

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).



2896 Genes from Step 1 [Revise](#)

Strategy: Text

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	Plasmodium																											
		Paderi	Pbergher	Pbillcollinsi	Pblacklocki	Pchabaudi	Pcoatneyi	Pcynomolgi (310)		P.falciparum (2896)										Pfragile	Pgabori (363)	Pgallina							
		G01	ANKA	G01	G01	chabaudi	Hackeri	strain B	strain III	3D7	7G8	CD01	Dd2	GA01	GB4	GN01	HB3	IT	KE01	KH01	KH02	ML01	SD01	SN01	TG01	strain nigin	strain G01	strain SY75	8A
7746	277	189	179	179	176	168	155	151	159	123	177	177	177	178	180	177	178	178	177	177	180	182	176	179	187	149	182	181	17

Gene Results [Genome View](#) [Analyze Results](#)

Rows per page: 20

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Found in	Score
PF3D7_0107600	PF3D7_0107600.1	P. falciparum 3D7	PI3D7_01_v3:314.618..319.405(+)	eukaryotic translation initiation factor 2-alpha kinase 2, putative	User Comments, RodMaIPhenotype, InterPro, Product, PubMed, Notes, GOTerms	13
PF3D7_0211700	PF3D7_0211700.1	P. falciparum 3D7	PI3D7_02_v3:469.790..473.491(+)	tyrosine kinase-like protein, putative	User Comments, Product, GOTerms, PubMed, InterPro	13
PF3D7_0217500	PF3D7_0217500.1	P. falciparum 3D7	PI3D7_02_v3:720.437..722.661(+)	calcium-dependent protein kinase 1	User Comments, Product, PubMed, GOTerms, InterPro, RodMaIPhenotype	13

- 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 Text Search form - **Identify Genes based on Text**, allows you to configure the search yourself, choosing parameters that best meet your needs. Use the Text 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.

Identify Genes based on Text

Search for Genes

expand all | collapse all

Find a search...

Text

- Text (product name, notes, etc.)
- Gene models
- Annotation, curation and identifiers
- Genomic Location

45 selected, out of 45

Filter list below...

Plasmodium

select all | clear all | expand all | collapse all

Text term (use * as wildcard)

kinase

Fields

- Alias
- EC descriptions
- Gene ID
- Gene notes
- Gene product
- Gene name
- GO terms and definitions
- Metabolic pathway names and descriptions
- Protein domain names and descriptions
- PubMed
- Rodent Malaria Phenotype
- Similar proteins (BLAST hits v. NRDB/Protein Data Bank)
- User comments

select all | clear all

Get Answer

kinase

Give this search a weight (optional)

Give your search a name for easy tracking

- 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-phosphofruktokinase”? 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. Searches can also be saved using Save As in the strategy action items.
- 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 correct 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.

Your searches named kinase and kinase*, but with an extra * added to the name

Which operation will return genes from step 1 (*kinase*) but not step 2

- 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. 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. For the second search choose **Identify Genes based on Protein Targeting, Predicted Signal Peptide**. You can find this by navigating through Genes -> Protein targeting and localization -> Predicted Signal Peptide.

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

Which operation will return genes that are in both search result sets?

Run Step

Signal Pep 44582 Genes

kinase 6388 Genes

453 Genes

Add Step

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!!

