Finding Genes, Building Search Strategies and Visiting a Gene Page

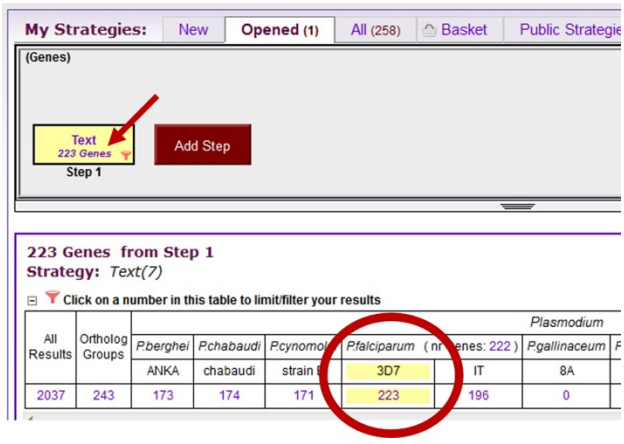
1. Finding a gene using text search.
   For this exercise use http://www.plasmodb.org

   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.



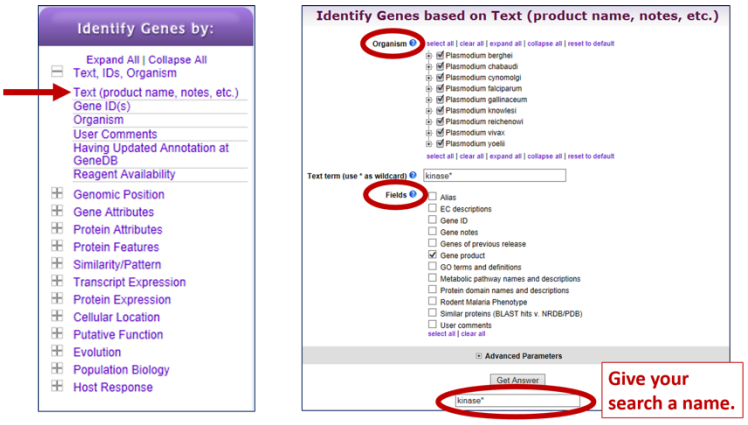
   - What happens if you search using the term **kinases** in the Gene Text Search box? How many results are returned?

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

   The search you ran in step 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.



