GENE PAGE EXERCISES
FINDING GENES, BUILDING SEARCH STRATEGIES AND VISITING A GENE PAGE

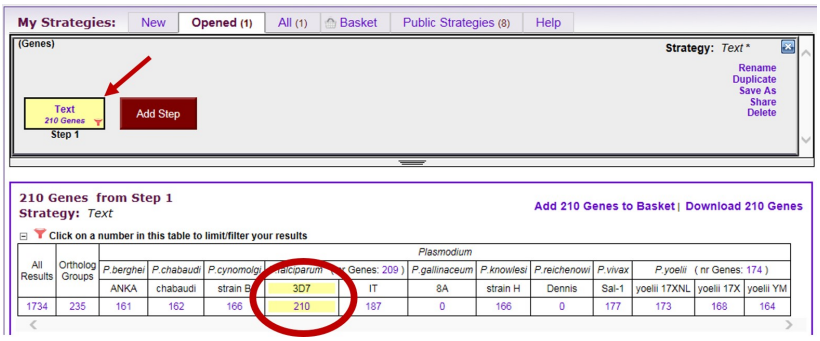
1. Finding a gene using text search.
For this exercise use http://www.plasmodb.org

a. Find all possible kinases in *Plasmodium*.

Hint: use the keyword "kinase" (without quotations) in the "Gene Text Search" box.

- How many genes did you get?
- Look closely at the sections of the result page. How many of those are in *P. falciparum*? How did you find this out?

Hint – the filter table is located between the strategy panel and the result table and shows the distribution of results across the organisms that you searched. Click on a number to display on that species' portion of the results.

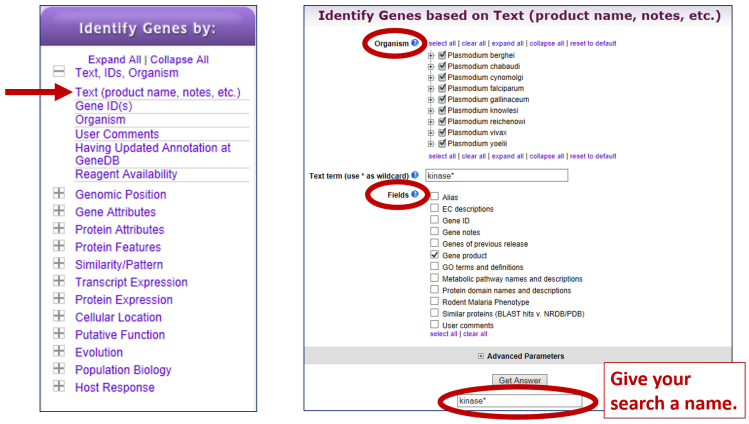


- What happens if you search using the term **kinases** in the Gene Text Search box? How many results are returned?

b. Find only the kinases that specifically have the word "kinase" in the gene product name.

The search you ran in step 1.1a using the Gene Text Search box initiates a preconfigured search. Initiating the search from the full text search form - **Identify Genes based on Text**, allows you to configure the search yourself, choosing parameters that best meet your needs. Use the search form to search for genes that have the word kinase in their **gene product** name/description.

- There are several ways to navigate to the **Identify Genes based on Text** page. Notice the sections of the search page. At the top are parameters and the Get Answer button followed by a search description and a list of datasets used by the search.



- How can you make sure to find your text term in plural form or in compound words like "kinases" or "6-phosphofructokinase". Adding a wild card in your search term will broaden your search. Use the full text search, the specific page where you can define the fields to be searched (Fields = Gene Product).