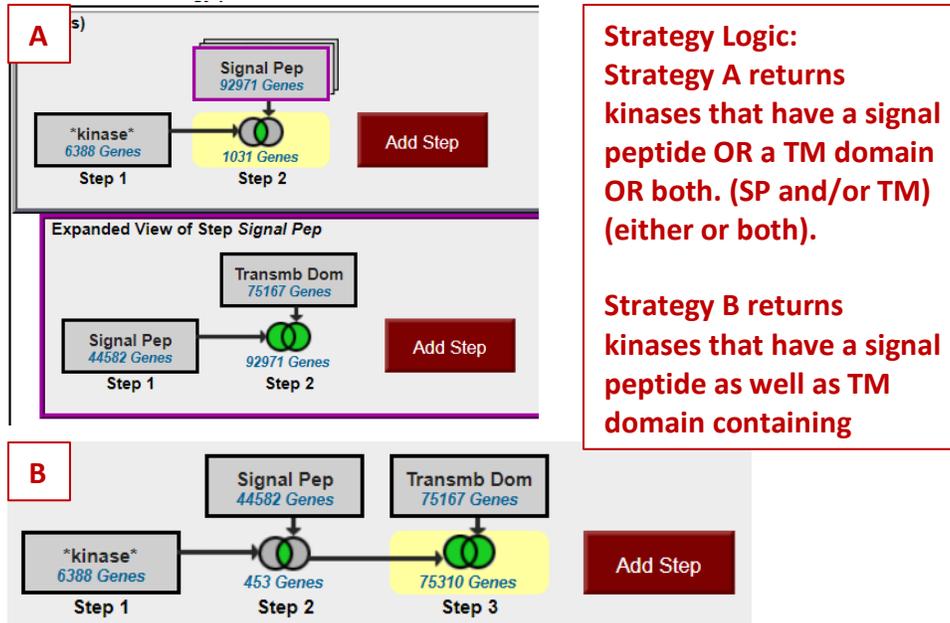
- From a biological standpoint what else would be interesting to know about these kinases? Add more searches to grow this strategy. Open the categories under Identify Genes By on the home page and explore the types of searches that are available. You can reduce (or expand) your result set by adding searches 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: $2 \times 3 + 5 = 11$
 Equation with nesting: $2 \times (3 + 5) = 16$

The image shows a bioinformatics strategy builder interface. At the top, a strategy diagram consists of two steps: Step 1 is a search for '*kinase*' (6388 Genes) and Step 2 is a search for 'Signal Pep' (44582 Genes). An arrow points from Step 1 to Step 2, and a red circle highlights the 'Signal Pep' step. A red arrow points from this circle to a detailed parameter window for the 'Signal Pep' step. This window lists various parameters such as 'Organism', 'Minimum SignalP-NN Conclusion Score', 'Minimum SignalP-NN D-Score', 'Minimum SignalP-HMM Signal Probability', and 'Matches any of the advanced parameters'. Below the strategy diagram, an 'Expanded View of Step Signal Pep' shows the step in isolation, with a red arrow pointing from the 'Signal Pep' step in the main diagram to this expanded view.

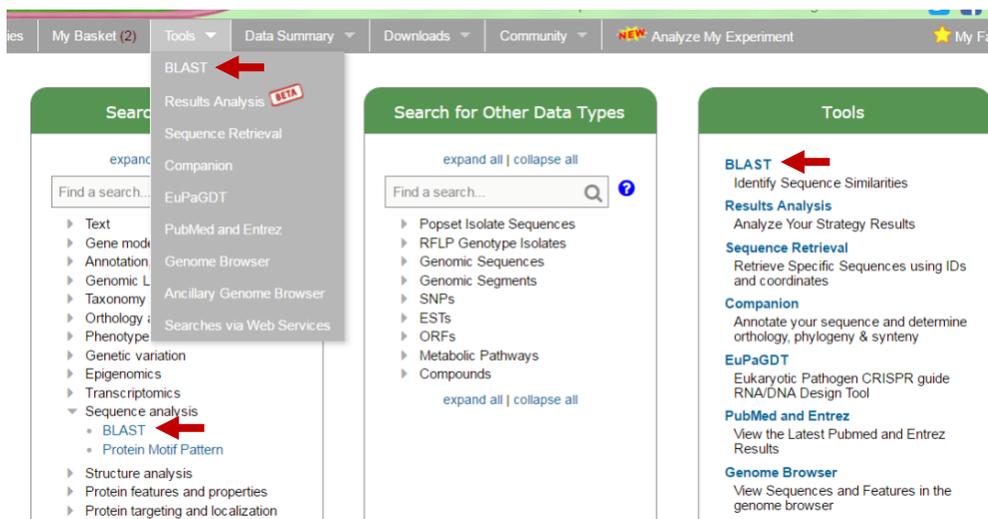


3. Finding a gene by BLAST Similarity.

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?

```
aaaggagagaaagataaaaaatacaaaaggtccccagagacacgatagtgttactgaca
catacagaatcaggtcgagcaatggaagaaccaagcaccggcgccagagattgaactcgc
ttggattccgtagcgtttatgagttgatagcttgctctaaaaaaacaaggctgaaa
atggaaaaaatgtctccaat
```



- Sequence is also available from this URL: <http://tinyurl.com/ex1blast>

- Explore the “shortcuts” section at the top of the gene page – try clicking on the magnifying glass. This option opens 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.