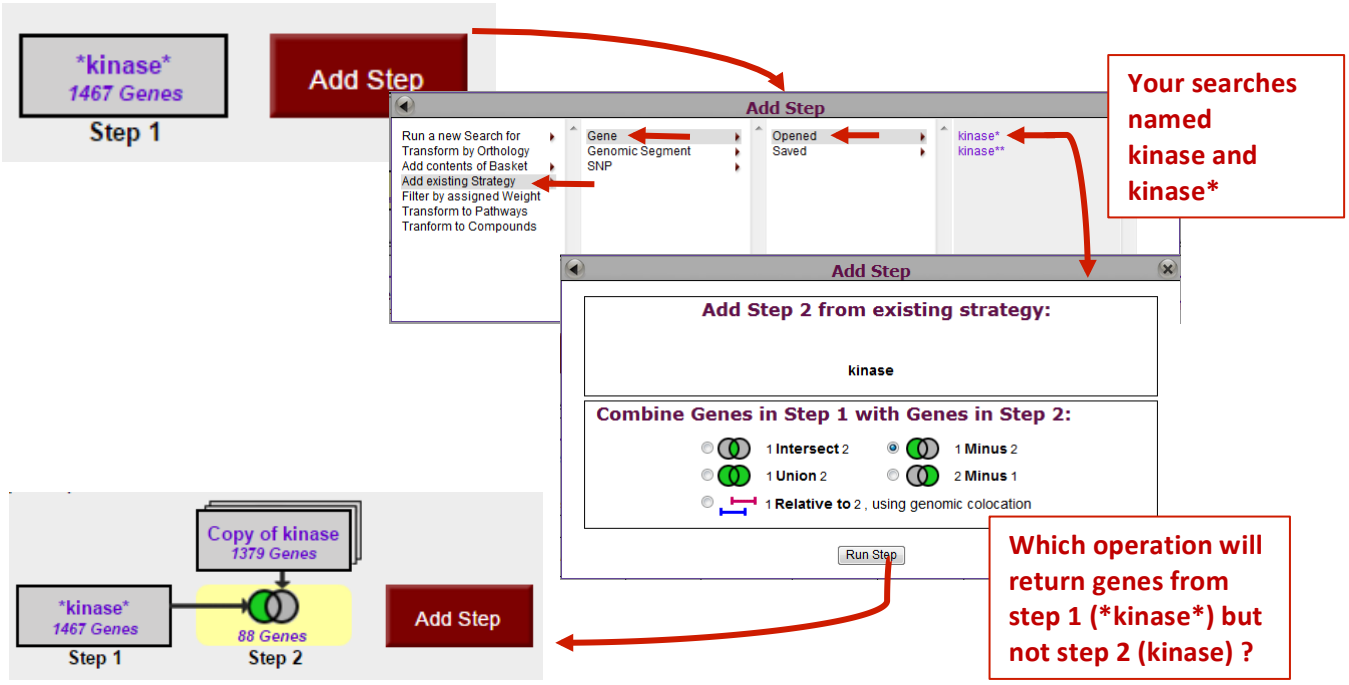
- How can you make sure to find your text term in plural form or in compound words like “kinases” or “6-phosphofructokinase”. Adding a wild card (wildcard = asterisk and means any character) in your search term will broaden your search. Use the full text search, the specific page where you can define 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. Notice that there is an extra asterisk at the end of an unsaved strategy name. The list of available searches will have an * at the end of the name.



- 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.

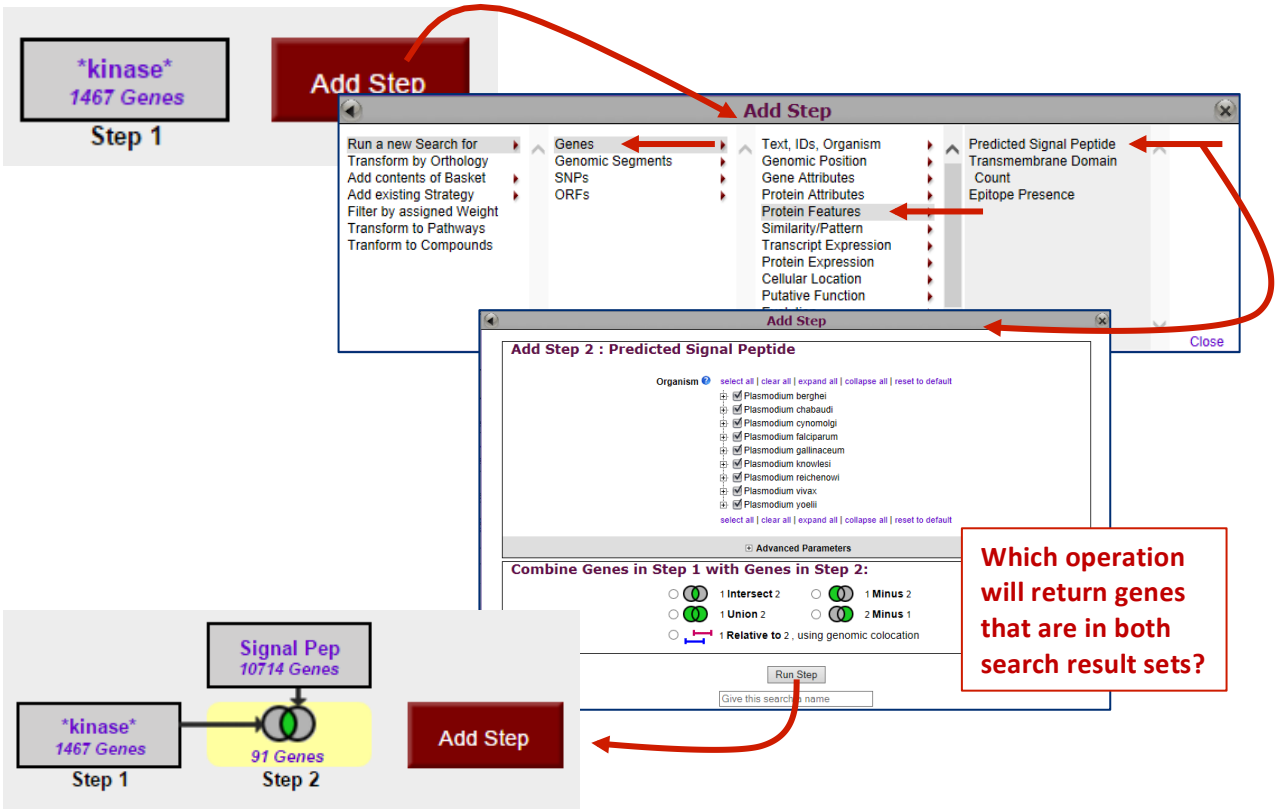
2. Combining text search results with results from other searches

