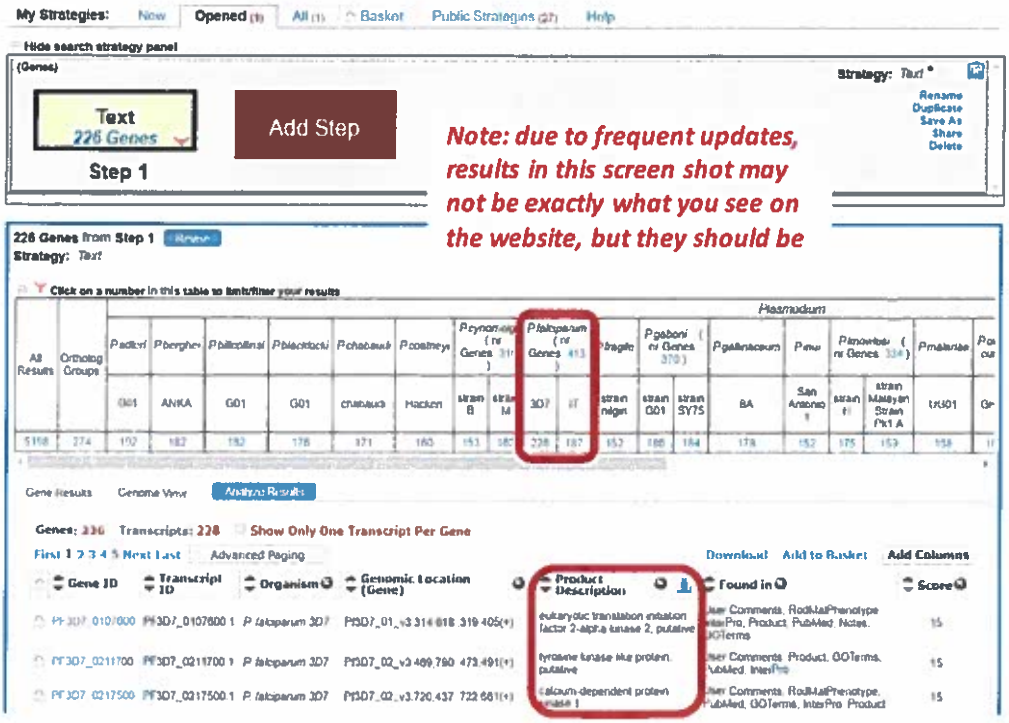
Finding Genes, Building Search Strategies and Visiting a Gene Page

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.

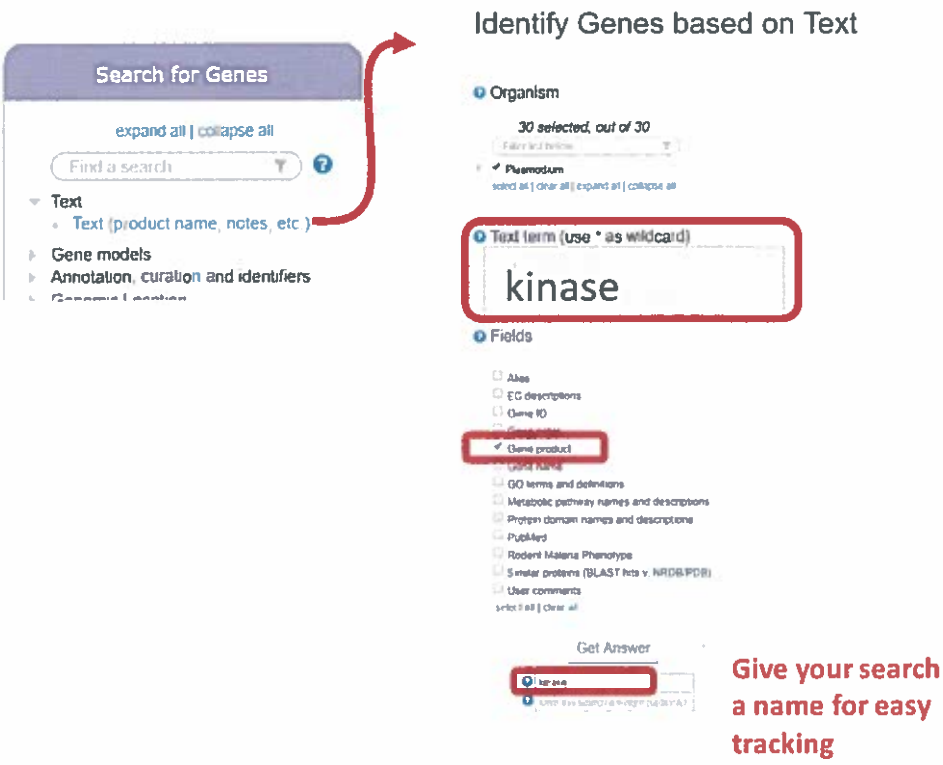
- How many genes did you get?
- Look closely at the sections of the result page. How many of those are in *P. falciparum*? How did you find this out? (*Hint* – the filter table is located between the strategy panel and the result table and shows the distribution of results across the organisms that you searched. Click on a number to 'filter' the result and display results from a specific species or strain).



- Do you believe that these genes are kinases? Find the Product Description in the Gene Result tab. Can you presume the gene encodes a kinase just by looking at the name?
- 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. Note that you can also revise the search from step 1a and configure the search parameters as described below.