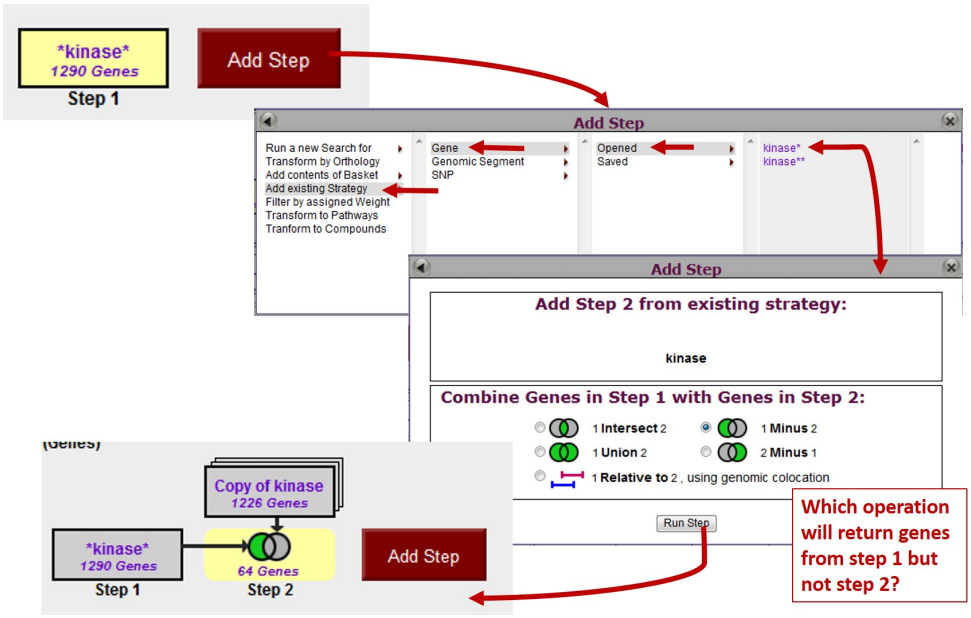
Try kinase *kinase *kinase*

- **Give each new search a name** to help you keep track of the searches.
- How did you get to the Text Search page?
- How does limiting the number of fields searched affect your results?
- Did you remember to use the wild card?
- How many genes have the word kinase in their product names?

c. Combine the results of two text searches.
Find genes that were identified using the key word *kinase* but not the word kinase?

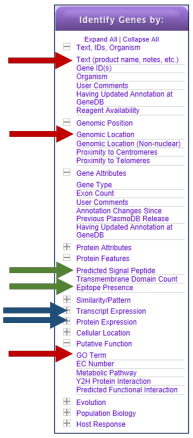
- Here we will build a search strategy that combines 2 of your searches. If you are not displaying the results of the ***kinase*** search (the strategy box will be highlighted in yellow), return to it by clicking on that step box in the strategy panel. To add your **kinase** search to this strategy, click on "Add Step" and select "existing strategy":
- Select the right strategy from your list of Gene Strategies and combine the strategies with the correct operation.

- Do the results make sense? Do all the product names contain the word kinase? From the result page look at the table of gene IDs returned by the search. The Product Description column contains the gene product name.



2. Find all genes between nucleotides 10,000 and 150,000 on chromosome 1 of the *P. berghei* ANKA genome.

In this example we will search for genes based on attributes other than text annotations. In our databases there are many attributes associated with genes and each attribute forms that basis of a search. For example, a gene's genomic location, Gene Ontology assignments, number of exons, number of transmembrane domains are attributes that form the basis of a search for genes within EuPathDB.

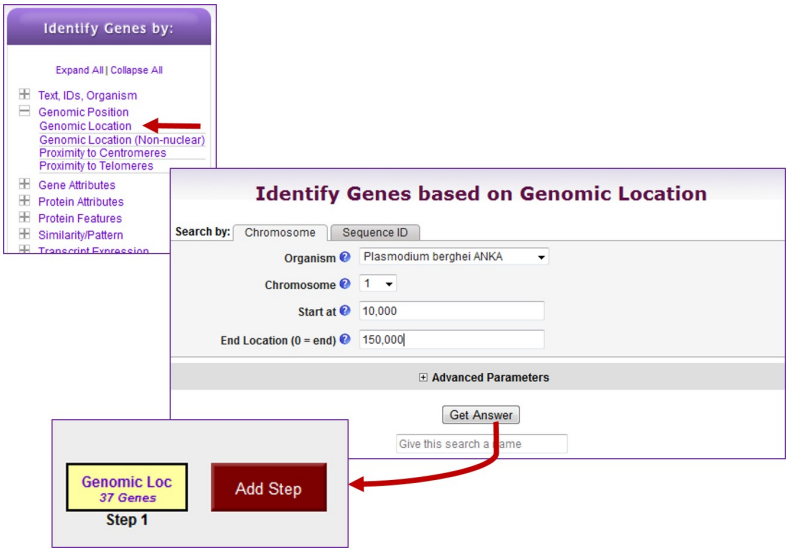


Find Genes Based on:
Annotation
- Text
- Gene ID
- Genomic Location
- Gene Ontology
- Enzyme Commission #
- etc.

