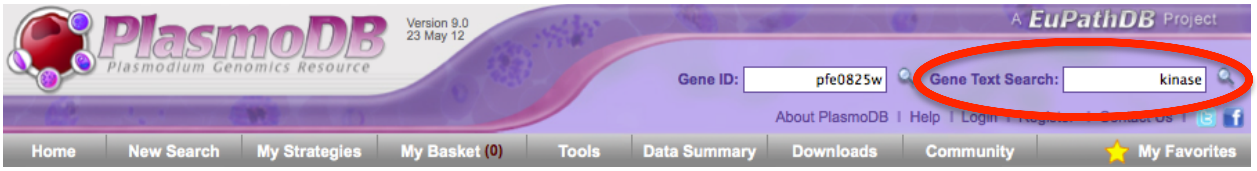
Finding genes and exploring the gene page
Exercise 1

1.1 Finding a gene using text search.
Note: 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.



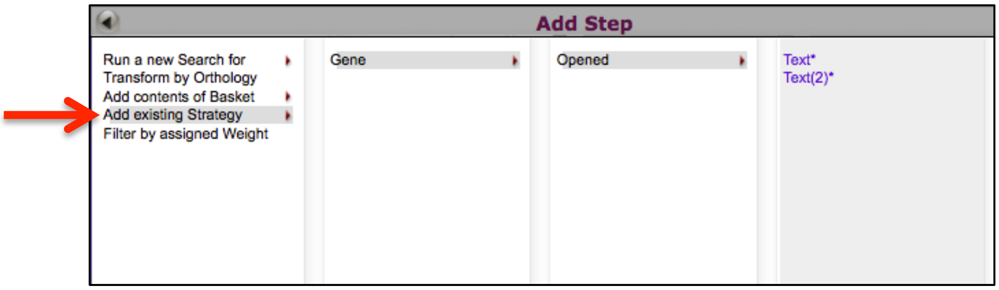
- How many genes did you get?
- How many of those are in *P. falciparum*? How did you find this out?
- What happens if you search using the word "kinases"? How many results did you return?

b. **How can you increase the number of possible kinases in your results**?

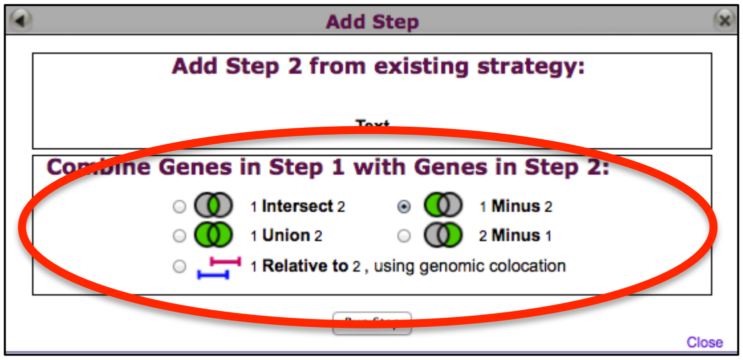
Hint: the search you did in 'a' will miss things like "6-phosphofructokinase" or "kinases" so you need to use a wild card in your search. Try "kinase*", "*kinase" and "*kinase*" (without quotations).

- Did you get more results?
- Which one of the above wild card combinations gave you the largest number of kinases?

- How can you quickly examine the genes that were identified using the key word "*kinase*" but not with the word "kinase"? Hint: You can easily do this by combining search strategies. Click on "Add Step" then select "existing strategy":



- Select the right strategy from the list of Gene Strategies. This about the type of operation you will apply when combing the strategies:



Add Step

Add Step 2 from existing strategy:

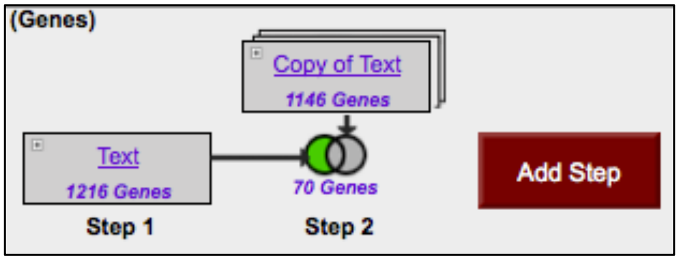
Text

Combine Genes in Step 1 with Genes in Step 2:

○ 1 **Intersect** 2 ● 1 **Minus** 2
○ 1 **Union** 2 ○ 2 **Minus** 1
○ 1 **Relative to** 2 , using genomic colocation

Close

Which operation did you choose?