- There are several ways to navigate to the **Identify Genes based on Text** page: home page 'Search for Genes' panel and the 'New Search' drop down menu. 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

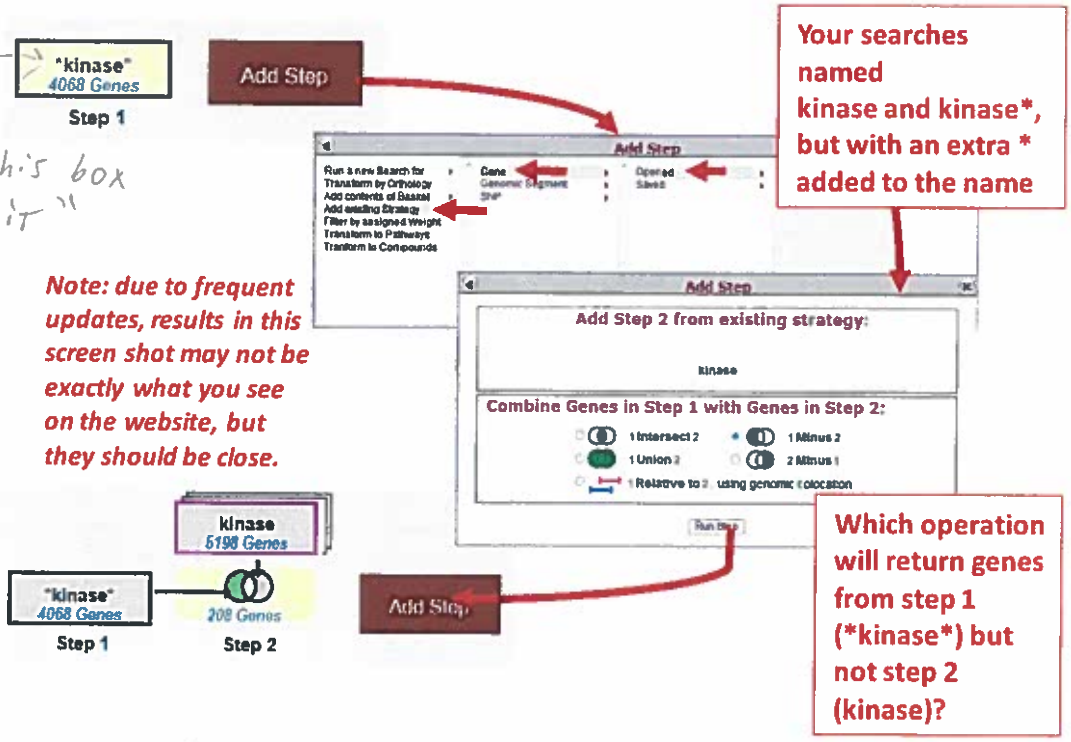
text search, the specific page where you can define the fields to be searched (Fields = Gene Product).

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.** must have both on Strategy Panel
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.

You can name this box by clicking "edit" and then the "rename" tab



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

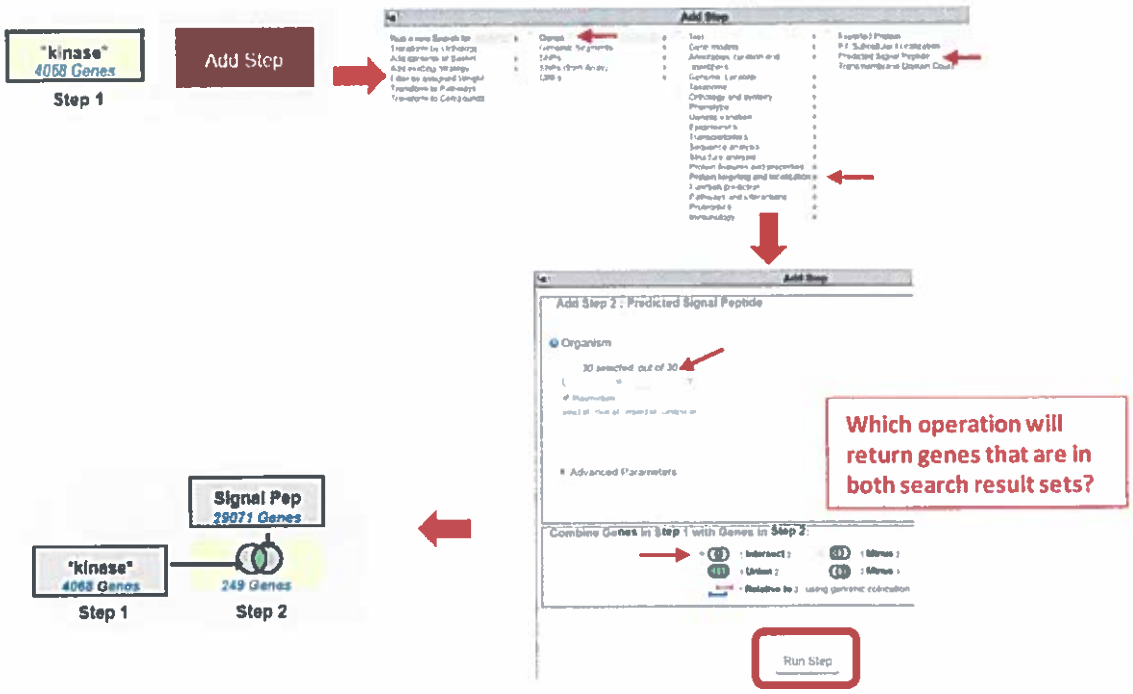
2. Combing text search results with results from other searches

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. 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?



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.