Genome Analysis Results
- Predicted Signal Peptide
- Epitope Presence
- Transmembrane Domains

Functional Data
- Microarray
- Proteomics
- RNA Sequencing

- Use the search **Identify Genes by Genomic Position, Genomic Location** to find *P. berghei* ANKA genes that are located between positions 10,000 and 150,000.



- How many genes are located on all of chromosome 2 in ANKA?
- The search offers a second way to define your area of interest. Use one of the genomic sequence IDs in the table below to find genes in other locations; or try a chromosomal location that is important to your research. This search is available on every EuPathDB site.

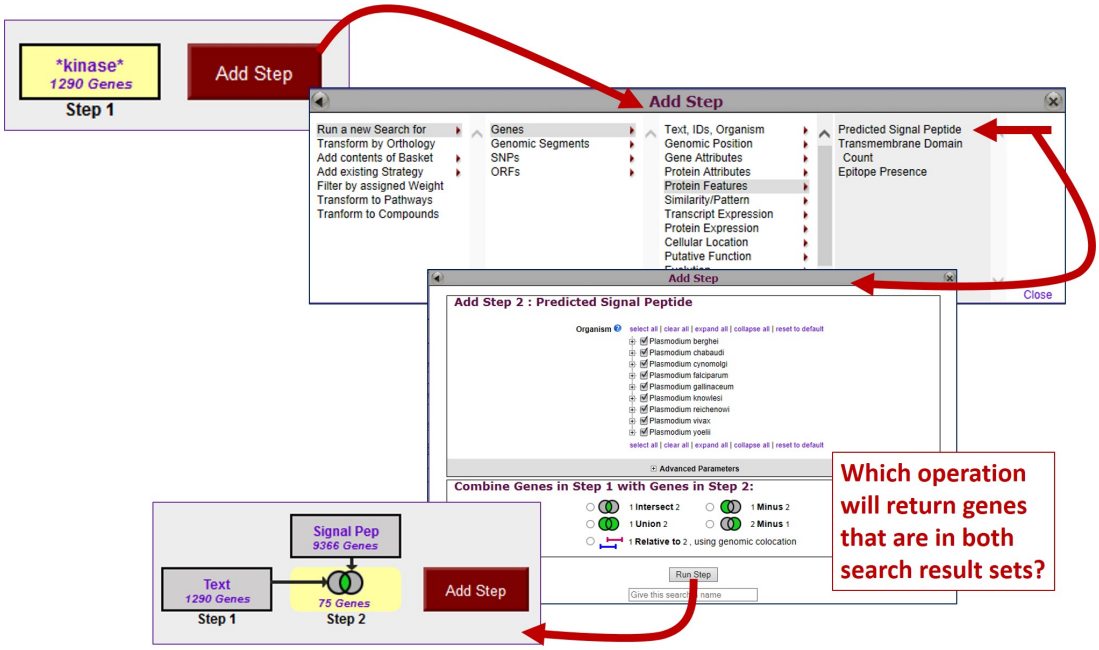
Organism and genomic sequence	Genomic Sequence ID
P. berghei ANKA chromosome 2	berg02
P. falciparum 3D7 chromsome 12	Pf3D7_12_v3
P. yoelii YM apicoplast	PyYM_API_v1

3. Combing text search results with results from other searches

a. Find kinase genes that are likely secreted.

In exercise 1.1b you identified genes that have the word **kinase** somewhere in their gene product name (searching *kinase* in gene product field). Grow your search strategy by adding a step that returns genes whose protein products are predicted to have a signal peptide. In this search you are querying the results of our genome-wide analysis that used the SignalP program to predict the presence and location of signal peptide cleavage sites in amino acid sequences.
http://www.cbs.dtu.dk/services/SignalP/

Focus your Strategies section on the ***kinase*** search and click Add Step. For the second search choose **Identify Genes based on Protein Features, Predicted Signal Peptide**
How did you combine the search results?
How many kinases are predicted to have a signal peptide?