a. Find kinase genes that are likely secreted.

In exercise 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. Navigate the Add Step popup to the Predicted Signal Peptide search. Choose intersect and click Run Step.

- How did you combine the search results?
- How many kinases are predicted to have a signal peptide?



Operator	:	Combined Result will contain:
1 INTERSECT 2	:	IDs in common between the two lists
1 UNION 2	:	IDs from list 1 and list 2
1 MINUS 2	:	IDs unique to 1
2 MINUS 1	:	IDs unique to 2
1 **Relative to** 2	:	IDs whose features are near each other (colocated) in the genome

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

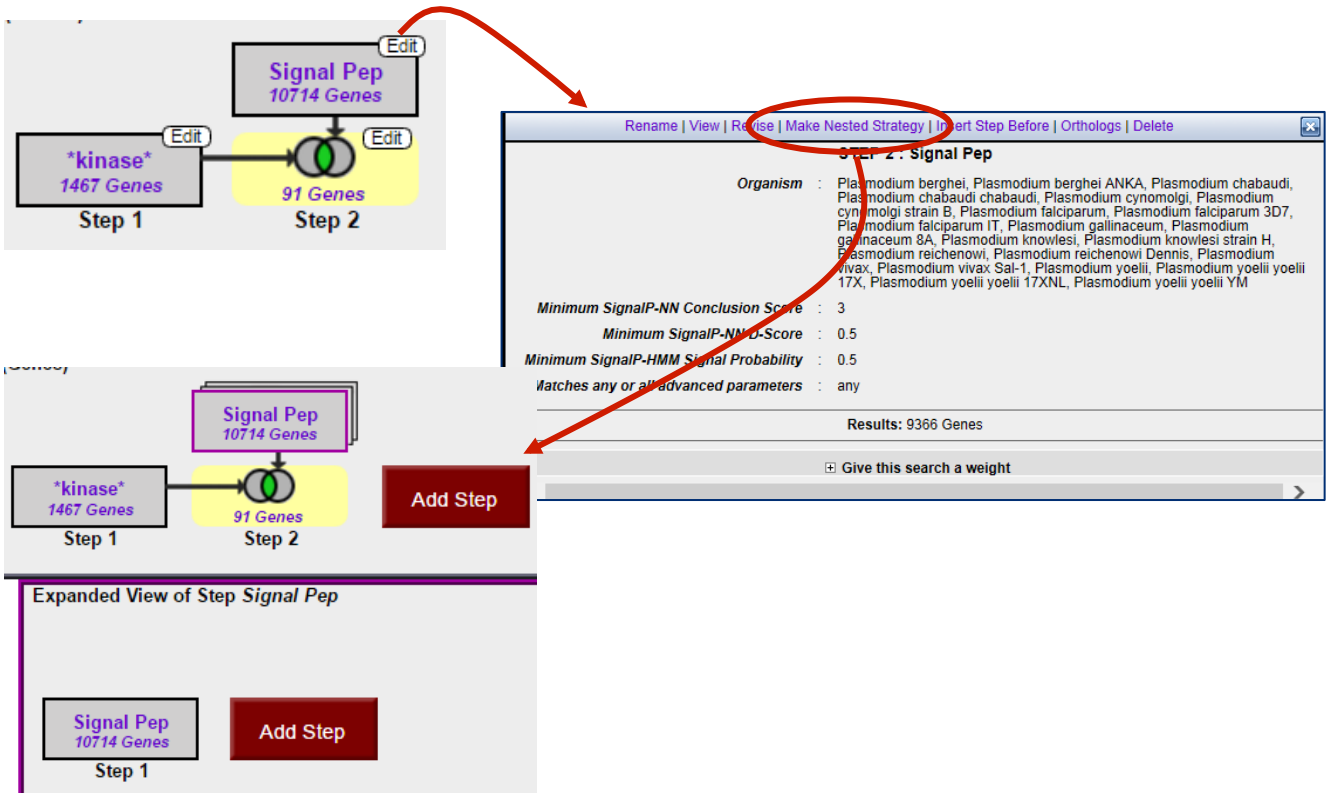
There is no wrong answer here!!