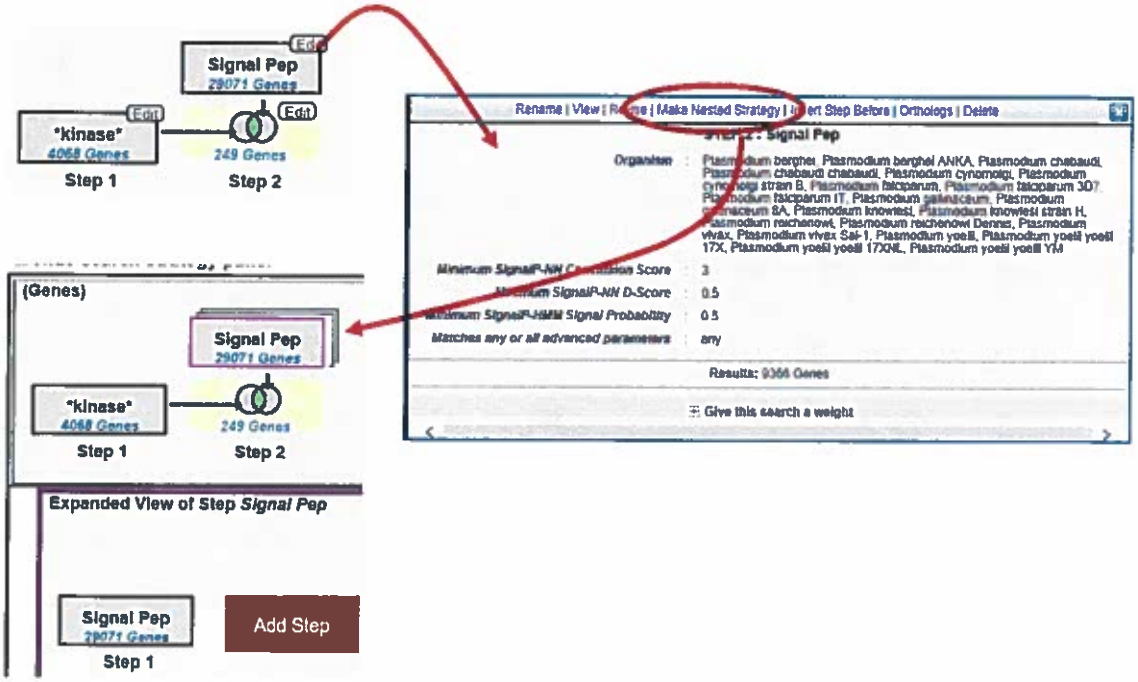
There is no wrong answer here!!

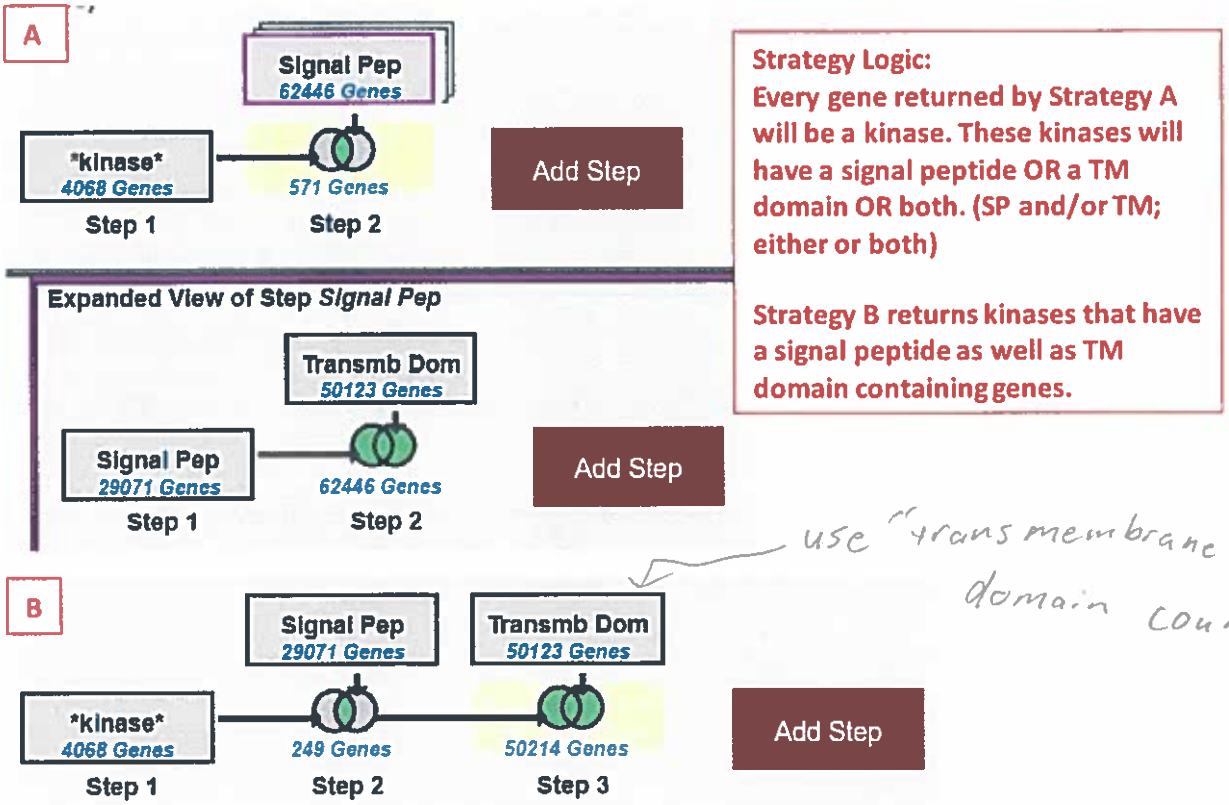
- From a biological standpoint what else would be interesting to know about these kinases? Add more searches to grow this strategy. Open the categories under Identify Genes By: on the home page and explore the types of searches that are available. You can reduce (or expand) your result set by adding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