Home New Search My Strategies My Basket (3) My Data Sets Tools Data Summary Downloads Community Analyze My Experiment

Add to basket Add to favorites Download Gene

PF3D7_1015200 cysteine--tRNA ligase

Name: CysRS
Type: protein coding
Chromosome: 10
Location: PF3D7_10_v3:614,872..617,736(-)

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

View this gene at GeneDB
 Add the first user comment

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 Protein Features Proteomics

24 data sets

Also see PF3D7_1015200 in the JBrowse Genome Browser or Protein Browser

PF3D7_1015200 cysteine--tRNA ligase SNPs

Name: CysRS
 Type: protein coding
 Chromosome: 10
 Location: PF3D7_10_v3:614,872..617,736(-)

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

View this gene at GeneDB
 Add the first user comment

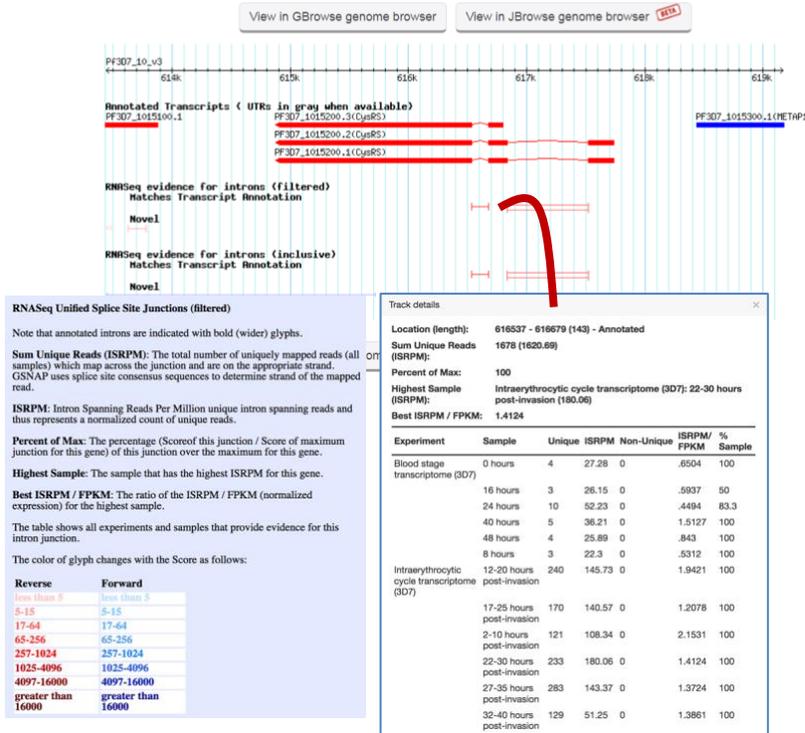
Contents

Exons in Gene
 Transcripts

Go to section on page

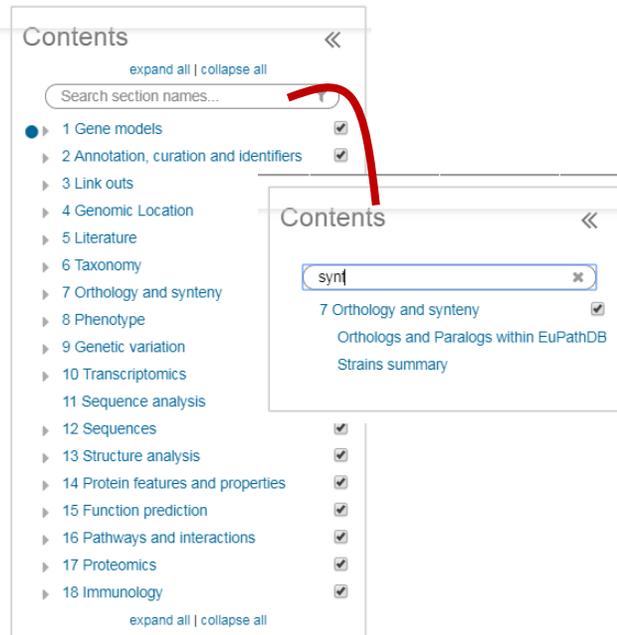
- Examine the “Gene Models” section of the gene page.
 - How many exons does this gene have?
 - How many transcripts does this gene encode?
 - What direction are the transcripts relative to the chromosome?
 - What does the “RNA Evidence for introns” information mean?
 - From what type of data are the “introns” determined?
 - How many nucleotides is the largest transcript? (hint: examine the transcripts table underneath the gene models).
 - Try out the ‘View in JBrowse genome browser’ button. Where does this link take you? What advantage do the genome browsers offer that the static images on the gene pages do not?

Exons in Gene 5
 # Transcripts 3
 ▾ Gene Models



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 JBrowse Genome Browser”** button right above the synteny section on the gene page.



View in JBrowse genome browser

- 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 few minutes to load).
- For each genome notice that there are two tracks: one called genes and the other span. 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?

b. Run a multiple sequence alignment comparing the protein sequence for this gene in Plasmodium falciparum 3D7 with orthologs from Plasmodium adleri G01, Plasmodium berghei ANKA and Plasmodium chabaudi chabaudi.