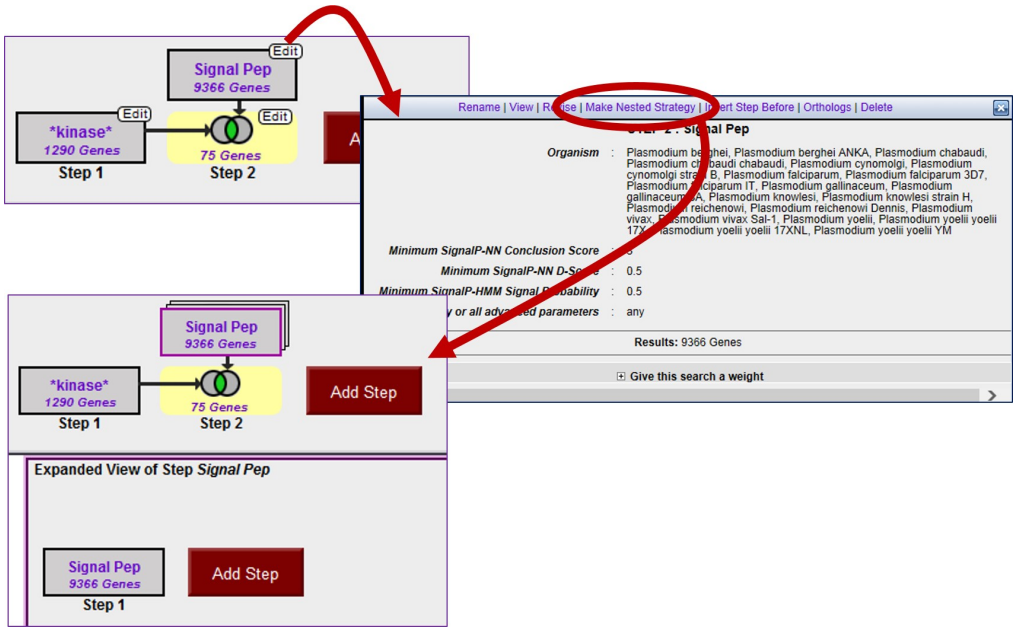


b. Now that you have a list of possible secreted kinases, expand this strategy even further.

There is no wrong answer here!!

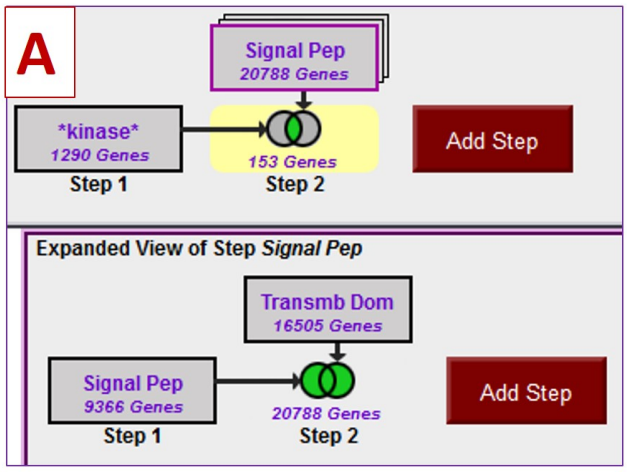
- From a biological standpoint what else would be interesting to know about these kinases? Add more searches to grow this strategy. Open the categories under Identify Genes By: on the home page and explore the types of searches that are available. You can reduce (or expand) your result set by adding searches that are based on many types of data.

- For example, how many of the secreted kinases also have transmembrane domains?



c. In the above example, how can you define kinases that have either a secretory signal peptide AND/OR a transmembrane domain(s)?

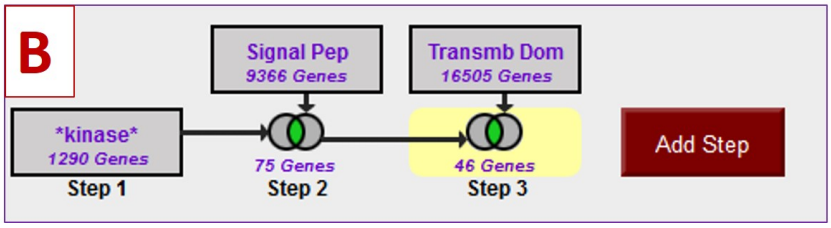
Hint: to do this properly you will have to employ the "Nested Strategy" feature. Nesting a strategy allows you to control the order in which your result sets are combined. Think about the difference between two mathematical equations.
Equation without nesting: 2 x 3 + 5 = 11
Equation with nesting: 2 x (3 + 5) = 16



Strategy Logic:

Strategy A returns kinases that have a signal peptide OR a TM domain OR both. (SP and/or TM)