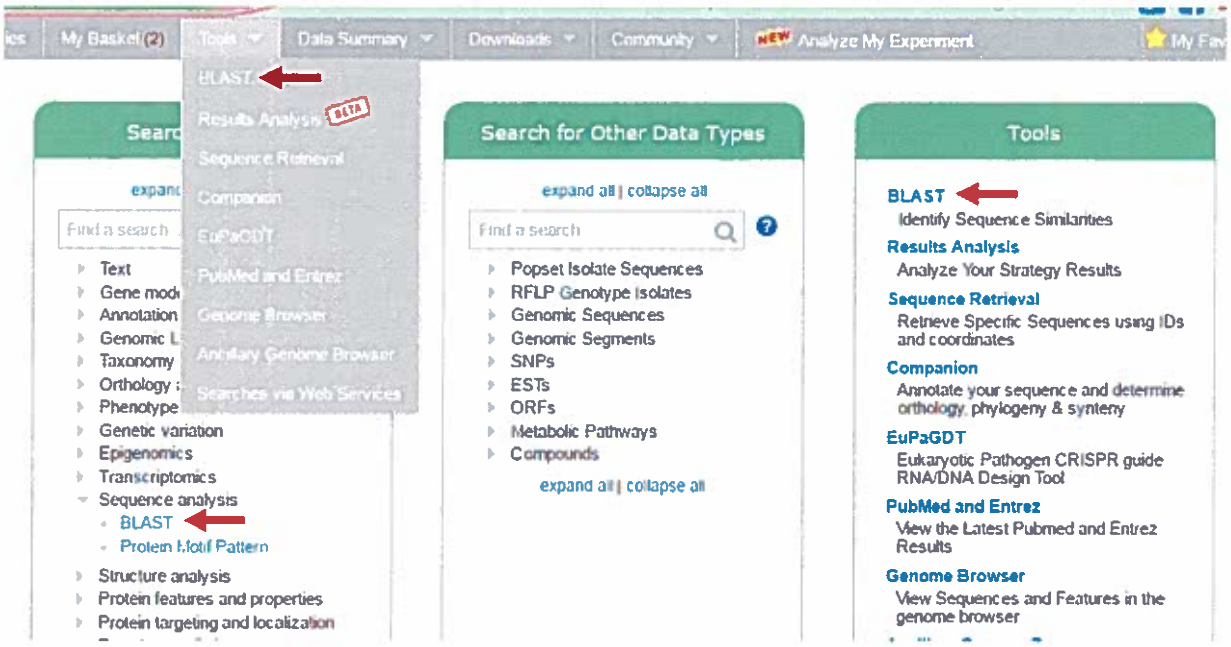
3. Finding a gene by BLAST Similarity.

Note: For this exercise start with http://toxodb.org/toxo/

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

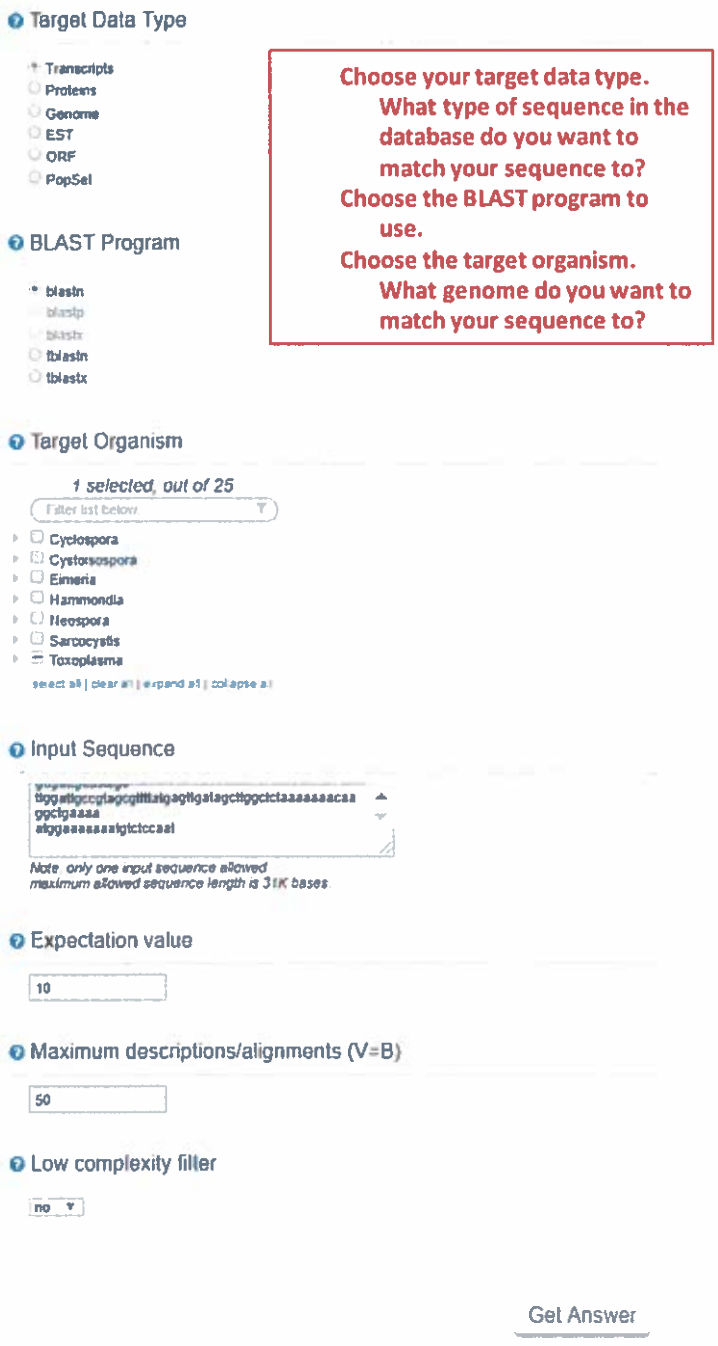
- Sequence is also available from this URL: http://tinyurl.com/ex1blast
- Navigate to the BLAST search and run the search with this sequence. The BLAST search will return records for sequences that are similar to your input sequence.