- Scroll up to the “Orthologs and Paralogs” within EuPathDB table. This table also functions as a multiple sequence alignment tool. Use the first column to check the strains you want to include in the alignment and then use the parameters at the bottom of the table to configure and run the alignment.

▼ Orthologs and Paralogs within EuPathDB [Data sets](#)

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the alignment with the 'Run Clustal Omega for selected genes' button.

Search this table... Showing 24 rows

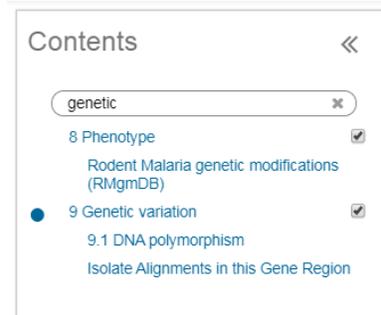
<input type="checkbox"/>	Gene	Organism	Product	is syntenic	has comments
<input checked="" type="checkbox"/>	PADL01_0105200	Plasmodium adleri G01	serine/threonine protein kinase, putative	yes	no
<input checked="" type="checkbox"/>	PBANKA_0205800	Plasmodium berghei ANKA	eukaryotic translation initiation factor 2-alpha kinase 2	yes	yes
<input checked="" type="checkbox"/>	PCHAS_0204200	Plasmodium chabaudi chabaudi	eukaryotic translation initiation factor 2-alpha kinase 2, putative	yes	yes
<input checked="" type="checkbox"/>	PF3D7_0107600	Plasmodium falciparum 3D7	eukaryotic translation initiation factor 2-alpha kinase 2, putative	yes	yes
<input type="checkbox"/>	PfDd2_010011200	Plasmodium falciparum Dd2	serine/threonine protein kinase, putative	yes	no

- How many mismatches do you find?
- How does this compare to an alignment run between P. falciparum strains? (try running an alignment with P. falciparum 3D7, Dd2, GB4, and KE01)

c. Does this gene contain Single Nucleotide Polymorphisms (SNPs)?

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 9.1.
 - 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 Alignments in this Gene Region’ 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 Strain’ quality allows you to choose individual isolates.
 - What do Ns indicate?



d. Is this gene expressed at the protein and/or transcript level?

```

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 TATTTTTTTC TTAGNNNNNN TNATATATAAT TATATATATA
RV_3635 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
RV_3675 600511 AAATATGTTT AATAAGTTGA AATTTTGTAA TTTATGAAAA TATTTTTTTC TTAGGAACTA TCTATATAAT TATATATATA
  
```

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’ on the bottom right of the page and then click on the ‘Shortcut’ image to navigate to that section of the page.