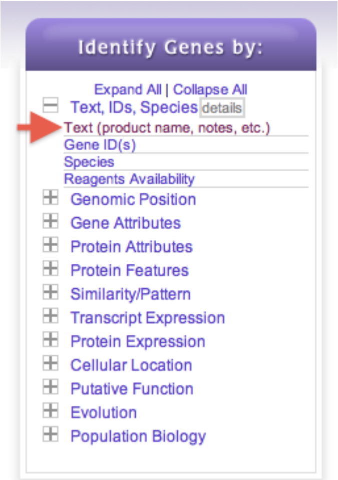


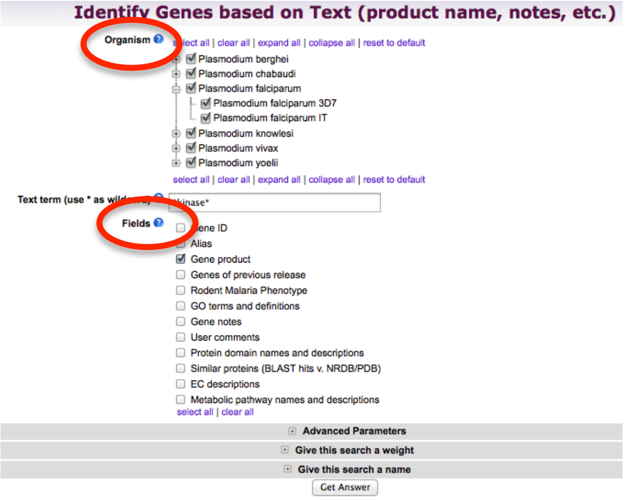
- Do the results make sense? Do all the product names contain the word kinase?

c. **Find only the kinases that specifically have the word “kinase” in the gene product name**.

Hint: use the text search page, the specific search page where you can define the fields to be included in the search. There are many ways to navigate to the Text Search page.

- How did you get there?
- How many kinases have the word kinase in their product names?
- Did you remember to use the wild card?




Identify Genes based on Text (product name, notes, etc.)
select all | clear all | expand all | collapse all | reset to default
select all | clear all | expand all | collapse all | reset to default
select all | clear all
Advanced Parameters
Give this search a weight
Give this search a name
Get Answer

Organism
- ☑ Plasmodium berghei
- ☑ Plasmodium chabaudi
- ☑ Plasmodium falciparum
  - ☑ Plasmodium falciparum 3D7
  - ☑ Plasmodium falciparum IT
- ☑ Plasmodium knowlesi
- ☑ Plasmodium vivax
- ☑ Plasmodium yoelii