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

Note on BLAST programs:

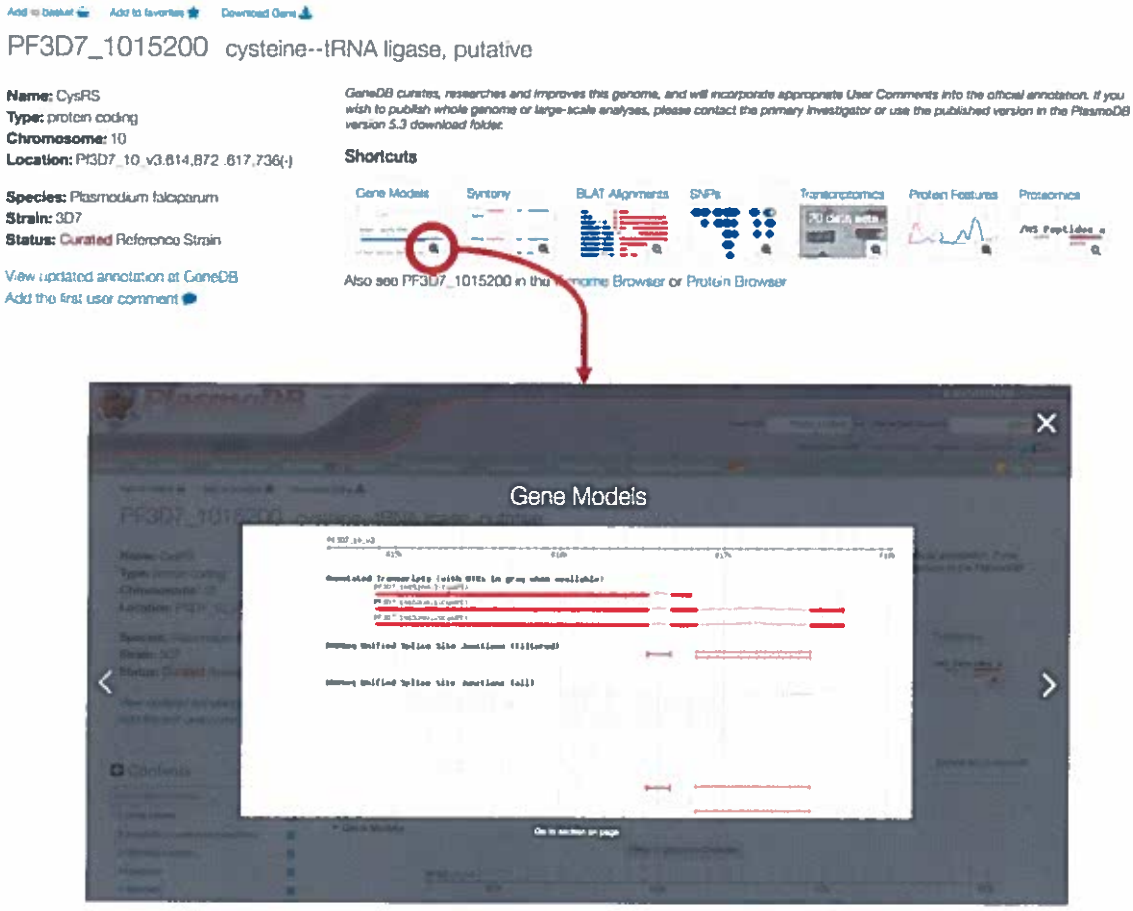
- blastp compares an amino acid sequence against a protein sequence database;
- blastn compares a nucleotide sequence against a nucleotide sequence database;
- blastx compares the six-frame conceptual translation products of a nucleotide sequence (both strands) against a protein sequence database;
- 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.



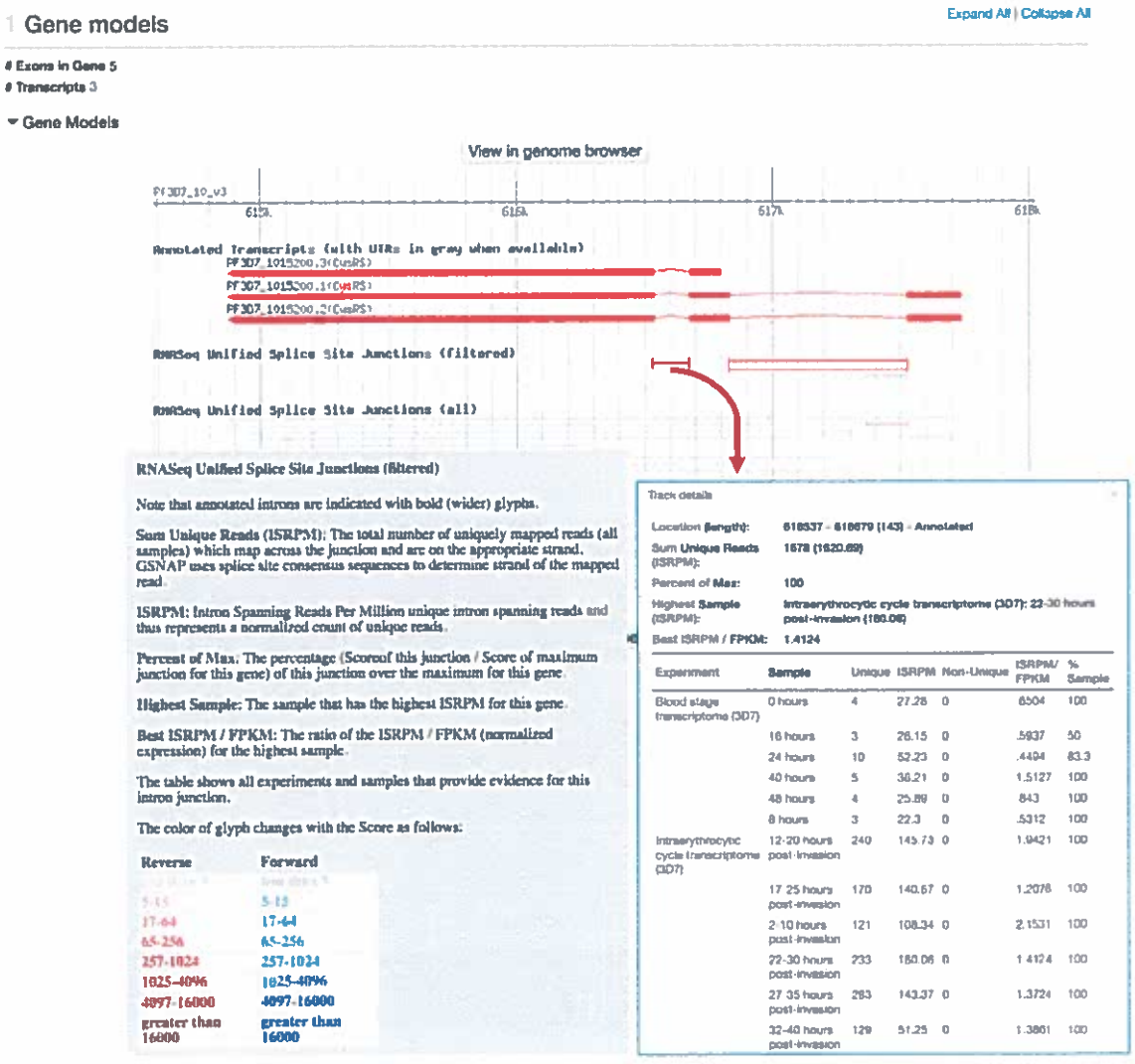
- Are you getting any results from blastx? tblastn? What about blastn?
- What is your gene? (hint: after running a blastn against *Toxoplasma* ME49 (Target organism) genomic sequence (Target Data Type), click on the "link to the genome browser". In the genome browser zoom out to see what gene is in the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

