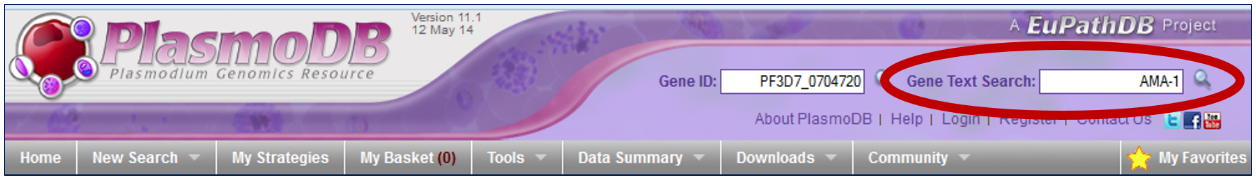
FINDING GENES AND BUILDING SEARCH STRATEGIES
(Exercise 1)

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.



Gene ID: PF3D7_0704720

Gene Text Search: AMA-1

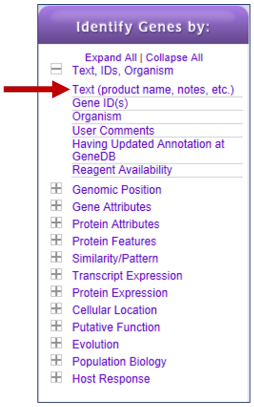
- How many genes did you get?
- Look closely at the sections of the result page. How many of those are in *P. falciparum*? How did you find this out?
- What happens if you search using “kinases”? How many results were returned?

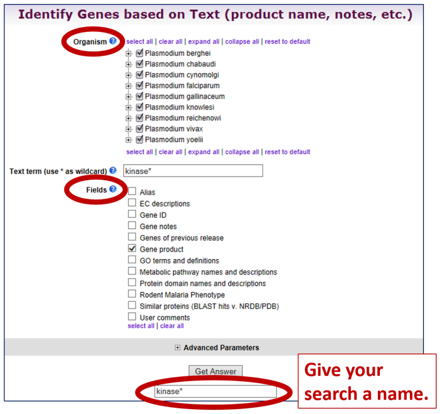
b. **Find only the kinases that specifically have the word “kinase” in the gene product name**.

Use the full text search, the specific page where you can define the fields to be searched. There are several ways to navigate to the Text Search page.

The search you did in ‘a’ will miss plurals and compound words like “kinases” or “6-phosphofructokinase”. Using a wild card in your search term will broaden your search – **try kinase kinase* *kinase *kinase***

- Try giving each new search a name to help you keep track of the searches.
- How did you get to the Text Search page?
- Did you remember to use the wild card?
- How many genes have the word kinase in their product names?





Identify Genes based on Text (product name, notes, etc.)

Organism
select all | clear all | expand all | collapse all | reset to default
☑ Plasmodium berghei
☑ Plasmodium chabaudi
☑ Plasmodium cynomolgi
☑ Plasmodium falciparum
☑ Plasmodium gallinaceum
☑ Plasmodium knowlesi
☑ Plasmodium reichenowi
☑ Plasmodium vivax
☑ Plasmodium yoelii
select all | clear all | expand all | collapse all | reset to default

