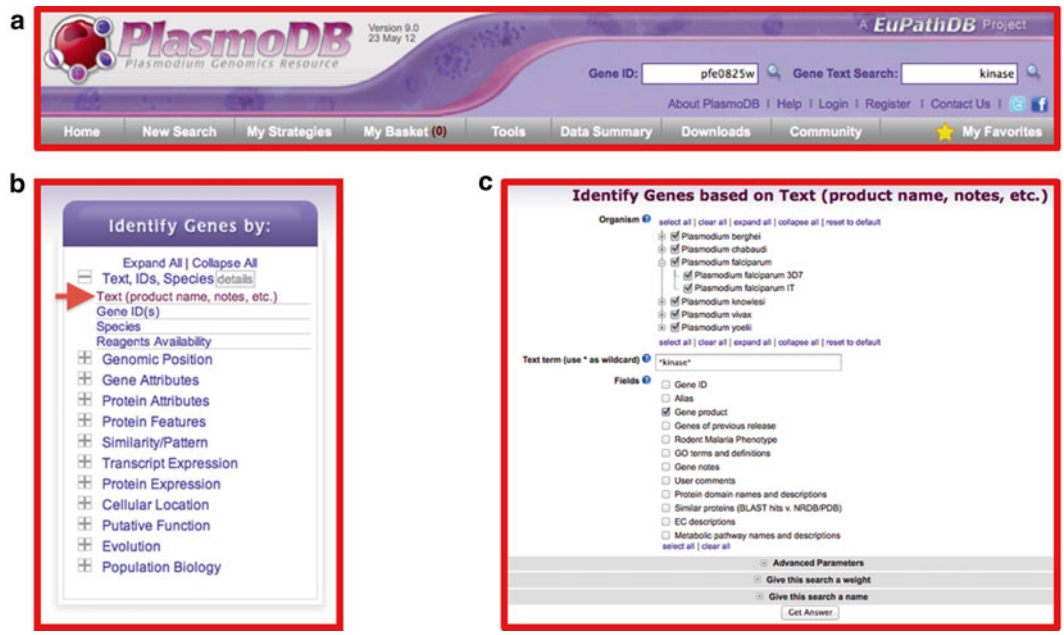


**Fig. 5** Screenshots from PlasmoDB depicting text search options. (**a**) The banner section of PlasmoDB, which includes a text search window in the upper right-hand corner. (**b**) The "Identify Genes by" section of the PlasmoDB home page with the "Text, IDs, Species" category expanded. (**c**) Selecting "Text (product names, notes etc.)" opens a pop-up window that enables specifying organisms to search, the text term, and the fields to search

(b) Via the text search query page which can be accessed by clicking on the Text query link under "Text, IDs, Species" section located in the "Identify Genes by:" column on the home page (Fig. 5b, c).

2. The search above will miss words like "6-phosphofructokinase" or "kinases." To retrieve genes containing such words you may use a wild card in your search—try "kinase*," "*kinase," and/or "*kinase*" (without quotations).

3. There are two places where a keyword may be entered to search for genes:

2. Finding genes that contain a predicted secretory signal peptide. Add a step to the kinase results that searches for genes with predicted secretory signal peptides. The search for signal peptides can be found under the "Cellular Location" search (Fig. 6a, b).

3. Adding genes with predicted transmembrane domains. Grow this search strategy to also include genes that have predicted