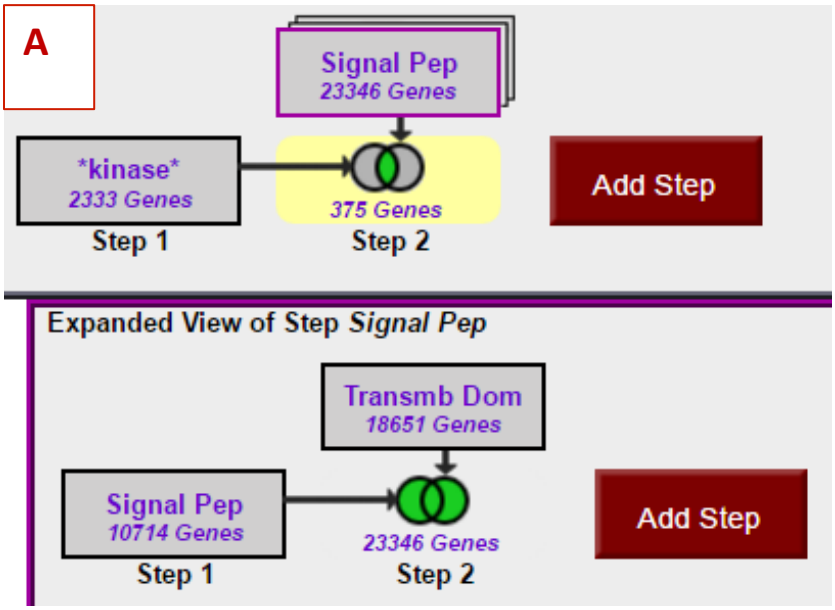
- From a biological standpoint what else would be interesting to know about these kinases? Add more searches to grow this strategy. Open the categories under Identify Genes By: on the home page and explore the types of searches that are available. You can reduce (or expand) your result set by adding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

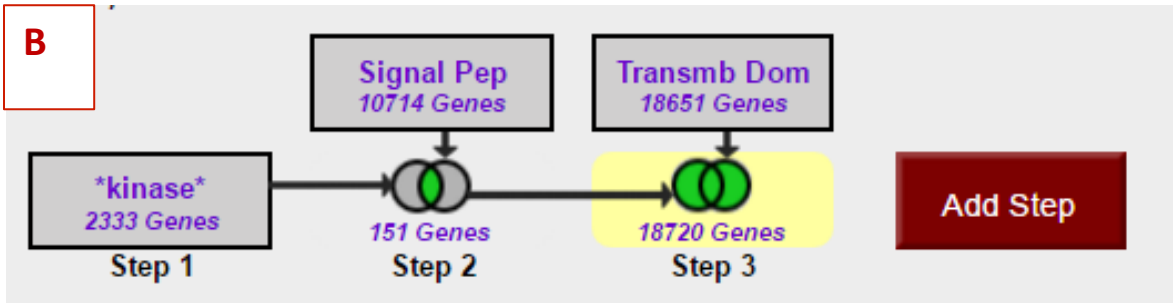




Strategy Logic:

Every gene returned by Strategy A will be a kinase. These kinases will have a signal peptide OR a TM domain OR both. (SP and/or TM; either or both)

Strategy B returns kinases that have a signal peptide as well as TM domain containing genes.



3. Finding a gene by BLAST Similarity.
Note: For this exercise start with http://www.toxodb.org

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

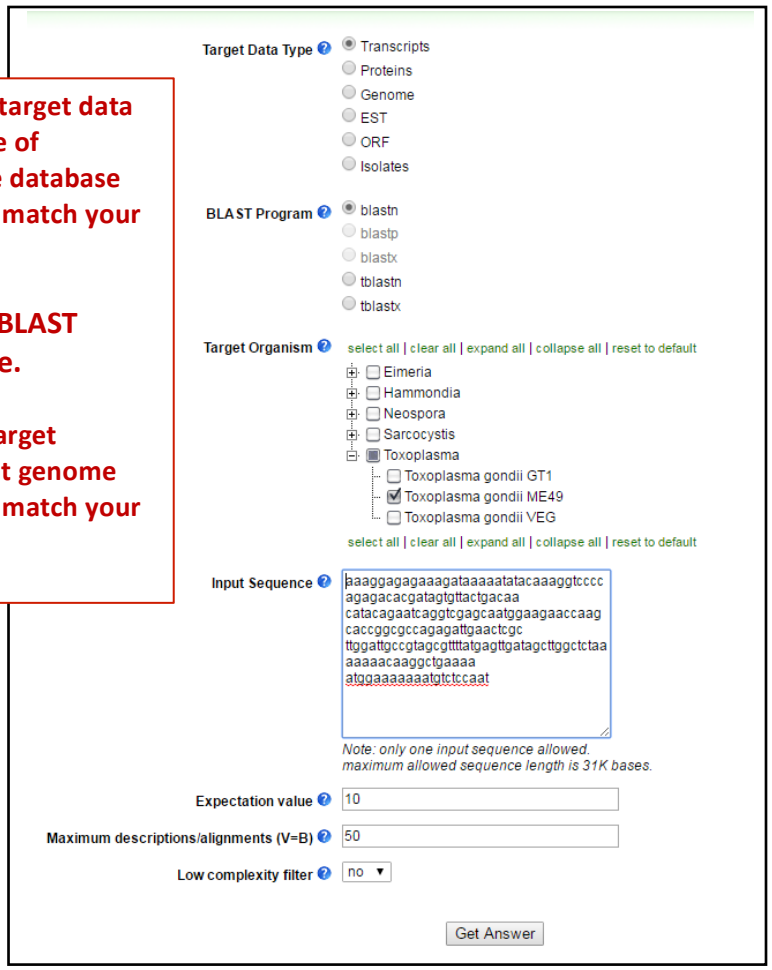
- Try using the BLAST search with this sequence

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

1. Choose your target data type. What type of sequence in the database do you want to match your sequence to?

2. Choose the BLAST program to use.

3. Choose the target organism. What genome do you want to match your sequence to?



Note on BLAST programs:

- blastn compares a nucleotide sequence against a nucleotide sequence database;
- blastp compares an amino acid sequence against a protein 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.