Strategy B returns kinases that have a signal peptide AND a TM domain



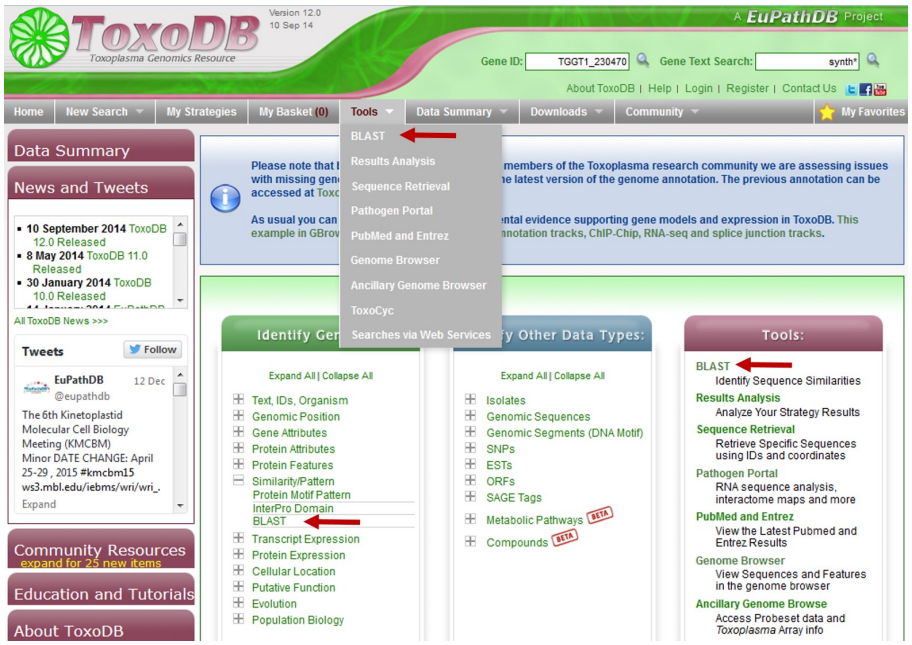
4. Finding a gene by BLAST Similarity.

Note: For this exercise use http://www.toxodb.org

a. Imagine that you generated an insertion mutant in *Toxoplasma* that is providing you with some of the most interesting results in your career! You sequence the flanking region and you are only able to get sequence from one side of the insertion (the sequence shown below). You immediately go to ToxoDB to find any information about this sequence. What do you do?

- aaaggagagaaagataaaaatatacaaaggtccccagagacacgatagtgttactgacaa catacagaatcaggtcgagcaatggaagaaccaagcaccggcgccagagattgaactcgc ttggattgccgtagcgttttatgagttgatagcttggctctaaaaaaacaaggctgaaaa atggaaaaaaatgtctccaat

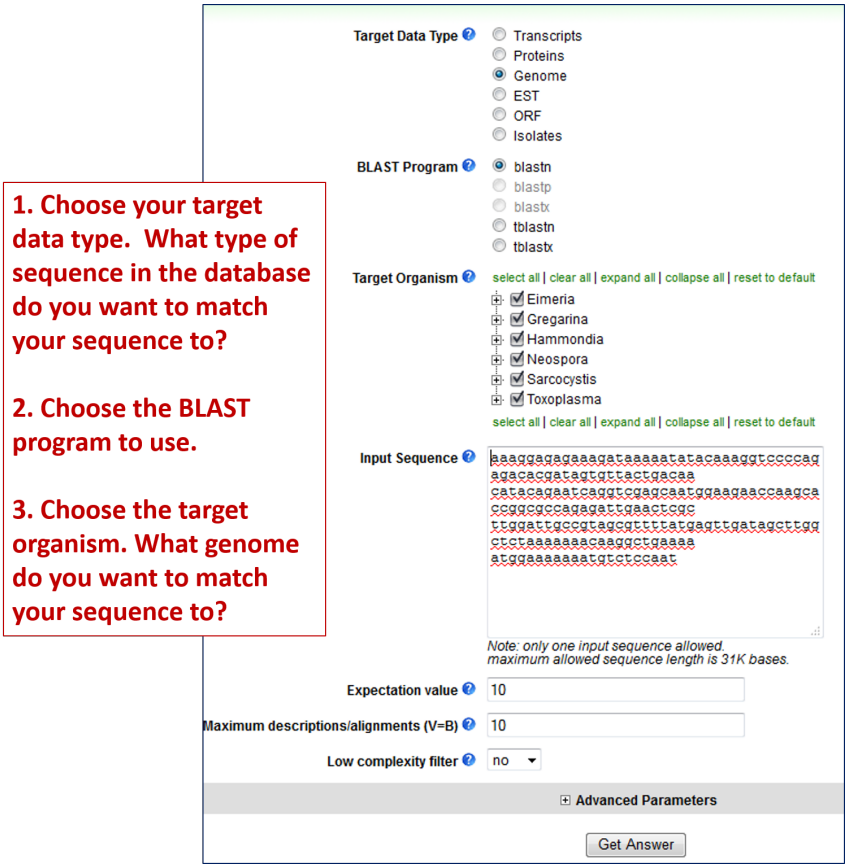
- Sequence is also available from this URL: **http://tinyurl.com/ex1blast**
- Try using the BLAST search with this sequence (hint: you can get to the BLAST tool by clicking on the BLAST link under tools on the home page).