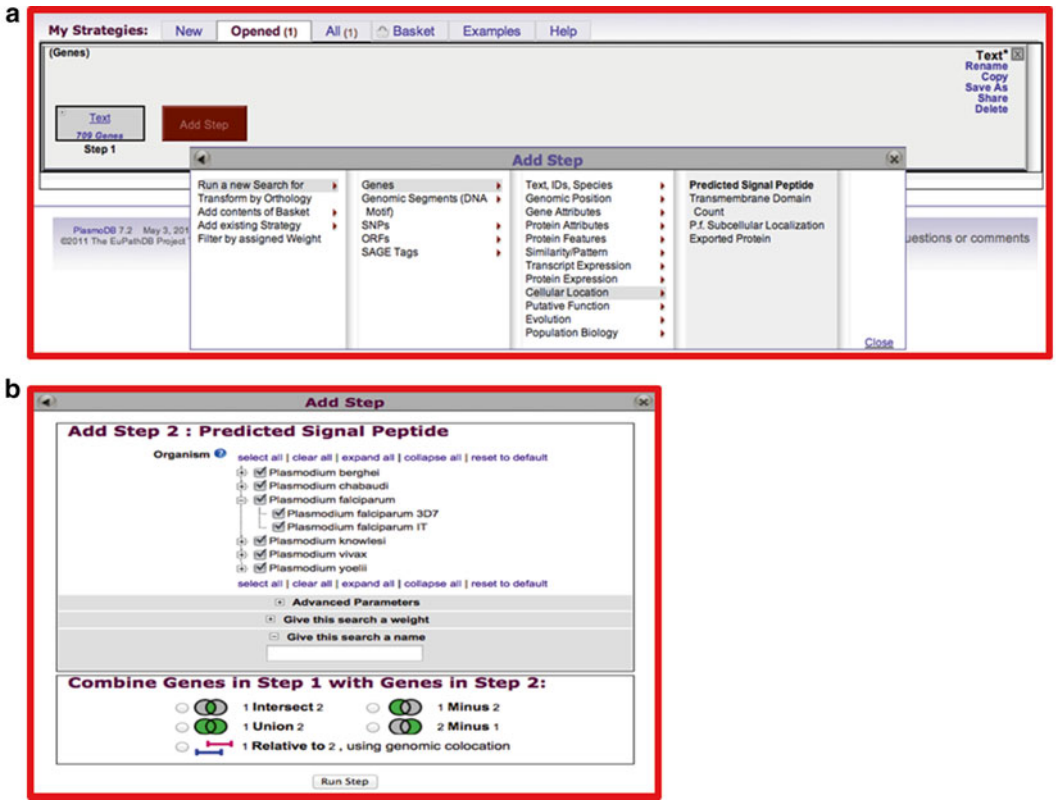


**Fig. 6** Screenshots from PlasmoDB depicting adding a step. (**a**) Clicking on the "Add Step" button reveals a pop-up window with all searches in PlasmoDB. Selecting "Predicted Signal Peptide" under the "Cellular Location" category reveals pop-up that enables the customization of this search (**b**)

transmembrane domains. In this case the goal is to define kinases that have a predicted signal peptide, at least one transmembrane domain or both. Hence, it is critical to expand the signal peptide step into a nested strategy, the results of which will be combined with the list of kinases (Fig. 7). Note that without the option of creating a nested strategy, results would be combined sequentially resulting in very different consequences (Fig. 7c, d).

### 3.4 Identifying Genes Based on Their IDs

Genes may be identified based on their unique identifiers (IDs). EuPathDB maps old IDs to new ones enabling searching with old archival IDs in updated versions of the databases. IDs may be entered one at a time or in bulk—the following protocol employs the ID search in http://PlasmoDB.org.

1. You can find genes based on their IDs, one at a time or in bulk. There are two places where you can enter a gene ID(s):
   (a) The "Gene ID" search box at the top of the home page (Fig. 5a).
   (b) Using the Gene ID query, which can be accessed by clicking on the Gene ID(s) query link under "Text, IDs, Species" section located in the "Identify Genes by:" column on the home page (Fig. 8).
2. When a single gene ID is entered you will be taken directly to the gene page. For example, enter the gene ID for the bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene (PF3D7_0417200) in the Gene ID search box and click on the search icon next to the box (note that EuPathDB databases provide ID mapping of old or alternative IDs to current official gene IDs).
3. Multiple gene IDs may be used as the input in the Gene ID query. This is useful if you have a list of gene IDs from your own experiments or a publication that you would like to further investigate in EuPathDB. In this example a list of gene IDs were obtained from a publication [13]:

   PFF0615c, Pf13_0338, PFE0395c, PF14_0201, PFF0995c, PF10_0346, PF10_0347, PF10_0348, PF10_0352, PF13_0197, PF13_0196, MAL13P1.174, PF13_0193, MAL13P1.173, Pf13_0191, PF13_0192, PF13_0194, PFL1385c, PFB0340c, MAL7P1.208, PF13_0348, PF10_0144, PF14_0102, PFE0080c, PFE0075c, PFD0955w

**Fig. 7** (continued) (**b**) Selecting "Make Nested Strategy" expands the step into a substrategy that can be expended as an independent branch of the search strategy. (**c**) Results of the nested strategy are combined with a step in the main search strategy. (**d**) An illustration of what the results would look like if a nested strategy is not used