Text term (use * as wildcard): kinase*

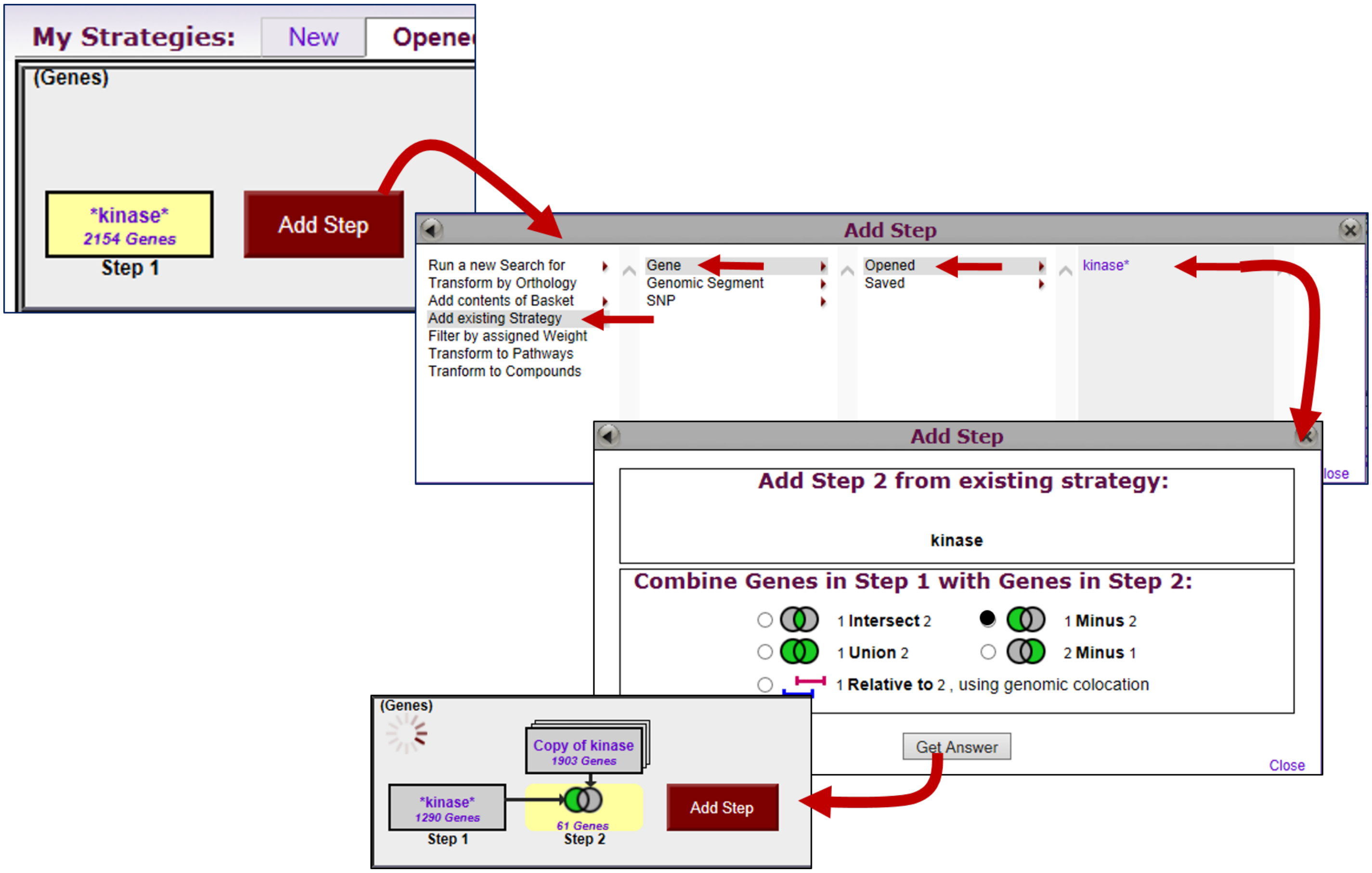
Fields
☐ Alias
☐ EC descriptions
☐ Gene ID
☐ Gene notes
☐ Genes of previous release
☑ Gene product
☐ GO terms and definitions
☐ Metabolic pathway names and descriptions
☐ Protein domain names and descriptions
☐ Rodent Malaria Phenotype
☐ Similar proteins (BLAST hits v. NRDB/PDB)
☐ User comments
select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

kinase*

Give your search a name.

c. **How can you quickly examine the genes that were identified using the key word *kinase* but not with the word kinase?**

- Hint: build a search strategy that combines 2 of your searches. Click on "Add Step" then 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?