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

a. Find the gene page for cysteine-tRNA ligase (PF3D7_1015200).
- There are several ways to navigate to the gene page using either the gene ID or the gene product name How did you navigate to this gene? What other ways could you get there?
- Examine the information at the top of the gene page:
  - What is the gene name?
  - What chromosome is this gene on?
- Explore the "shortcuts" section at the top of the gene page – try clicking on the magnifying glass. This option opens up a preview of various sections of the gene page for quick access. Clicking the image itself will take you to that section of the gene page.



- Examine the "Gene Models" section of the gene page.
  - How many exons does this gene have?
  - How many transcripts does this gene encode?
  - What direction are the transcripts relative to the chromosome?
  - What does the "splice site junctions" information mean?
  - From what type of data are the "splice site junctions" determined?
  - How many nucleotides is the largest transcript? (hint: examine the transcripts table underneath the gene models).



1 Gene models

Expand All | Collapse All

Exons in Gene 5

Transcripts 3

▾ Gene Models

View in genome browser

PF3D7_10_v3

Annotated Transcripts (with UTRs in gray when available)

RNASeq Unified Splice Site Junctions (filtered)

RNASeq Unified Splice Site Junctions (all)

RNASeq Unified Splice Site Junctions (filtered)

Note that annotated introns are indicated with bold (wider) glyphs.

Sum Unique Reads (ISRPM): The total number of uniquely mapped reads (all samples) which map across the junction and are on the appropriate strand. GSNAP uses splice site consensus sequences to determine strand of the mapped read.

ISRPM: Intron Spanning Reads Per Million unique intron spanning reads and thus represents a normalized count of unique reads.

Percent of Max: The percentage (Score of this junction / Score of maximum junction for this gene) of this junction over the maximum for this gene.

Highest Sample: The sample that has the highest ISRPM for this gene.

Best ISRPM / FPKM: The ratio of the ISRPM / FPKM (normalized expression) for the highest sample.

The table shows all experiments and samples that provide evidence for this intron junction.

The color of glyph changes with the Score as follows:

Reverse	Forward
5-16	5-16
17-64	17-64
65-256	65-256
257-1024	257-1024
1025-4096	1025-4096
4097-16000	4097-16000
greater than 16000	greater than 16000

Track details

Location (length): 618537 - 618679 (143) - Annotated

Sum Unique Reads (ISRPM): 1678 (1920.69)

Percent of Max: 100

Highest Sample (ISRPM): Intraerythrocytic cycle transcriptome (3D7): 22-30 hours post-invasion (180.06)

Best ISRPM / FPKM: 1.4124

Experiment	Sample	Unique	ISRPM	Non-Unique	ISRPM/ FPKM	% Sample
Blood stage transcriptome (3D7)	0 hours	4	27.28	0	6504	100
	16 hours	3	26.15	0	.5937	50
	24 hours	10	52.23	0	.4494	83.3
	40 hours	5	36.21	0	1.5127	100
	48 hours	4	25.89	0	843	100
	8 hours	3	22.3	0	.5312	100
Intraerythrocytic cycle transcriptome (3D7)	12-20 hours post-invasion	240	145.73	0	1.9421	100
	17-25 hours post-invasion	170	140.67	0	1.2078	100
	2-10 hours post-invasion	121	108.34	0	2.1531	100
	22-30 hours post-invasion	233	180.06	0	1.4124	100
	27-35 hours post-invasion	283	143.37	0	1.3724	100
	32-40 hours post-invasion	129	51.25	0	1.3861	100

b. What does the synteny of this gene look like? How did you find/navigate to this section? (hint: you can use the "Contents" menu on the left side of the gene page to find/navigate to the different sections. You can also click on the images in the Shortcuts section to navigate to the image within the data section of the page).