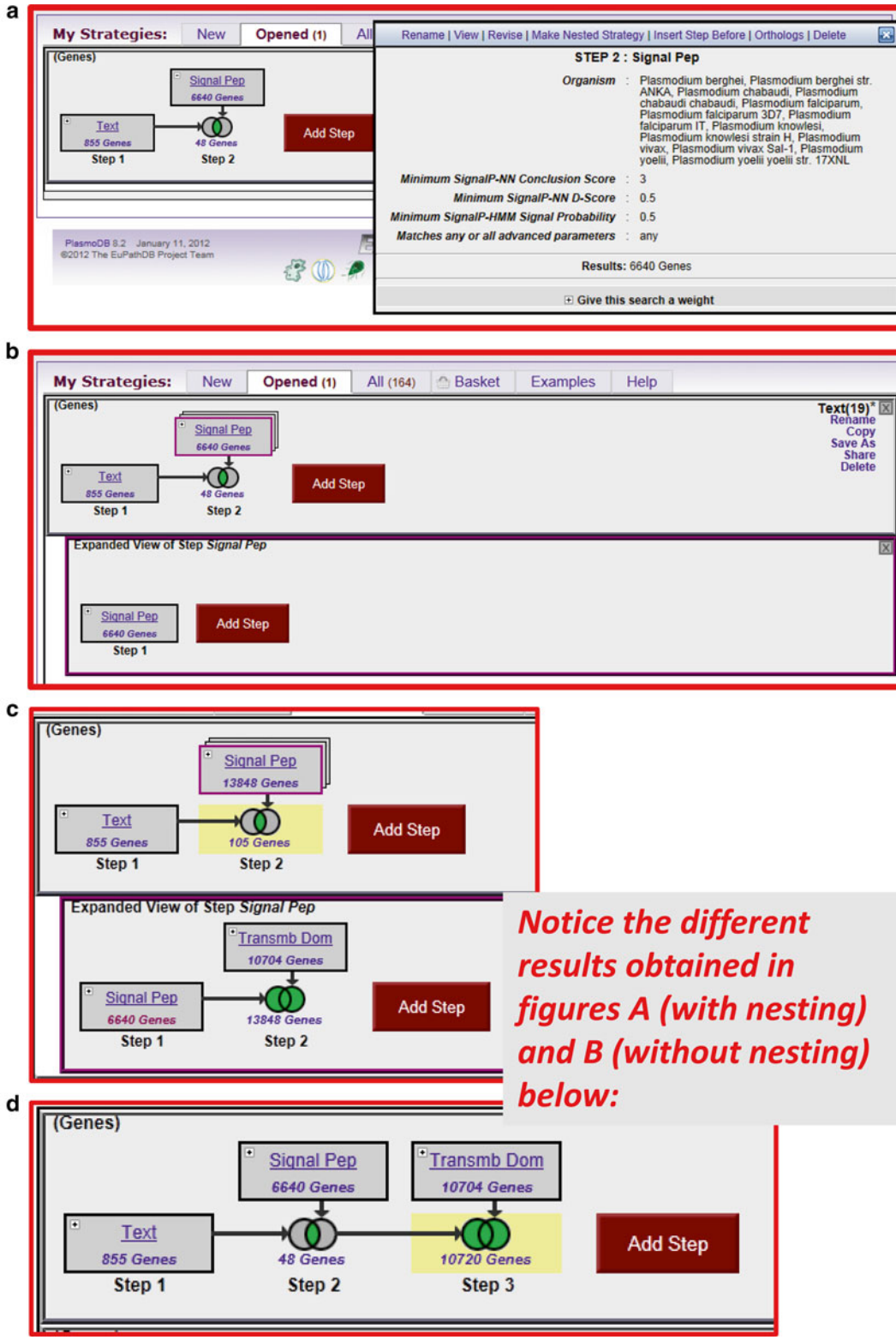


**Fig. 7** Screenshots from PlasmoDB depicting the conversion of a step into a nested strategy. (**a**) Click on the name of the step to be made into a nested strategy. In this image, the signal peptide step was selected. A pop-up window enables the selection of several options including revise, delete, insert step before, and make nested strategy.