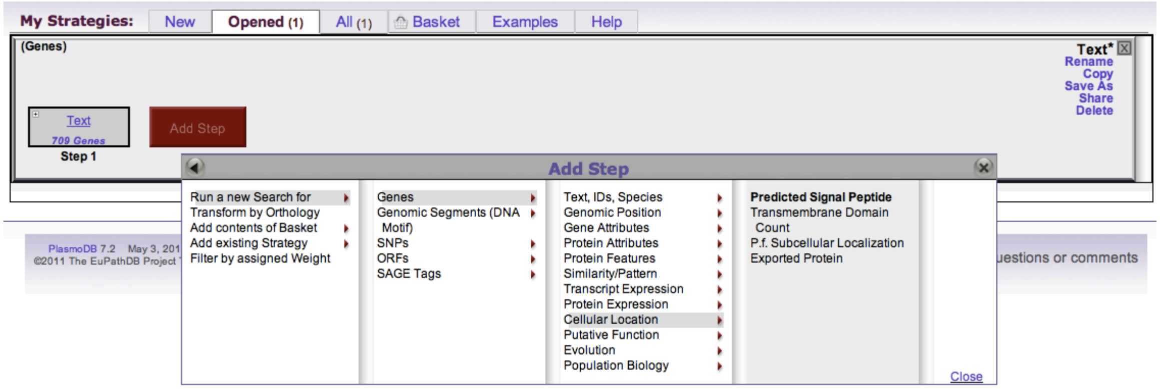
Text term (use * as wildcard): kinase*

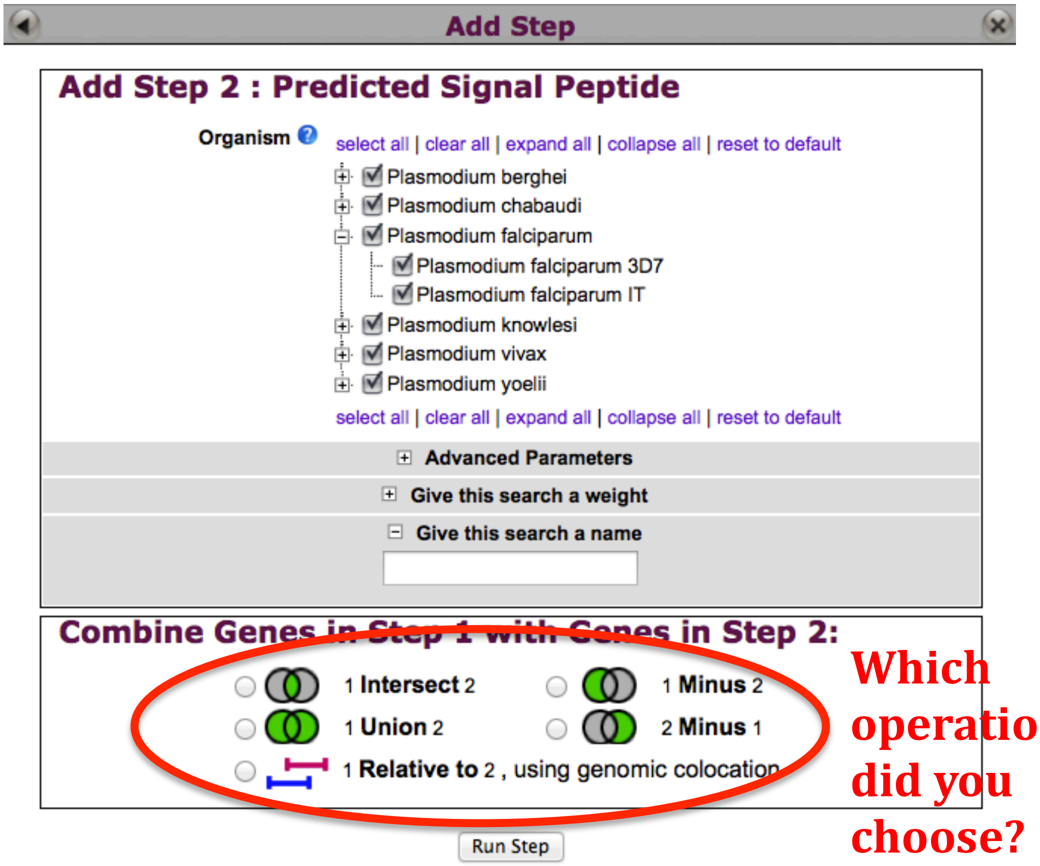
Fields
- ☐ Gene ID
- ☐ Alias
- ☑ Gene product
- ☐ Genes of previous release
- ☐ Rodent Malaria Phenotype
- ☐ GO terms and definitions
- ☐ Gene notes
- ☐ User comments
- ☐ Protein domain names and descriptions
- ☐ Similar proteins (BLAST hits v. NRDB/PDB)
- ☐ EC descriptions
- ☐ Metabolic pathway names and descriptions

1.2 Combing text search results with results from other searches

a. In exercise 1.1 you identified genes that have the word “kinase” somewhere in their product name. Can you now find out how many of these kinases are likely secreted?

Hint: grow your search strategy by adding a step. Choose a search that identifies genes with likely secretory signal peptides. How did you combine the search results?





Add Step

Add Step 2 : Predicted Signal Peptide

Organism select all | clear all | expand all | collapse all | reset to default

- ☑ Plasmodium berghei
- ☑ Plasmodium chabaudi
- ☑ Plasmodium falciparum
  - ☑ Plasmodium falciparum 3D7
  - ☑ Plasmodium falciparum IT
- ☑ Plasmodium knowlesi
- ☑ Plasmodium vivax
- ☑ Plasmodium yoelii

select all | clear all | expand all | collapse all | reset to default

⊞ Advanced Parameters

⊞ Give this search a weight

⊟ Give this search a name

Combine Genes in Step 1 with Genes in Step 2:

○ 1 Intersect 2 ○ 1 Minus 2

○ 1 Union 2 ○ 2 Minus 1

○ 1 Relative to 2 , using genomic colocation

Run Step

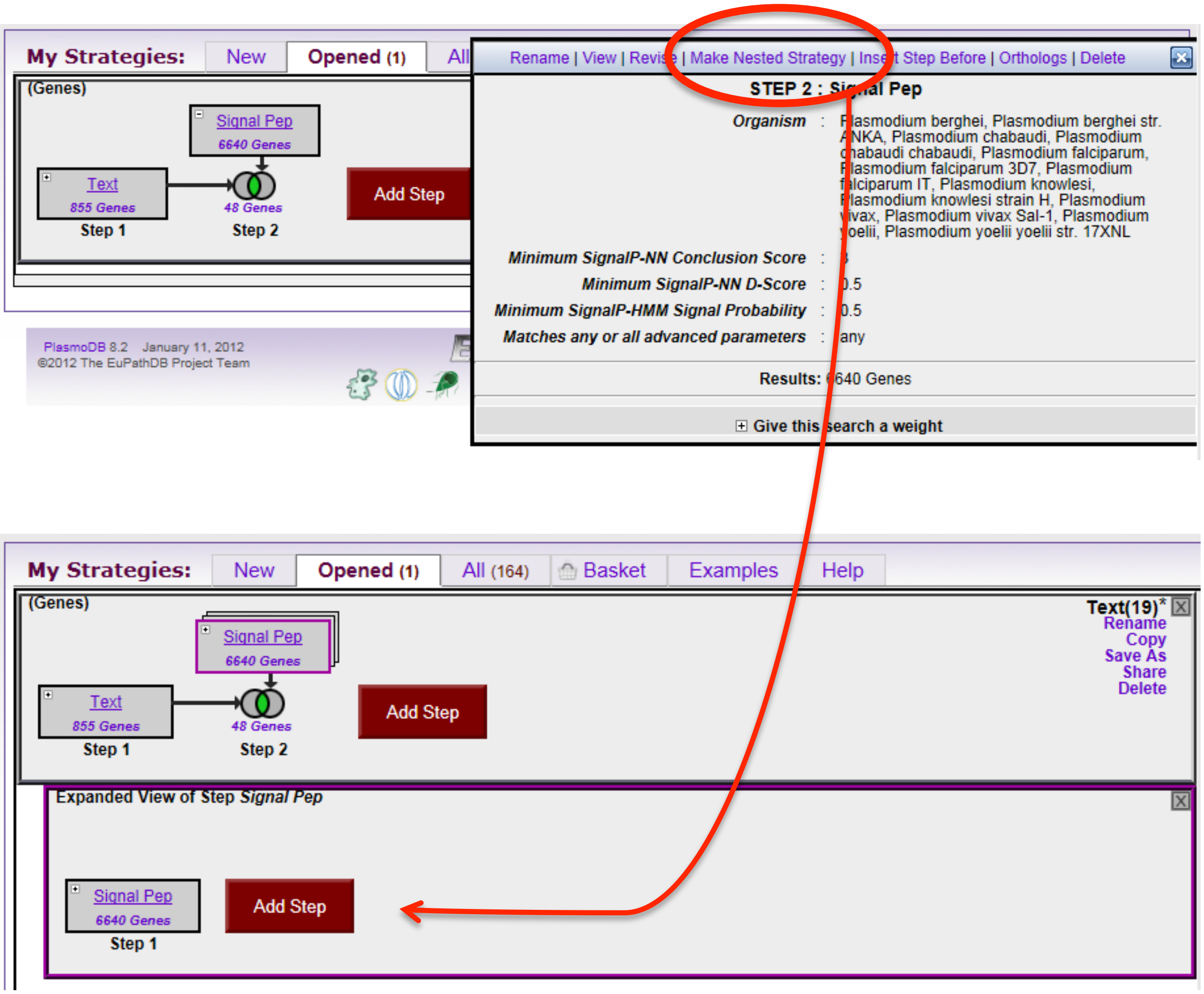
Which operation did you choose?

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

Hint: 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.
- For example, how many of these 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. Why?