- Is synteny (chromosome organization) in this region maintained in other species? Hint: compare gene organization between the different species in the synteny section.
- What does the shading between genes indicate?
- What does synteny look like across the entire chromosome? To do this:
  - Click on the **"View in Genome Browser"** button right under the synteny section on the gene page.

  - 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 (this may take a minute to load).
  - For each genome notice that there are two tracks: one called genes and the other contig. 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 this gene contain Single Nucleotide Polymorphisms (SNPs)? (return to the gene page using the browser back button)

In gene pages, SNPs are represented in a section called "Genetic variation". This section includes an isolate alignment tool for displaying SNPs between chosen isolates and a DNA polymorphism browser with textual and graphical SNP representations.

- Examine the DNA polymorphism section 9.2.
  - 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 this gene?

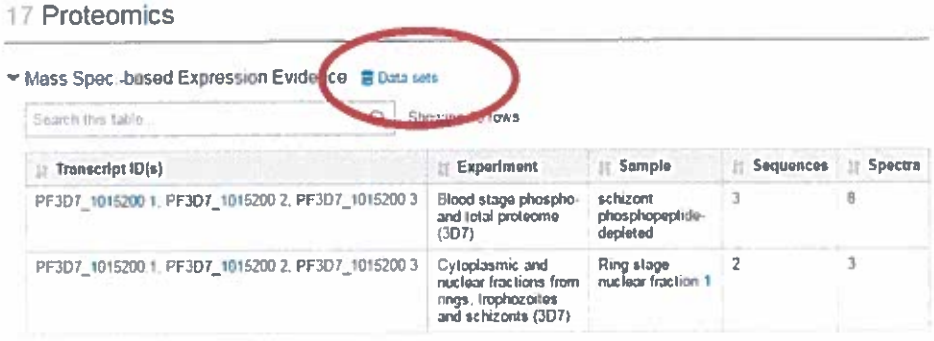
- What do the different color diamonds in the browser view signify? (Hint: move your cursor over a diamond – without clicking - to get more information in a popup).

- Compare Specific isolates to each other:
  - Using the isolate alignment tool, run an alignment between several isolates: 303.1, 383.1, 7G8, GB4, N011-A, O222-A, PS097, PS206_E11, RV_3635, RV_3675
  - This tool can produce a multiple sequence alignment of all isolates or a subset of isolates. Use the Select strains feature to choose an isolate quality from the left panel and then use the right side panel to define the range of the quality. The 'Parasite Organism' quality allows you to choose individual isolates.
  - What do Ns indicate?

```
Pf3D7_10_v3 600512 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
      303.1 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
      383.1 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATANN
      7G8_2 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
        GB4 600511 NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
     N011-A 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
     O222-A 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
      PS097 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
  PS206_E11 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTN TTAGNNNNTN TNTATATANN TATATATATA
    RV_3635 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
    RV_3675 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
```

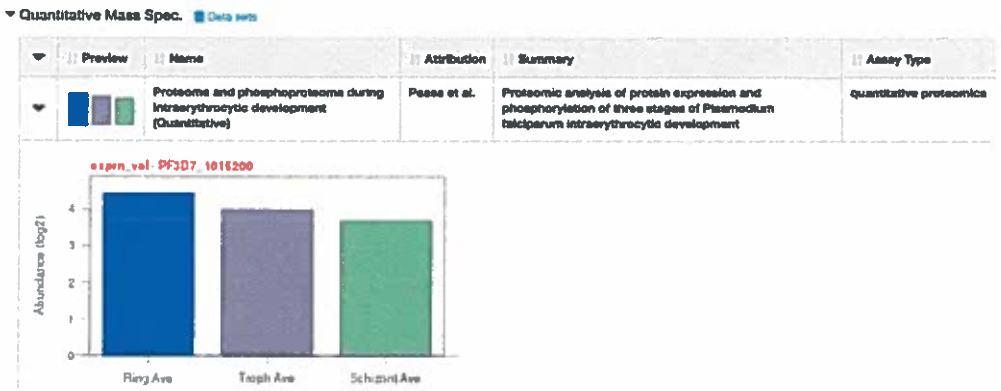
d. Is this gene expressed at the protein and/or transcript level?
Look at the gene page sections entitled "Proteomics" and "Transcriptomics". You can use the contents panel to navigate to those sections. Or you can return to the top of the page with the 'back to top button' and then click on the 'Shortcut' image to navigate to that section of the page.

- What kinds of data in PlasmoDB provide evidence for protein expression? (Hint, view the Mass Spec.-based Expression Evidence table).
- Is this gene expressed at the protein level in salivary gland sporozoites?
- Does is contain any post-translational modifications?
- Can you quickly link to the data set record for proteomics experiments?