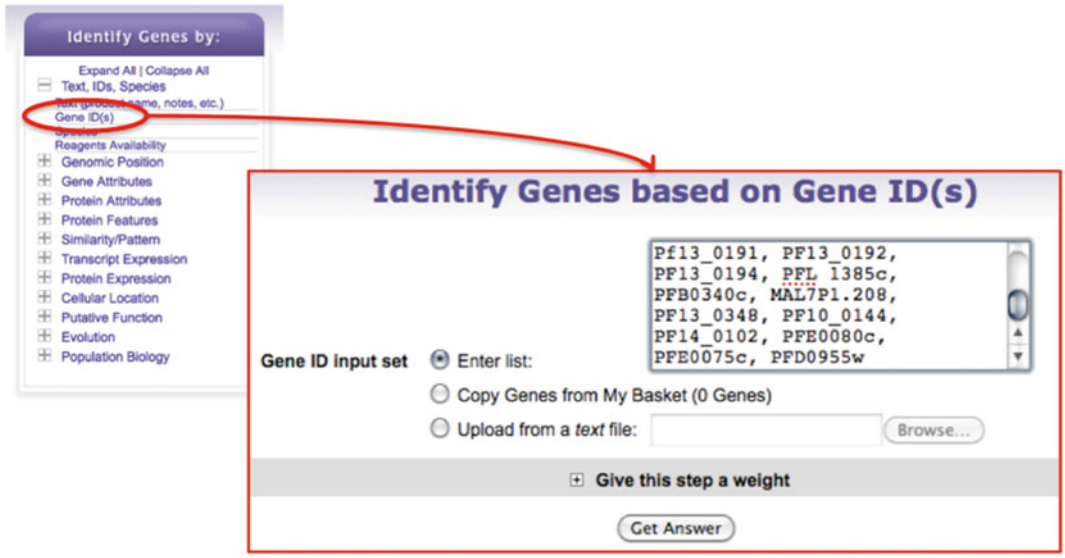


**Fig. 8** Screenshot depicting the Gene ID search window. This option allows searching for a list of gene IDs in one lump sum. IDs may be pasted from another document (i.e., a publication), uploaded from a file, or imported from your basket

Paste these IDs into the ID query and click on the get answer button to retrieve this list of genes in PlasmoDB (Fig. 8).

### *3.5 Using the Colocation Tool to Find Genes Within a Defined Distance from a DNA Motif*

The colocation tool enables the identification entities that can be mapped on a genome (i.e., genes, restriction sites, transcription factor-binding sites, single-nucleotide polymorphisms) based on their relative location to each other (genomic colocation). For example, genes located within 500 nucleotides of transcription factor-binding sites may be defined since both genes and transcriptions factors can be mapped to specific coordinates on a genome. In the protocol presented below, all genes located within 500 nucleotides of a BamHI restriction site are identified using http://MicrosporidiaDB.org.

1. Find all BamHI restriction sites in all microsporidia genomic sequences available in MicrosporidiaDB. BamHI sites are defined by the DNA motif GGATCC. The DNA motif search is under the heading "Genomic Segments" (Fig. 9a). Selecting "DNA Motif Pattern" reveals a pop-up window where the specific nucleotide motif may be defined (Fig. 9b). Note that in addition to entering a DNA motif as a string of IUPAC code, regular expressions may be utilized to defined less stringent motifs. Take a look at your results; notice the Genomic location and the Motif columns (Fig. 9c).
2. Find genes that are 500 nucleotides downstream of the BamHI sites: Add a "Genes by Organism" step to the motif search,