Add Step 2 from existing strategy:

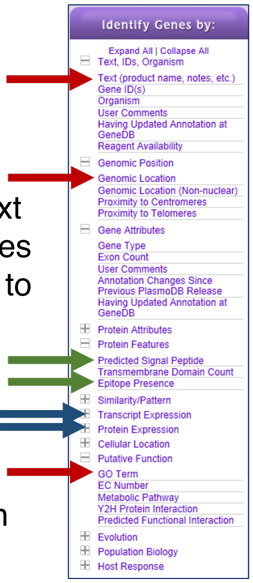
kinase

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

d. Find *P. berghei* ANKA genes that have between 2 and 5 transmembrane domains.

In this example we will search for genes based on attributes other than text annotations. There are many attributes associated with genes that can be used to identify genes. A gene's genomic location, Gene Ontology assignments, number of exons, number of transmembrane domains are all examples of gene 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.**

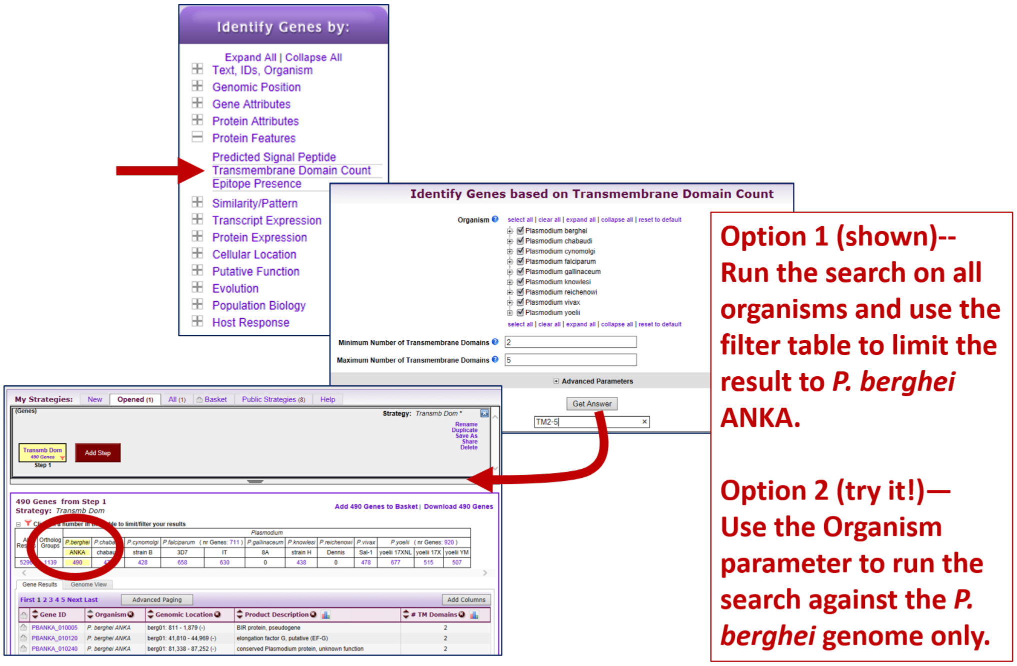
Genome Analysis Results

- **Predicted Signal Peptide**
- **Epitope Presence**
- **Transmembrane Domains**

Functional Data

- **Microarray**
- **Proteomics**
- **RNA Sequencing**

- How many *P. berghei* ANKA genes have between 2 and 5 transmembrane domains?
- Hint – use the Transmembrane Domain Count search under Protein Features.
- There are two options for setting the Organism parameter. Which option did you use? Using Option 1 takes advantage of the filter table to easily toggle between search results of different species.