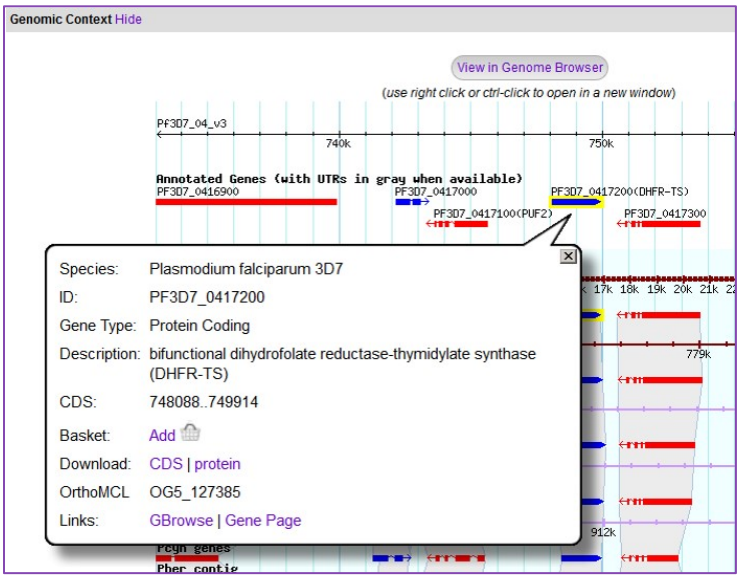
- Are you getting any results from blastx? tblastn? What about blastn?
- What is your gene? (hint: after running a blastn against *Toxoplasma* ME49 (Target organism) genomic sequence (Target Data Type), click on the "link to the genome browser". In the genome browser zoom out to see what gene is in the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

4. Viewing data on a gene page.
Note: For this exercise use http://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.

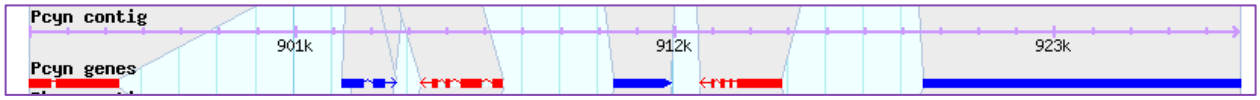
Look at the information on the gene page.
- What chromosome is this gene on?
- How many exons does this gene have? Hint: look at the graphic in the Genomic Context data track and mouse over the glyph representing the gene.
- What direction is the gene relative to the chromosome?
- How many nucleotides of coding sequence?
- Do you see a way to quickly download the coding and protein sequences?
- 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. Does the contig track contain gaps or truncations? What does this imply about the genome assembly?