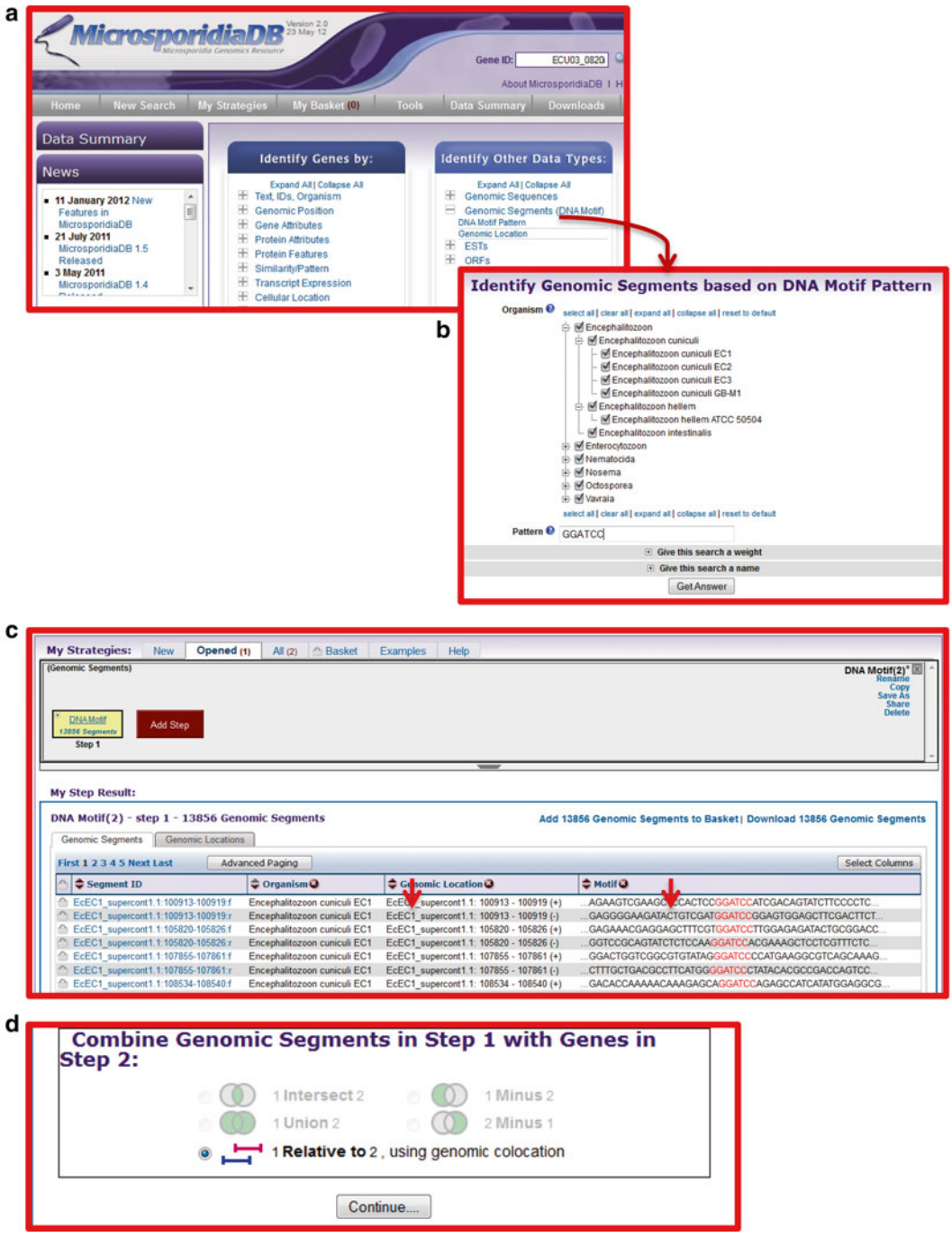


**Fig. 9** Screenshots from MicrosporidiaDB depicting a search for a DNA motif. (**a**) A portion of the MicrosporidiaDB home page with the "Genomic Segments" category expanded. (**b**) Selecting "DNA Motif Pattern" reveals a pop-up window where the specific nucleotide motif may be defined. (**c**) Results of a DNA motif query are displayed as a search strategy. DNA motif records are dynamically generated and displayed as a list of results under the search strategy. (**d**) Results from a DNA motif query may be combined with other types of results (i.e., genes) using the genomic colocation option