Identify Genes based on Transmembrane Domain Count

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

- ☑ Plasmodium berghei
- ☑ Plasmodium chabaudi
- ☑ Plasmodium cynomolgi
- ☑ Plasmodium falciparum
- ☑ Plasmodium gallinaceum
- ☑ Plasmodium knowlesi
- ☑ Plasmodium reichenowi
- ☑ Plasmodium vivax
- ☑ Plasmodium yoelii

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

Minimum Number of Transmembrane Domains: 2

Maximum Number of Transmembrane Domains: 5

⊞ Advanced Parameters

Get Answer

TM2-5

My Strategies: New | Opened (1) | All (1) | Basket | Public Strategies (8) | Help
Rename
Duplicate
Save As
Share
Delete
Add Step

(Genes) Strategy: *Transmb Dom* *

Transmb Dom — *490 Genes* — Step 1

490 Genes from Step 1
Strategy: *Transmb Dom*

Add 490 Genes to Basket | Download 490 Genes

Cli a number in ble to limit/filter your results

Al Res s	Ortholog Groups	*Plasmodium*											
		P.berghei	*P.chaba*	*P.cynomolgi*	*P.falciparum* (nr Genes: 711)		*P.gallinaceum*	*P.knowlesi*	*P.reichenowi*	*P.vivax*	*P.yoelii* (nr Genes: 920)		
		ANKA	chabau	strain B	3D7	IT	8A	strain H	Dennis	Sal-1	yoelii 17XNL	yoelii 17X	yoelii YM
5296	1139	490	47	428	658	630	0	438	0	478	677	515	507

Gene ID	Organism	Genomic Location	Product Description	# TM Domains
PBANKA_010005	*P. berghei ANKA*	berg01: 811 - 1,879 (-)	BIR protein, pseudogene	2
PBANKA_010120	*P. berghei ANKA*	berg01: 41,810 - 44,969 (-)	elongation factor G, putative (EF-G)	2
PBANKA_010240	*P. berghei ANKA*	berg01: 81,338 - 87,252 (-)	conserved Plasmodium protein, unknown function	2

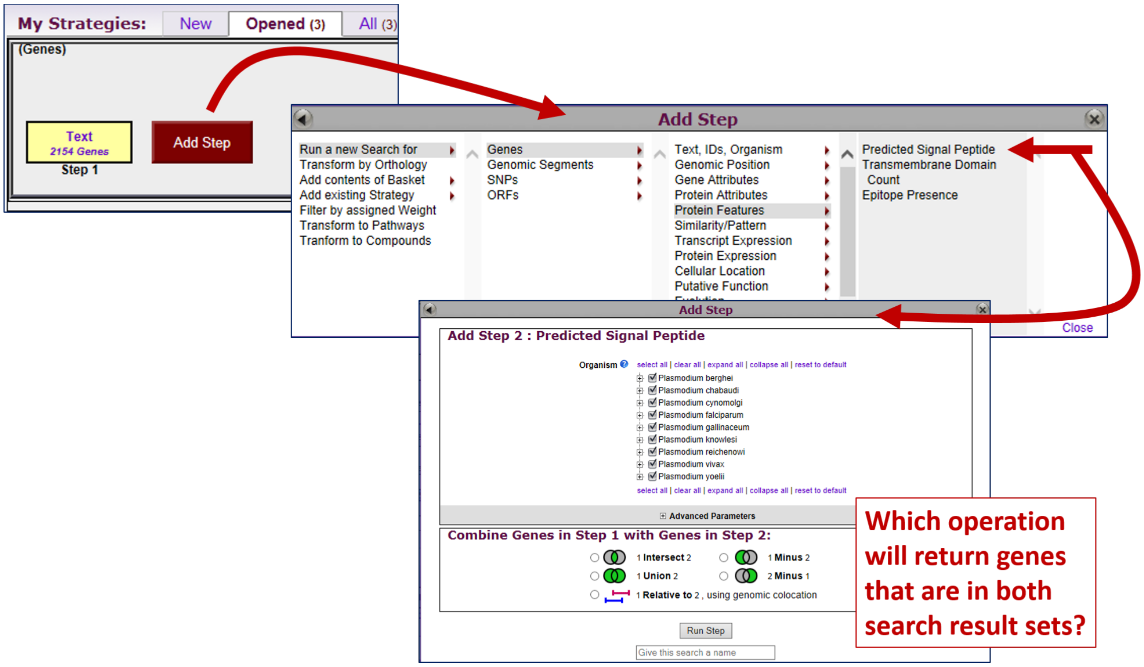
Option 1 (shown)-- Run the search on all organisms and use the filter table to limit the result to *P. berghei* ANKA.