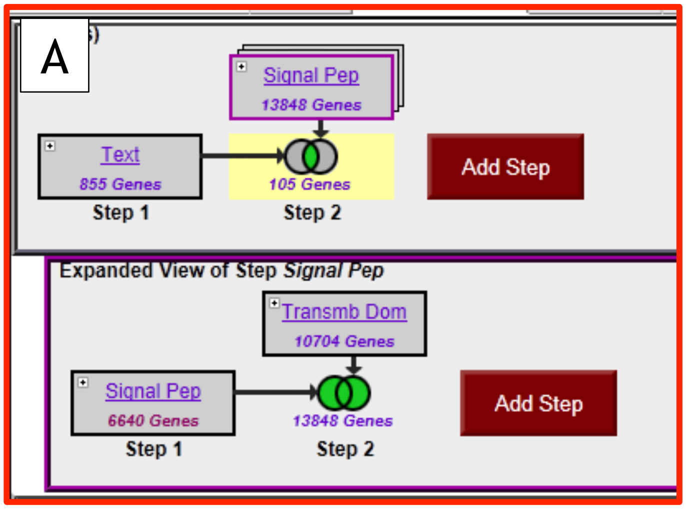


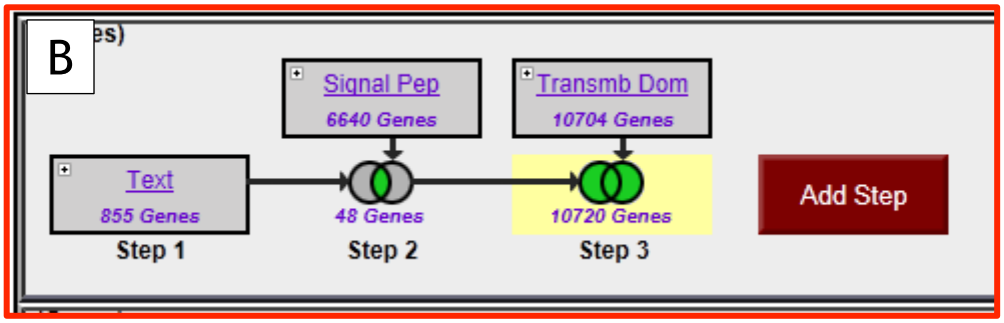
STEP 2 : Signal Pep

Organism	:	Plasmodium berghei, Plasmodium berghei str. ANKA, Plasmodium chabaudi, Plasmodium chabaudi chabaudi, Plasmodium falciparum, Plasmodium falciparum 3D7, Plasmodium falciparum IT, Plasmodium knowlesi, Plasmodium knowlesi strain H, Plasmodium vivax, Plasmodium vivax Sal-1, Plasmodium yoelii, Plasmodium yoelii yoelii str. 17XNL
Minimum SignalP-NN Conclusion Score	:	3
Minimum SignalP-NN D-Score	:	0.5
Minimum SignalP-HMM Signal Probability	:	0.5
Matches any or all advanced parameters	:	any

Results: 6640 Genes

Notice the different results obtained in figures A (with nesting) and B (without nesting) below:





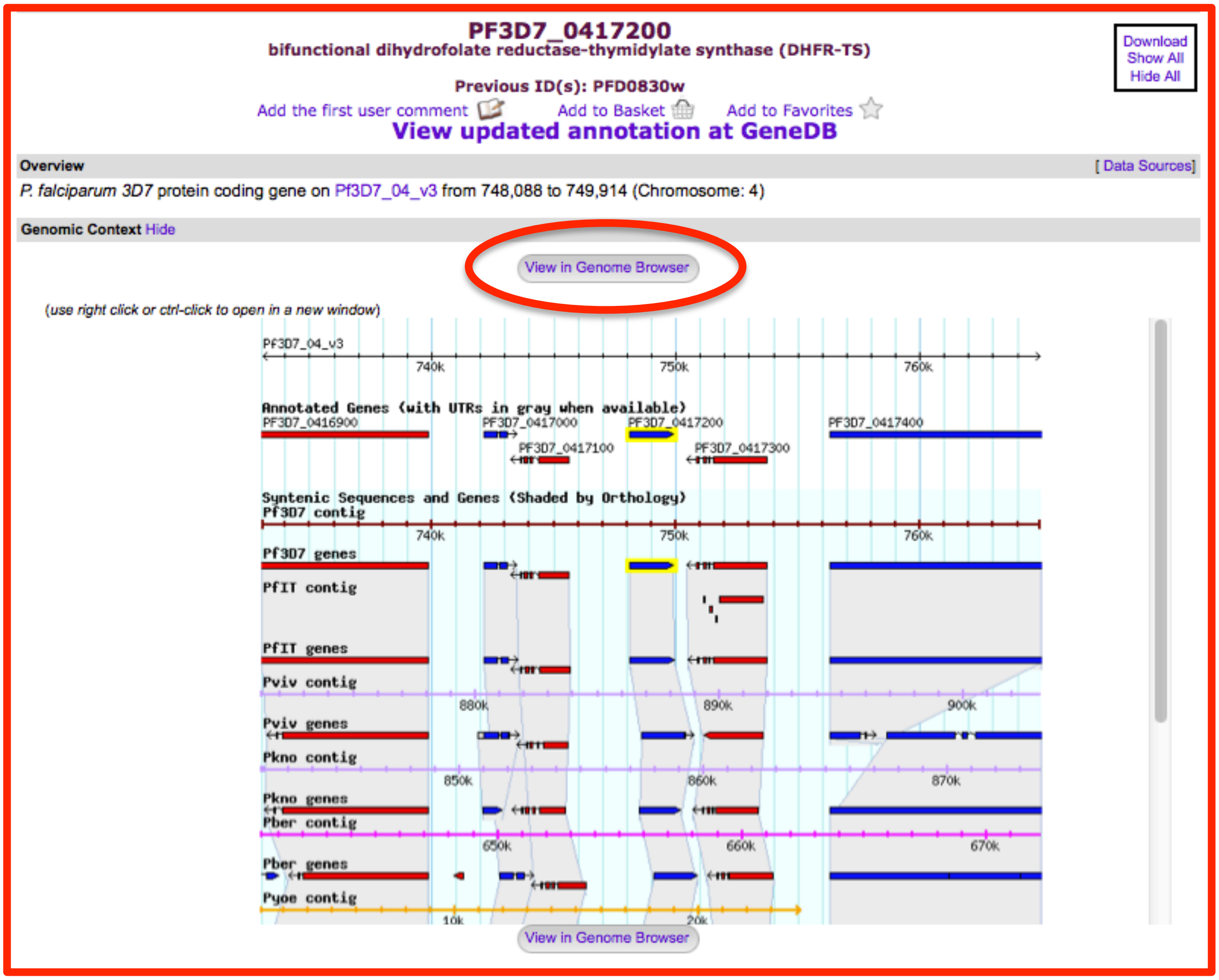
1.3 Visiting a specific gene page.
Note: For this exercise use http://www.plasmodb.org