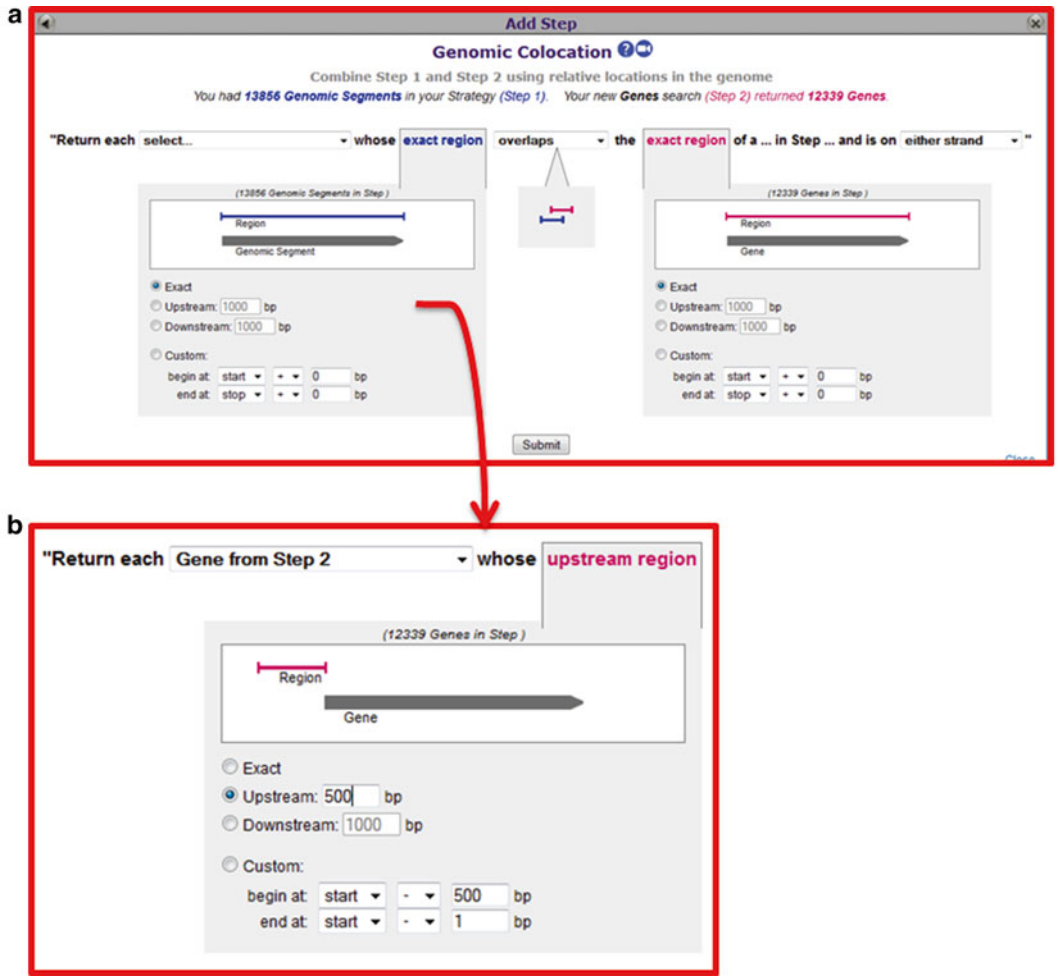


**Fig. 10** A screenshot of the genomic colocation pop-up window. (**a**) The genomic colocation pop-up includes a logic statement that allows selecting the results to return and to define the relationship between the results based on their relative genomic locations. (**b**) An enlarged section of the genomic colocation pop-up window showing a dynamic graphical interface that illustrates the selected relationship

select the "1 relative to 2, using genomic locations" option (Fig. 9d), and click on continue.

3. Use the logic statement at the top of the pop-up window (Fig. 10a) to define which results to return (genes or motifs) and the desired relationship between the genes and the motifs. For this example select genes from step two whose upstream 500 nucleotides contain the motif (BamHI) (Fig. 10b)

### *3.6 Defining Proteins with Specific Amino Acid Motifs*

Genes which translated products contain a specific amino acid motif may be identified using the protein motif pattern search. This query allows defining a motif based on an exact string of amino acids or using a regular expression. The protein motif pattern search can be

found under the heading "Similarity/Pattern" in the "Identify gene by" section of EuPathDB home pages.

Regular expressions are straightforward to compose as illustrated in the example below that finds all proteins in *Trypanosoma cruzi* that contain a signature motif for trans-sialidases using http://TriTrypDB.org. The search strategy described in this protocol may be accessed here:http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42

1. *T. cruzi* has an expanded family of trans-sialidases. In fact, if you run a text search for any gene with the word "trans-sialidase," you return over 3,500 genes among the strains in the database!!! Try this and see what you get (Fig. 11a).