- Which blast program should you use? (hint: try different combinations, just keep in mind that you have a nucleotide sequence so you have to use an appropriate BLAST program).

Note on BLAST programs:

- blastp compares an amino acid sequence against a protein sequence database;
- blastn compares a nucleotide sequence against a nucleotide sequence database;
- blastx compares the six-frame conceptual translation products of a nucleotide sequence (both strands) against a protein sequence database;
- tblastn compares a protein sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands);
- tblastx compares the six-frame translations of a nucleotide sequence against the six-frame translations of a nucleotide sequence database.



- Are you getting any results from blastx? tblastn? What about blastn?
- What is your gene? (hint: after running a blastn against *Toxoplasma* ME49 genomic sequence, click on the "link to the genome browser". In the genome browser zoom out to see what gene is in the area).

5. More BLASTing in EuPathDB (OPTIONAL).
Note: for this exercise use http://www.eupathdb.org

- The first thing we will need to do is get a sequence to use for BLAST. Search for the keyword "dihydrofolate" (without quotations). (Hint: use the Gene Text Search on the

upper right hand side of the EuPathDB home page).

- You should get multiple hits. Find the first one that is annotated as "dihydrofolate reductase-thymidylate synthase" (Look in the product description column).
- Once you find one, click on the gene ID and go to the gene page. It might be helpful to open the gene page in a new window or tab.
- Scroll down to the bottom of the page to the “Sequences” section.
- Copy the amino acid sequence and go back to EuPathDB (if you have not done so already, it might be helpful to open EuPathDB in a new window or tab).
- Go to the BLAST page from the EuPathDB home page. (hint: under Tools on the EuPathDB home page).
- Paste the amino acid sequence into the input window.
- Select target data type (start with "Proteins").
- Select BLAST Program. (Hint: BLASTP).
    - Expectation value : 10
    - Maximum descriptions/alignments (V=B) : 100
- Select the target organism. Click on "Get Answer".

a. Based on the results you should have identified excellent hits in almost all pathogens in EuPathDB but can you find good hits in *Giardia* or *Trichomonas*? *Let's try a different BLAST method:*

- Go back to the BLAST window. Change the target data type to Genome.
- Select the BLAST Program. Notice you cannot select BLASTP anymore. Try the other options. Notice how your input sequence type has to change when you select a different program. (Hint: TBLASTN is the one you need).
- Select all target organisms. Click on "Get Answer".

b. Note that the results are still missing a dihydrofolate from *Giardia* and *Trichomonas*. *Let's try a different BLAST method.*

- Go to your gene page window (in CryptoDB) and copy the nucleotide coding sequence.
- Go to the BLAST window and paste the nucleotide sequence into the input window.
- Select the target data type (try different ones) and the BLAST program. Notice you can only select TBLASTX or BLASTN when your input sequence is nucleotide. (Hint: select TBLASTX).
- Select the target organisms. This time let's specifically only select *Giardia* and *Trichomonas*. Click on "Get Answer".

Getting frustrated?

Not getting a hit for *Giardia* in this case is actually the correct answer! This organism does not have dihydrofolate reductase or thymidylate synthetase activity.