Option 2 (try it!)— Use the Organism parameter to run the search against the *P. berghei* genome only.

1.2 Combing text search results with results from other searches

a. In exercise 1.1b 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 2 : Predicted Signal Peptide

Organism

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

- ☑ Plasmodium berghei
- ☑ Plasmodium chabaudi
- ☑ Plasmodium cynomolgi
- ☑ Plasmodium falciparum
- ☑ Plasmodium gallinaceum
- ☑ Plasmodium knowlesi
- ☑ Plasmodium reichenowi
- ☑ Plasmodium vivax
- ☑ Plasmodium yoelii

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

Advanced Parameters

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

Give this search a name

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

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

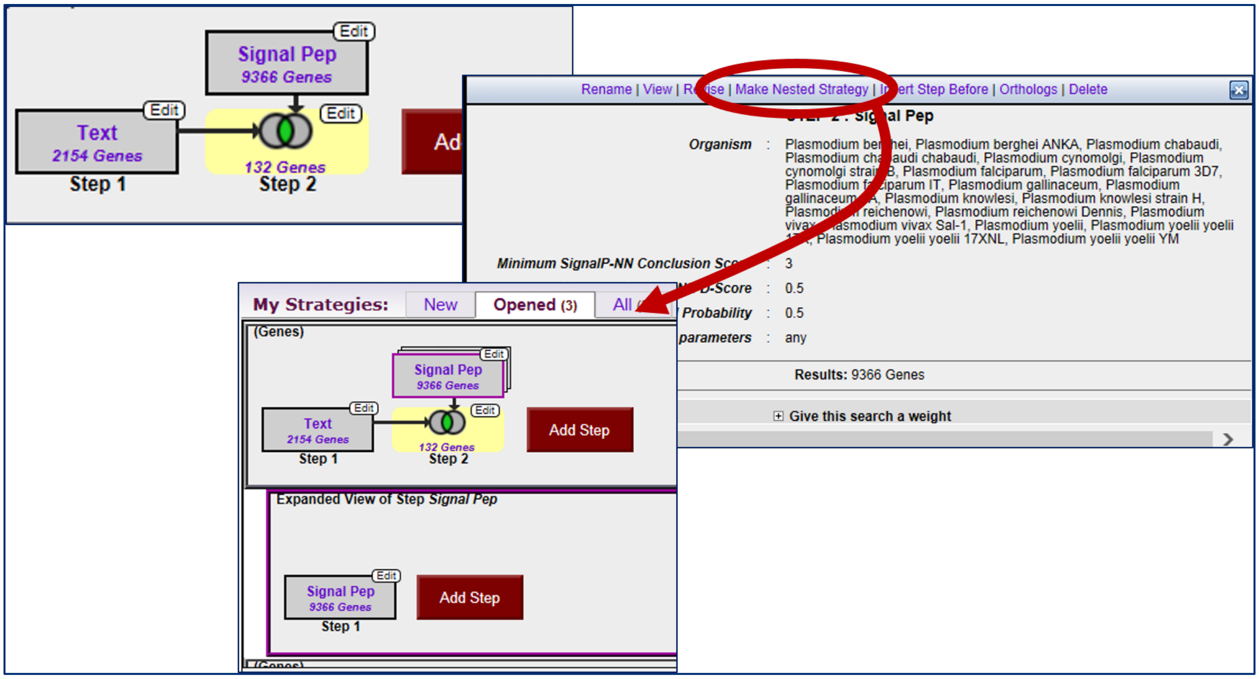
There is no wrong answer here!!

- 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 these secreted kinases also have transmembrane domains?