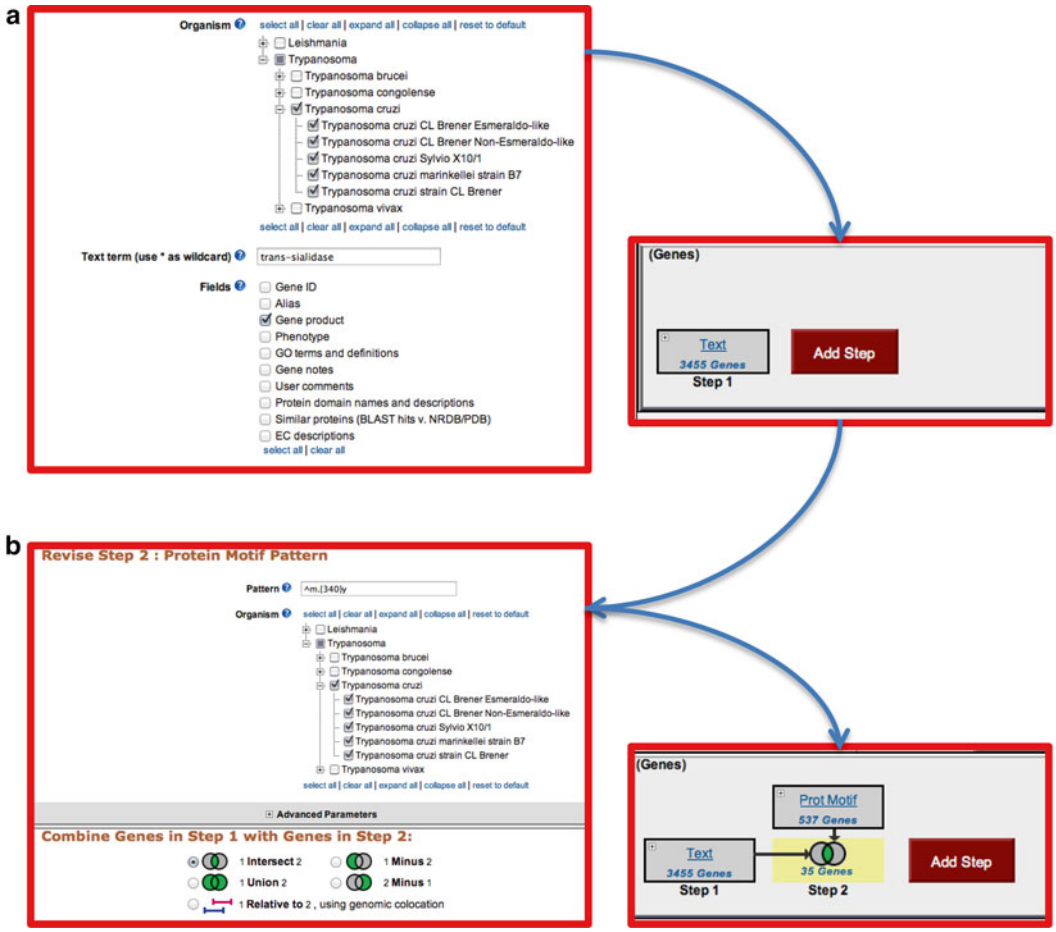


**Fig. 11** Screenshots representing a text search in (**a**) combined with the protein motif search in (**b**). (**a**) A text search for all gene products containing the keyword "trans-sialidase" in *Trypanosoma cruzi* returns 3,455 genes. (**b**) A protein motif pattern search for all *T. cruzi* proteins that start with a methionine, followed by 340 amino acids of any kind and a tyrosine (Y) at position 342—represented by the regular expression ^m.{340}y—returns 537 genes. The intersection of both searches is 35 genes

2. However, not all of these are predicted to be active. It is known that active trans-sialidases have a signature tyrosine (Y) at position 342 in their amino acid sequence. Add a motif search step to the text search in "a" to identify only the active trans-sialidases. Note that for this regular expression the first amino acid should be a methionine (start of the protein), followed by 340 of any amino acid, followed by a tyrosine "Y" (Fig. 11b).

### 3.7 Identifying Isolates Based on Associated Metadata

EuPathDB sites integrate isolate data from multiple sources including GenBank. The genetic background of isolates may be defined by single-locus sequencing, single-nucleotide profiling (SNP-Chip), or high-throughput genomic sequencing. Isolate searches are available under the "Isolates" heading in the "Identify Other Data Types" section of EuPathDB home pages (Fig. 12a). The following protocol uses http://CryptoDB.org to identify Cryptosporidium isolates from Europe that were isolated from feces.

1. Find all Cryptosporidium isolates identified from Europe. This is achieved by running an isolate by geographic location search (Fig. 12a, b) and defining Europe as the geographic location.
2. Add a search for isolates based on isolation source (Fig. 12a), select "feces" (Fig. 12c), and combine the results of this search with those from **step 1** (Fig. 12d).
3. Isolate data is displayed in tabular format and can also be viewed graphically on a dynamic world map by clicking on the "Isolate Geographic Location" tab.

## 4 Notes

- Additional exercises used in EuPathDB workshops: http://workshop.eupathdb.org/current/
- Online tutorials: http://tinyurl.com/eupathdbTutorials
- Updated EuPathDB data content summary: http://tinyurl.com/eupathdbSummary
- EuPathDB data set information: http://tinyurl.com/eupathdbdatasource
- EuPathDB news: http://eupathdb.org/eupathdb/aggregateNews.jsp
- EuPathDB data submission standard operating procedure: http://eupathdb.org/EuPathDB_datasubm_SOP.pdf
- Request a workshop or webinar: help@eupathdb.org