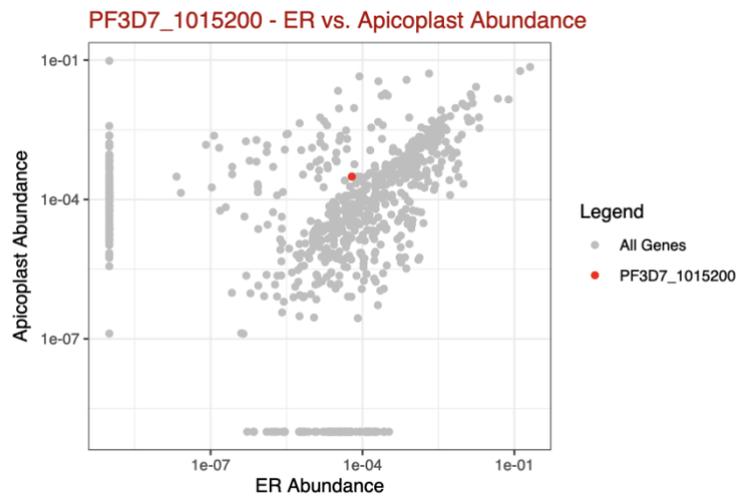
- What kinds of data in PlasmoDB provide evidence for protein expression? (Hint, view the Mass Spec.-based Expression Evidence table).
- Is this gene expressed at the protein level in salivary gland sporozoites?
- Does it contain any post-translational modifications?
- Can you quickly link to the data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

17 Proteomics

▼ Mass Spec.-based Expression Evidence [Data sets](#)

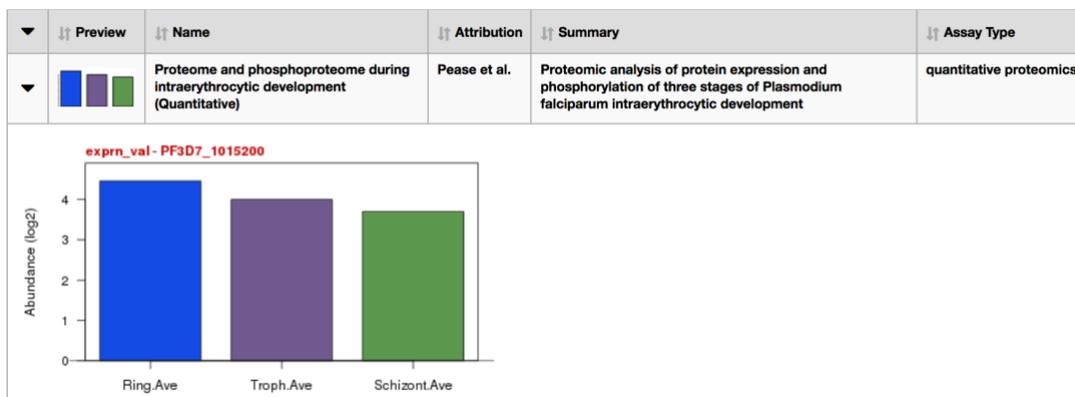
Search this table... Showing 2 rows

Transcript ID(s)	Experiment	Sample	Sequences	Spectra
Pf3D7_1015200.1, PF3D7_1015200.2, PF3D7_1015200.3	Blood stage phospho- and total proteome (3D7)	schizont phosphopeptide-depleted	3	8
Pf3D7_1015200.1, PF3D7_1015200.2, PF3D7_1015200.3	Cytoplasmic and nuclear fractions from rings, trophozoites and schizonts (3D7)	Ring stage nuclear fraction 1	2	3