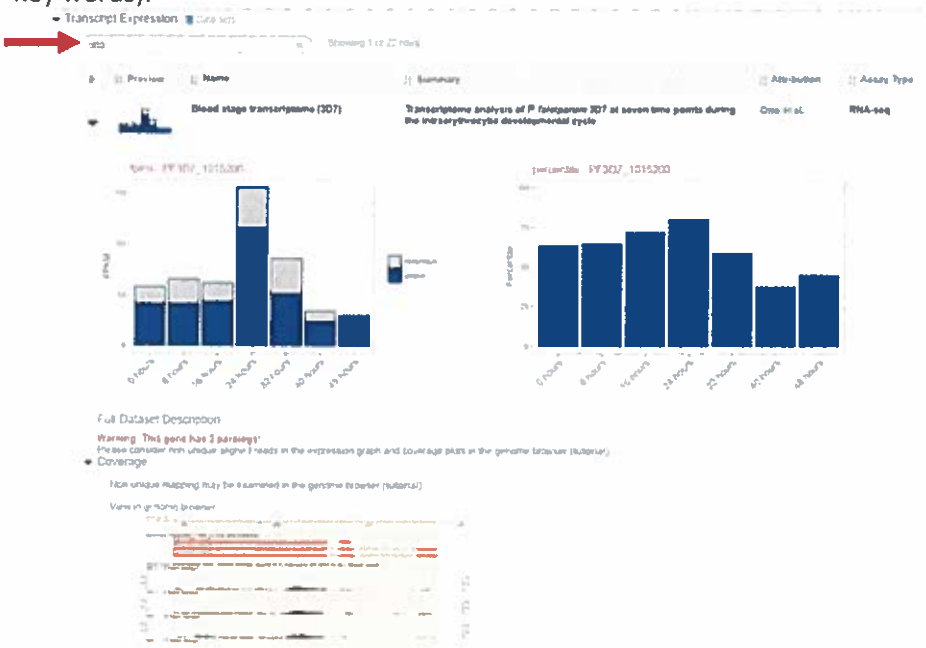


Some answers are in Proteomics, others are in transcriptomics

- How abundant is this protein? How confident are you of this analysis? Abundance can be estimated by counting the number of spectra supporting a peptide. 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)**.
- Look at the Expression data track labeled **Life cycle expression data (3D7)**. Based on this data, at what life cycle stage is this protein most abundant?

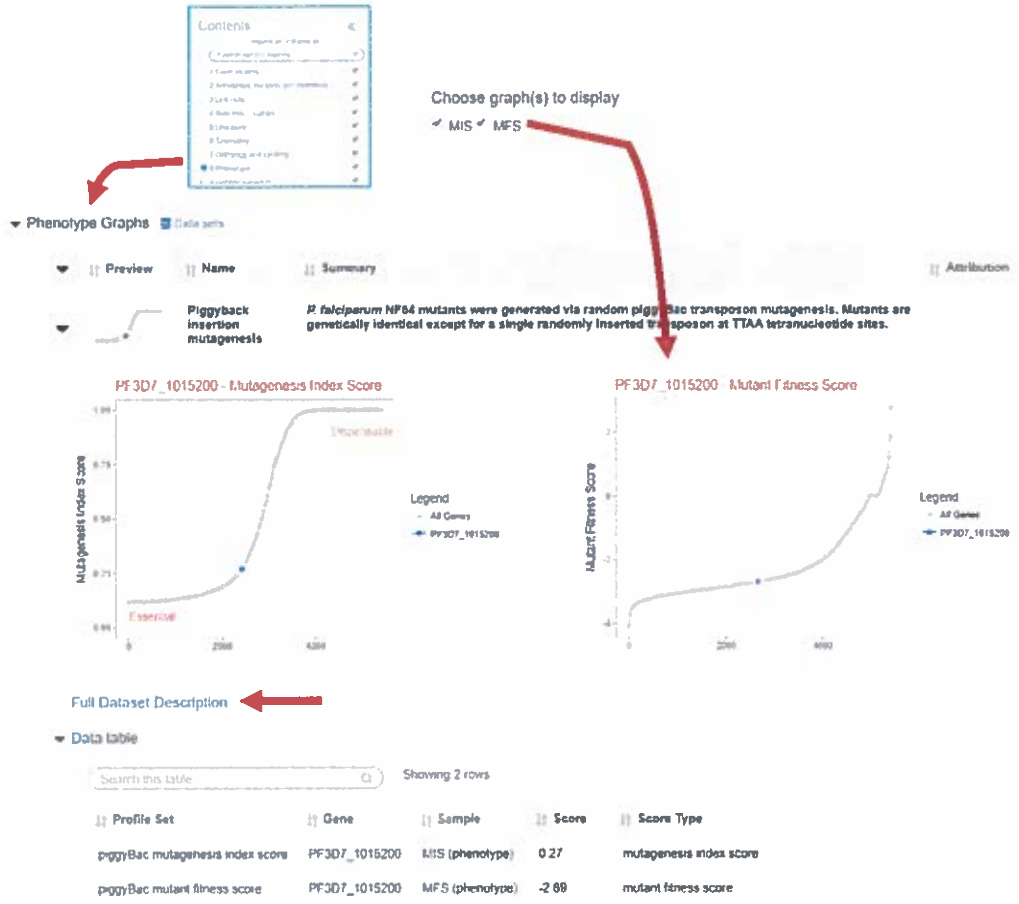


- Does the proteomic data agree with the available transcriptomic data? (Hint, navigate to the transcriptomic section – remember you can use the contents table on the left side of your screen).
- Find the RNAseq experiment by Otto et al. Where is this gene most highly expressed? How did you find this experiment? (Hint, you can search the transcriptomic table with key words).



- How does the RNAseq data compare with the microarray data?
- What does the polysomal data look like?

e. Is cysteine-tRNA ligase essential to Plasmodium? Does mutating the gene reduce fitness?
   - Navigate to the phenotype section and notice the Piggyback insertion metagenesis data in the Phenotype Graphs. The section opens with the Mutagenesis Index Score (MIS) graph displayed. Turn on the Mutational Fitness Score graph too by checking the box next to MFS in the Choose graphs to display (below the description and axis labels).



   - Explore the data and data descriptions in to gain understanding of the data and its meaning. Visit the data description page for an overview of the data set, links to the publication, etc.
   - What are the MIS and MFS scores for this gene? open data table
   - How do these scores compare to scores for the rest of the genome?
   - What do these scores mean?

- How do the MIS and MFS scores for other known essential genes such as PF3D7_0417200: bifunctional dihydrofolate reductase-thymidylate synthase, or PF3D7_1343500 conserved Plasmodium protein, unknown function? (Hint: visit their gene pages and compare their scores with our gene)
- How does it compare to PF3D7_1343700 kelch protein K13?