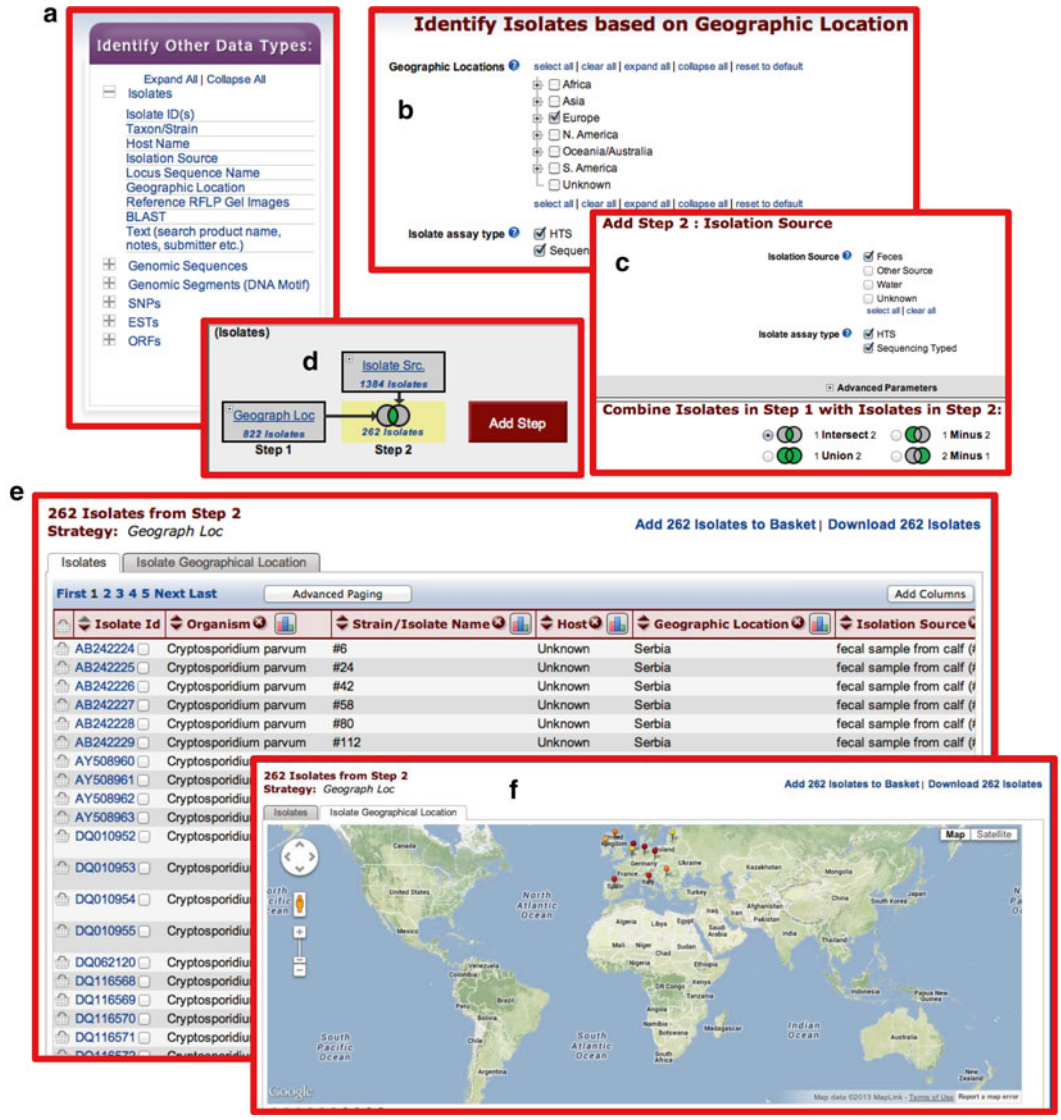


**Fig. 12** Screenshots of an isolate query in CryptoDB. (**a**) A number of searches for isolates are available under the "Isolates" heading in the "Identify Other Data types" section of EuPathDB home pages. (**b**) The search for isolates based on geographic location allows the selection of entire continents or specific countries. (**c**) Isolates may be identified based on their isolation source. (**d**) A combination of geographic location and an isolation source defining 262 Cryptosporidium isolates identified in Europe from feces. (**e**) Isolate search results are listed in a dynamic table that can be sorted and expanded. (**f**) Isolate results may also be visualized graphically on a world map by clicking on the "Isolate Geographic Map" tab above the result list

## References


1. Aurrecoechea C, Brestelli J, Brunk BP et al (2010) EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res 38:D415–9. doi:10.1093/nar/gkp941
2. Aslett M, Aurrecoechea C, Berriman M et al (2009) TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res 38:D457–D462. doi:10.1093/nar/gkp851
3. Logan-Klumpler FJ, De Silva N, Boehme U et al (2012) GeneDB—an annotation database for pathogens. Nucleic Acids Res 40:D98–108. doi:10.1093/nar/gkr1032
4. Fischer S, Aurrecoechea C, Brunk BP, et al. (2011) The strategies WDK: a graphical search interface and web development kit for functional genomics databases. Database (Oxford) 2011:bar027. doi: 10.1093/database/bar027
5. Stajich JE, Harris T, Brunk BP et al (2012) FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res 40:D675–81. doi:10.1093/nar/gkr918
6. Zerlotini A, Heiges M, Wang H et al (2009) SchistoDB: a Schistosoma mansoni genome resource. Nucleic Acids Res 37:D579–82. doi:10.1093/nar/gkn681
7. Galagan JE, Sisk P, Stolte C et al (2010) TB database 2010: overview and update. Tuberculosis (Edinb) 90:225–235. doi:10.1016/j.tube.2010.03.010
8. Mazzarelli JM, Brestelli J, Gorski RK et al (2007) EPConDB: a web resource for gene expression related to pancreatic development, beta-cell function and diabetes. Nucleic Acids Res 35:D751–5. doi:10.1093/nar/gkl748
9. Tarun AS, Peng X, Dumpit RF et al (2008) A combined transcriptome and proteome survey of malaria parasite liver stages. Proc Natl Acad Sci 105:305–310. doi:10.1073/pnas.0710780104
10. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. doi:10.1101/gr.1224503
11. Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–8. doi:10.1093/nar/gkj123
12. Fischer S, Brunk BP, Chen F, et al. (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics 1–19 Chapter 6:Unit 6.12. 1–19
13. Tetteh KKA, Stewart LB, Ochola LI et al (2009) Prospective identification of malaria parasite genes under balancing selection. PLoS One 4:e5568. doi:10.1371/journal.pone.0005568