a. Find the bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene (or, if you prefer, the apical membrane antigen 1 gene; AMA1) in *P. falciparum*.

- How did you navigate to this gene? What other ways could you get there? (Hint: what about using the gene ID? PF3D7_0417200)
- How many exons in this gene?
- How many nucleotides of coding sequence?

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?
- How complete is the genome assembly for other species? (Hint: it may help to view in the genome browser.) The genome browser can be accessed from gene pages by clicking on the "View in Genome Browser" link (see below). In the genome browser, data tracks can be loaded from the "Select Tracks" page. Tracks are automatically added to the browser image when you select them on the "Select Tracks" page. Just go back to the "Browser" page to view the data.
- Which tracks did you turn on?



PF3D7_0417200
bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS)

Previous ID(s): PFD0830w

Overview
P. falciparum 3D7 protein coding gene on Pf3D7_04_v3 from 748,088 to 749,914 (Chromosome: 4)

Genomic Context

(use right click or ctrl-click to open in a new window)

c. **How many Single Nucleotide Polymorphisms (SNPs) can you identify within the *P. falciparum* DHFR-TS (or AMA1) gene?** (Note: Go back to the gene page using the back arrow in your browser or by clicking on the gene highlighted in yellow).
   - How many of these SNPs are in coding sequence?
   - How many impact the predicted protein sequence?
   - How many alleles are there for each SNPs?

- What is the maximum number of SNPs per strain?
- Is this likely to define the full spectrum of sequence variation in these particular strains?
- What about in field isolates? (Hint: mouse over the diamond shapes to get more information about the SNPs.)
- How do these results compare with SNP distribution in other genes?

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

Hint: look at the gene page sections entitled "Protein" and "Expression" – you may have to click on the show link to reveal the underlying data).

- What kinds of data in PlasmoDB provide evidence for expression?
- At what life cycle stage is DHFR-TS (AMA1) most abundant?
- Does this make sense?
- How do the different life cycle microarray expression profiles compare to each other?
- Are the results similar? What about RNA-sequence data, does it agree with microarray data?
- How abundant is DHFR-TS (AMA1) protein? How confident are you of this analysis?

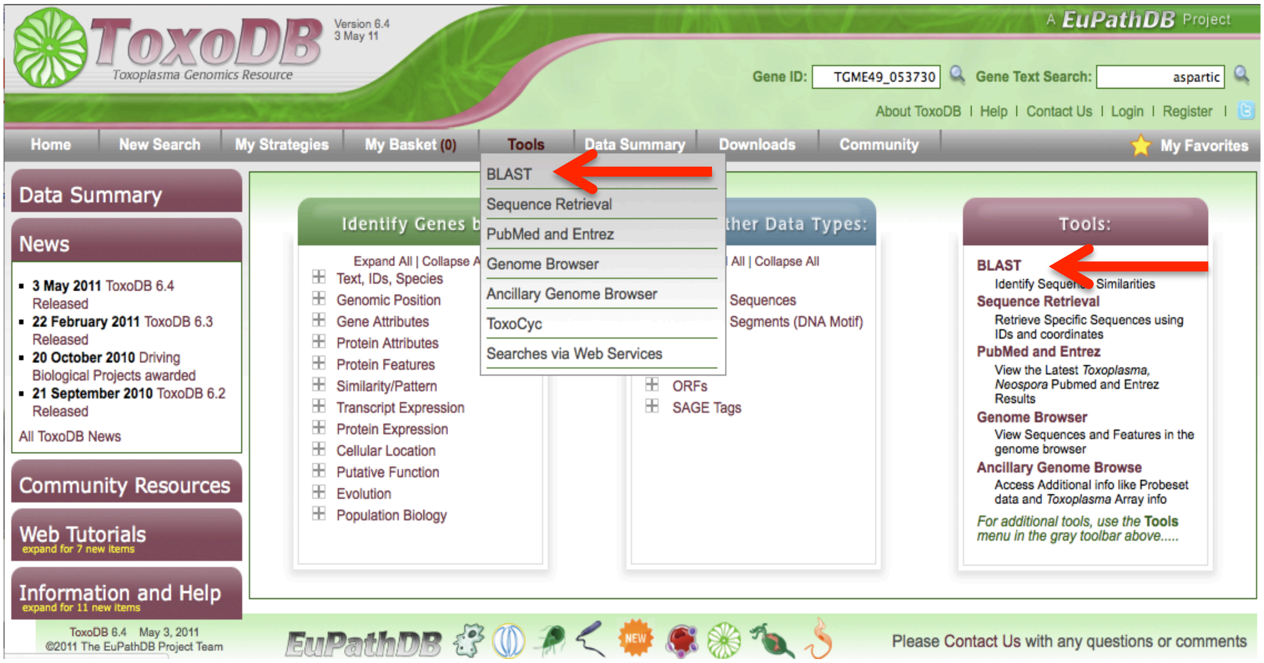
1.4 Finding a gene by BLAST.