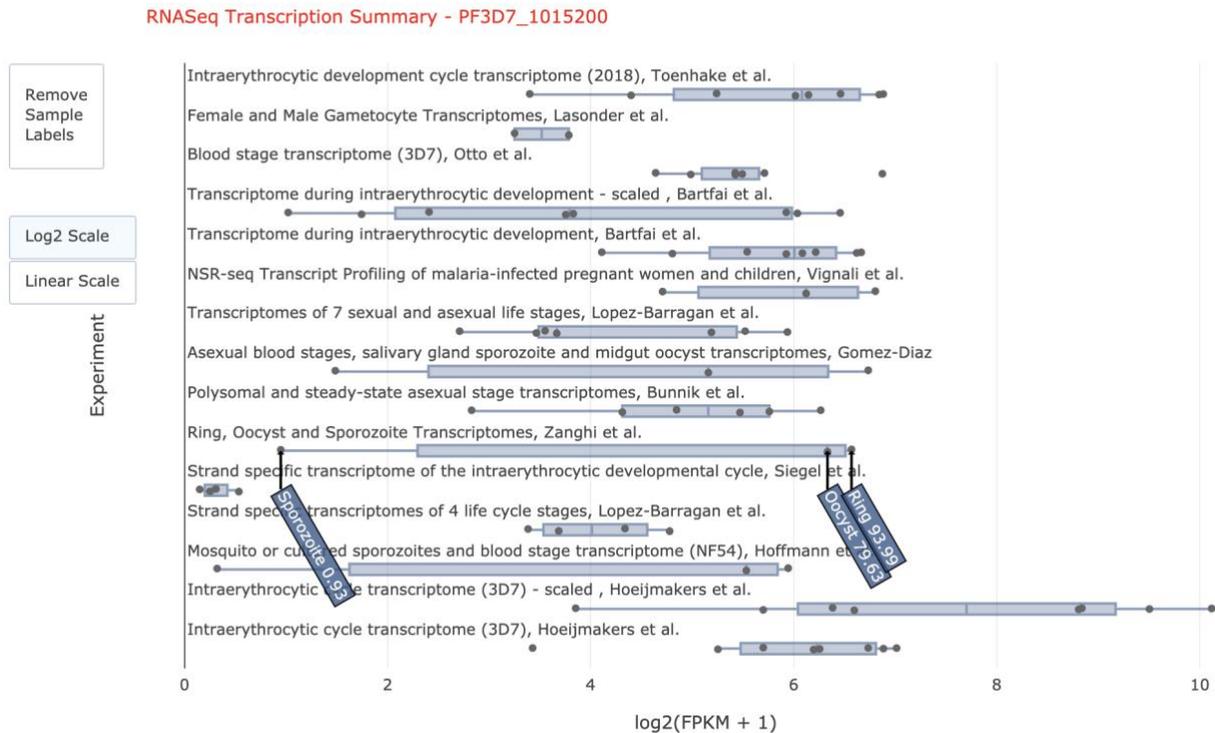


- Is this protein targeted to the apicoplast? What available experimental evidence supports your conclusion? (hint: examine apicoplast and ER proteome from Boucher et al.)
- How abundant is this protein? How confident are you of this analysis? Abundance can be estimated by counting the number of spectra 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 experiment called **Proteome and phosphoproteome during intraerythrocytic development (Quantitative)**.