6. **Visiting a specific gene page.**
   Note: For this exercise use http://www.plasmodb.org

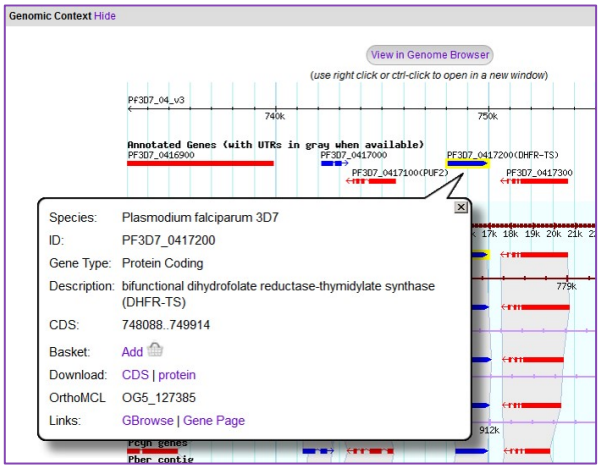
a. Find the gene page for one of the following *P. falciparum* genes and explore the information there to answer these questions.

1. bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS, PF3D7_0417200)
2. apical membrane antigen 1 gene (AMA1, PF3D7_1133400)

- How did you navigate to this gene? What other ways could you get there? I can think of 4 ways to reach the gene page)

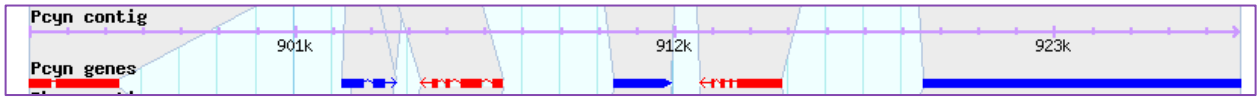
Look at the information on the gene page.

- What chromosome is this gene on?
- How many exons does this gene have? Hint: look at the graphic in the Genomic Context data track and mouse over the glyph representing the gene.
- What direction is the gene relative to the chromosome?
- How many nucleotides of coding sequence?
- Does this gene have a user comment?



b. What genes are located upstream & downstream of DHFR-TS (AMA1) in P. falciparum?

- Is synteny (chromosome organization) in this region maintained in other species? Hint: look in the genomic context section of the gene page – what does the shading mean?
- How complete is the genome assembly for other species? Each genome is displayed as two tracks – the genomic sequence (chromosome or contig) on top and the gene models underneath. Do the contigs contain gaps or truncations?



- What does synteny look like across the entire chromosome? To do this:
  - Click on the **"View in GBrowse"** button in the genomic context section.

- Zoom out to the entire chromosome. There are a few ways to do this – for example, drag your cursor across the entire chromosome then select "zoom" from the popup menu.
- Click on the tab called "Select tracks". Select the track called "Syntenic Sequences and Genes (Shaded by Orthology)". Go back to the Browser tab (this may take a minute to load).
- Which genome is composed of the most fragments? Are there any other interesting observations you can back by looking at synteny over large genomic regions?

c. Does the P. falciparum DHFR-TS (or AMA1) gene contain any single Nucleotide Polymorphisms (SNPs)?

SNPs are represented graphically in the genomic context section and also in a table called "SNP Overview". Using the SNP Alignment track you can view an alignment showing SNPs between specific strains/isolates.

- Examing the SNP track in the Genomic Context graphic. What do the different color diamonds represent? Mouse over the diamonds to get more information.
- What is the total number of SNPs in the gene?
- How many impact the predicted protein sequence?
- Is this likely to define the full spectrum of sequence variation in these particular strains?
- Compare the SNP characteristics of this gene to upstream and downstream genes. How do these results compare with SNP distribution in other genes?

d. Is the DHFR-TS (or AMA1) gene expressed?

Look at the gene page sections entitled "Protein" and "Expression". You may have to click on the **show** link to reveal the data associated with that data track.

- What kinds of data in PlasmoDB provide evidence for expression?
- Is this gene expressed at the protein level in salivary gland sporozoites? – in the blood stage phosphoproteome? Look at the Protein context graphic and the table of Mass Spec.-based Expression Evidence.
- How abundant is DHFR-TS (AMA1) protein? How confident are you of this analysis? Abundance can be estimated by counting the number of peptide spectra that map to a protein, or by using the RPKM value from RNA sequencing data.
- Look at the Expression data track labeled Life cycle expression data (3D7). At what life cycle stage is DHFR-TS (AMA1) most abundant? Does this make sense?
- Do the life cycle microarray expression profiles from different data tracks (and thus different experiments/data sets) give the same results? What tracks?
- What about RNA-sequence data, does it agree with microarray data? See these two data tracks – Strand specific transcriptomes of 4 life cycle stages; Transcriptomes of 7 sexual and asexual life stages.