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

Also, you need to open another window in your browser to retrieve the sequence we will use in our BLAST search, got to:

http://goo.gl/6BGjZ

- 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 in the link above). You immediately go to ToxoDB to find any information about this sequence. What do you do?
- Try running a 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.)



- **3 May 2011** ToxoDB 6.4 Released
- **22 February 2011** ToxoDB 6.3 Released
- **20 October 2010** Driving Biological Projects awarded
- **21 September 2010** ToxoDB 6.2 Released

All ToxoDB News

BLAST
Identify Sequence Similarities
Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates
PubMed and Entrez
View the Latest *Toxoplasma*, *Neospora* Pubmed and Entrez Results
Genome Browser
View Sequences and Features in the genome browser
Ancillary Genome Browse
Access Additional info like Probeset data and *Toxoplasma* Array info

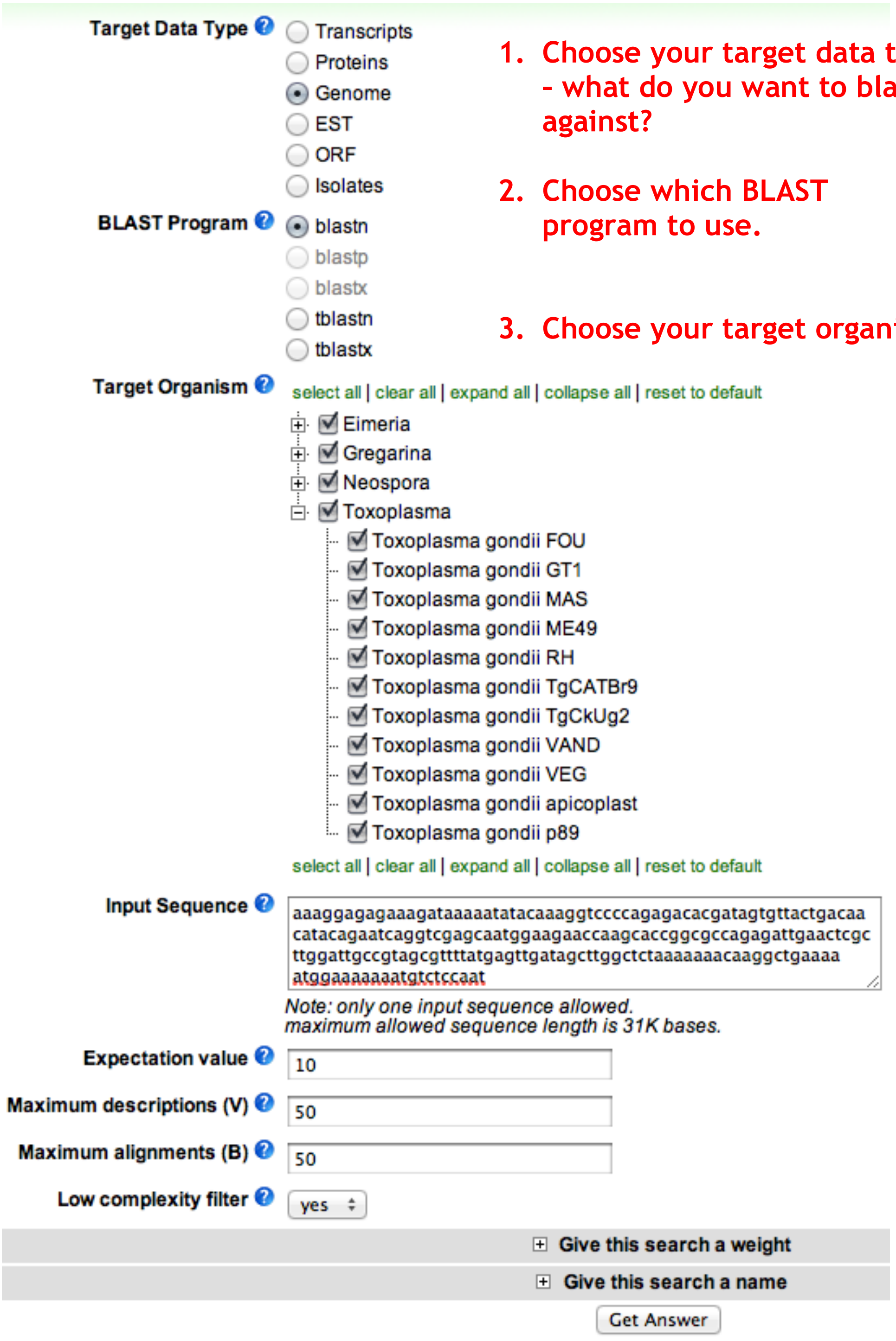
*For additional tools, use the **Tools** menu in the gray toolbar above.....*