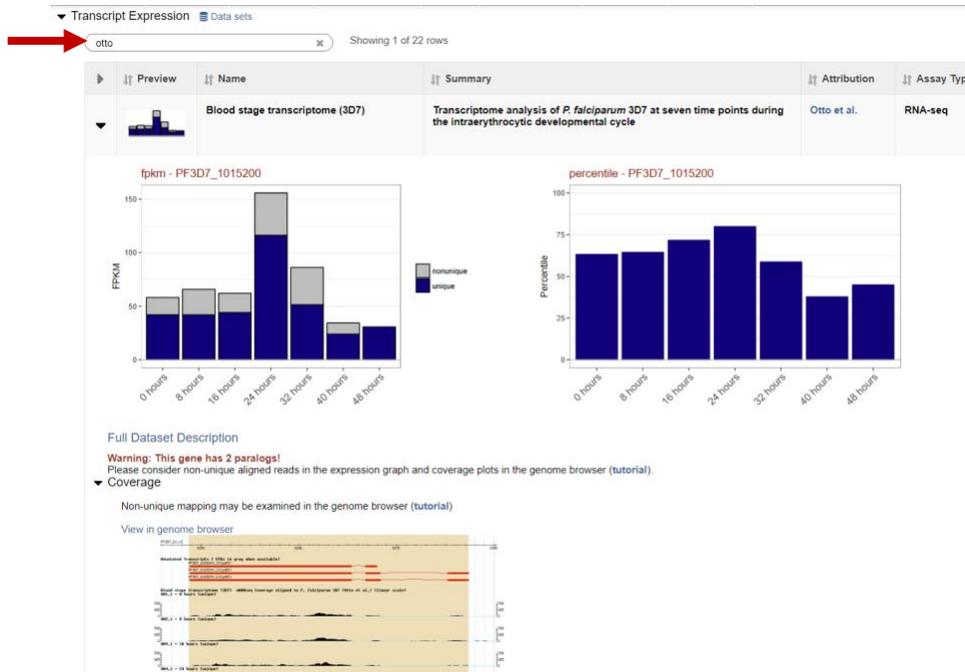
▼ Quantitative Mass Spec. [Data sets](#)



- Examine the Transcript Expression section of the gene page. Notice that the first part of the transcriptomics section includes a graph that summarizes all the available RNAseq experiments. This an interactive graph - try clicking on one of the dots in one of the experiments, what happens? Can you use this graph to determine in which experiment this gene is most differentially expressed? what about least?



- Scroll down and find Expression data experiment labeled **Life cycle expression data (3D7)**. Based on this data, at what life cycle stage is this gene most highly expressed?
- Does the proteomic data (from above) agree with the available transcriptomic data? (Hint, examine different transcriptomics experiments, microarrays and RNAseq – remember you can use the contents table on the left side of your screen). Should transcript expression and protein expression coincide perfectly?
- 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).



- Can you find microarray experiments? Hint: you can use the same method as above and instead of typing “Otto” you can type “microarray”.
 - How does the RNAseq data compare with the microarray data? Does the gene’s expression follow a similar pattern/trend in both experiment types?
 - How does the data from the Polysomal and steady-state asexual transcriptomes experiment compare? Why is this experiment interesting to look at?
- e. Is cysteine-tRNA ligase essential to Plasmodium? Does mutating the gene reduce fitness?**

Contents

- 1 Gene models
- 2 Annotation, curation and identifiers
- 3 Link outs
- 4 Genomic Location
- 5 Literature
- 6 Taxonomy
- 7 Orthology and synteny
- 8 Phenotype
- 9 Genetic variation

Choose graph(s) to display

MIS MFS

▼ Phenotype Graphs Data sets

Preview	Name	Summary	Attribution
	Piggyback insertion metagenesis	<i>P. falciparum</i> NF54 mutants were generated via random piggyBac transposon mutagenesis. Mutants are genetically identical except for a single randomly inserted transposon at TTA tetranucleotide sites.	

PF3D7_1015200 - Mutagenesis Index Score

Mutagenesis Index Score

Legend

- All Genes
- PF3D7_1015200

PF3D7_1015200 - Mutant Fitness Score

Mutant Fitness Score

Legend

- All Genes
- PF3D7_1015200

Full Dataset Description

▼ Data table

Search this table... Showing 2 rows

Profile Set	Gene	Sample	Score	Score Type
piggyBac mutagenesis index score	PF3D7_1015200	MIS (phenotype)	0.27	mutagenesis index score
piggyBac mutant fitness score	PF3D7_1015200	MFS (phenotype)	-2.69	mutant fitness score

- Navigate to the phenotype section and notice the Piggyback insertion metagenesis data in the Phenotype Graphs.
- Explore the data and data descriptions in order to gain an understanding of the data and its meaning. Visit the data description page for an overview of the data set, links to the publication, etc.
- Open the Data Table. What are the MIS and MFS scores for this gene?
- How do these scores compare to scores for the rest of the genome?
- What do these scores mean? (hint: see experiment description)
- 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? Note: K13 is a protein that is involved in resistance to one of the frontline malaria drugs.