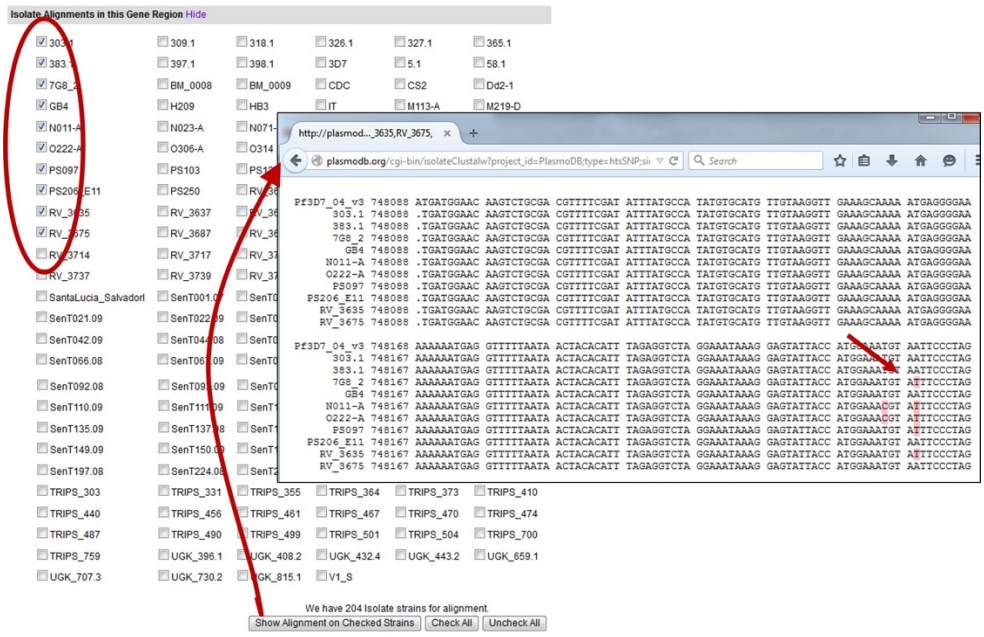


- What does synteny look like across the entire chromosome? To do this:
  - Click on the **"View in Genome Browser"** button in the genomic context section.
  - Zoom out to the entire chromosome. There are a few ways to do this. For example, drag your cursor across the entire chromosome in the Overview panel and 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 support by looking at synteny over large genomic regions?

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

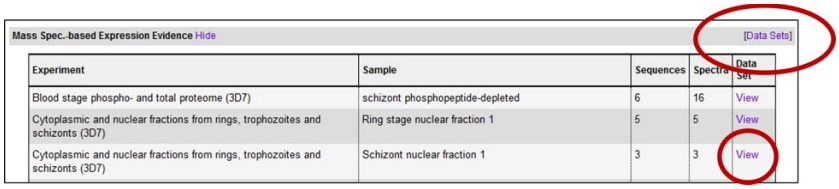
SNPs are represented in a table called "SNP Overview" and using the "Isolate Alignments in this Gene Region" track you can view an alignment showing SNPs between specific strains/isolates.

- Examine the SNP Overview table.
- What is the total number of SNPs in the gene?
- How many SNPs 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?
- Open the Isolate Alignments in this Gene Region data track and run an alignment between several isolates: 303.1, 383.1, 7G8_2, GB4, N011-A, O222-A, PS097, PS206_E11, RV_3635, RV_3675

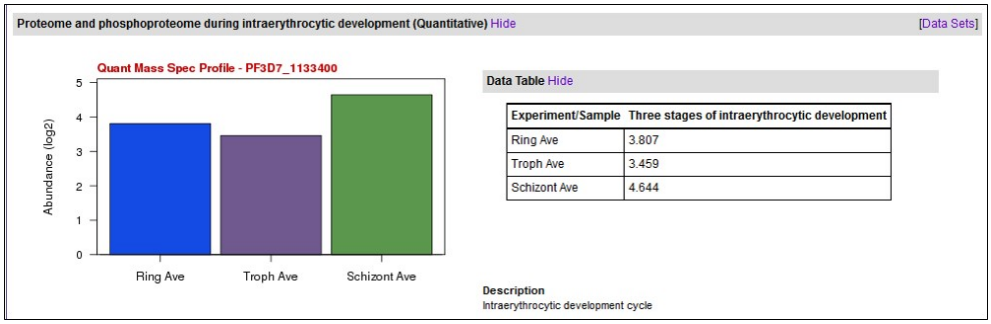


d. Is the DHFR-TS (or AMA1) gene expressed at the protein and/or transcript level?
Look at the gene page sections entitled "Protein" and "Expression". You may have to click on the **show** link to reveal the data associated with that data track.

- What kinds of data in PlasmoDB provide evidence for protein expression? Hint: open the Protein Features graphic, which is the first data track in the Protein section.
- 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.
- Can you quickly link to the data set record for proteomics experiments?



- How abundant is DHFR-TS (AMA1) protein? How confident are you of this analysis? Abundance can be estimated by counting the number of spectra supporting a peptide spectra that maps to the protein. Where do you find information about the number of spectra?
- Is the protein more abundant in the ring or schizont life cycle stage? Hint: open the quantitative proteomics track called **Proteome and phosphoproteome during intraerythrocytic development (Quantitative)**.



(PF3D7_1133400 (AMA1))

- Look at the Expression data track labeled Life cycle expression data (3D7). Based on this data, 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 did you use?
- 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.