ToxoDB 6.4 May 3, 2011
©2011 The EuPathDB Project Team

Please Contact Us with any questions or comments

- 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.*



Target Data Type
○ Transcripts
○ Proteins
● Genome
○ EST
○ ORF
○ Isolates

BLAST Program
● blastn
○ blastp
○ blastx
○ tblastn
○ tblastx

Target Organism
select all | clear all | expand all | collapse all | reset to default

☑ Eimeria
☑ Gregarina
☑ Neospora
☑ Toxoplasma
- ☑ Toxoplasma gondii FOU
- ☑ Toxoplasma gondii GT1
- ☑ Toxoplasma gondii MAS
- ☑ Toxoplasma gondii ME49
- ☑ Toxoplasma gondii RH
- ☑ Toxoplasma gondii TgCATBr9
- ☑ Toxoplasma gondii TgCkUg2
- ☑ Toxoplasma gondii VAND
- ☑ Toxoplasma gondii VEG
- ☑ Toxoplasma gondii apicoplast
- ☑ Toxoplasma gondii p89

select all | clear all | expand all | collapse all | reset to default

Input Sequence

```
aaaggagagaaagataaaaatatacaaaggtccccagagacacgatagtgttactgacaa
catacagaatcaggtcgagcaatggaagaaccaagcaccggcgccagagattgaactcgc
ttggattgccgtagcgttttatgagttgatagcttggctctaaaaaaacaaggctgaaaa
atggaaaaaaatgtctccaat
```

Note: only one input sequence allowed.
maximum allowed sequence length is 31K bases.

Expectation value: 10

Maximum descriptions (V): 50

Maximum alignments (B): 50

Low complexity filter: yes

⊞ Give this search a weight

⊞ Give this search a name

Get Answer

1. Choose your target data type - what do you want to blast against?

2. Choose which BLAST program to use.

3. Choose your target organism.

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

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

a. 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.)
b. You should get multiple hits. Find the first one that is annotated as "dihydrofolate reductase-thymidylate synthase" (Look in the product description column.)
c. Once you find one, click on the gene and go to the gene page. It might be helpful to open the gene page in a new window or tab.
d. Scroll down to the bottom of the page to the “Sequences” section.
e. 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.)
f. Go to the BLAST page from the EuPathDB home page. (Hint: look under Tools on the EuPathDB home page.)
g. Paste the amino acid sequence into the input window.
h. Select target data type (start with "Proteins").
i. Select BLAST Program. (Hint: BLASTP).
j. Select the target organism. (Let's start by selecting all.) Click on "Get Answer".

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

k. Go back to the BLAST window. Change the target data type to Genome.
l. 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.)
m. Select all target organisms. Click on "Get Answer".

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

n. Go to your gene page window (in CryptoDB) and copy the nucleotide coding sequence.
o. Go to the BLAST window and paste the nucleotide sequence into the input window.
p. 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.)
q. Select the target organisms. This time let's specifically only select *Giardia*, *Entamoeba* and *Trichomonas*. Click on "Get Answer".

Getting frustrated?

Not getting a hit for *Giardia*, *Entamoeba* and *Trichomonas* in this case is actually the correct answer! These organisms do not have dihydrofolate reductase or thymidylate synthetase activity.