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

Hint: to do this properly you will have to employ the “Nested Strategy” feature. Why?



STEP 2 : Sig al Pep

Organism : Plasmodium ber hei, Plasmodium berghei ANKA, Plasmodium chabaudi, Plasmodium cha audi chabaudi, Plasmodium cynomolgi, Plasmodium cynomolgi strai B, Plasmodium falciparum, Plasmodium falciparum 3D7, Plasmodium fa ciparum IT, Plasmodium gallinaceum, Plasmodium gallinaceum A, Plasmodium knowlesi, Plasmodium knowlesi strain H, Plasmodi reichenowi, Plasmodium reichenowi Dennis, Plasmodium vivax, lasmodium vivax Sal-1, Plasmodium yoelii, Plasmodium yoelii yoelii 17 , Plasmodium yoelii yoelii 17XNL, Plasmodium yoelii yoelii YM

Minimum SignalP-NN Conclusion Sc : 3

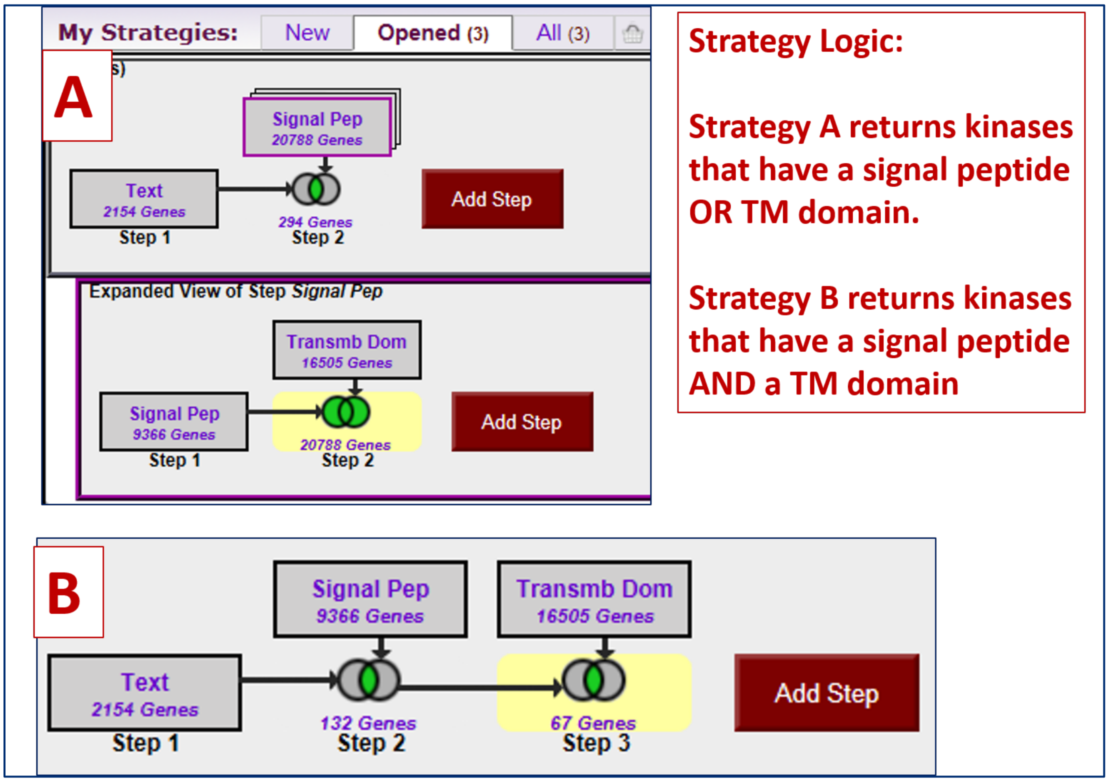
D-Score : 0.5

Probability : 0.5

parameters : any

Results: 9366 Genes

⊞ Give this search a weight



Strategy Logic:

Strategy A returns kinases that have a signal peptide OR TM domain.

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

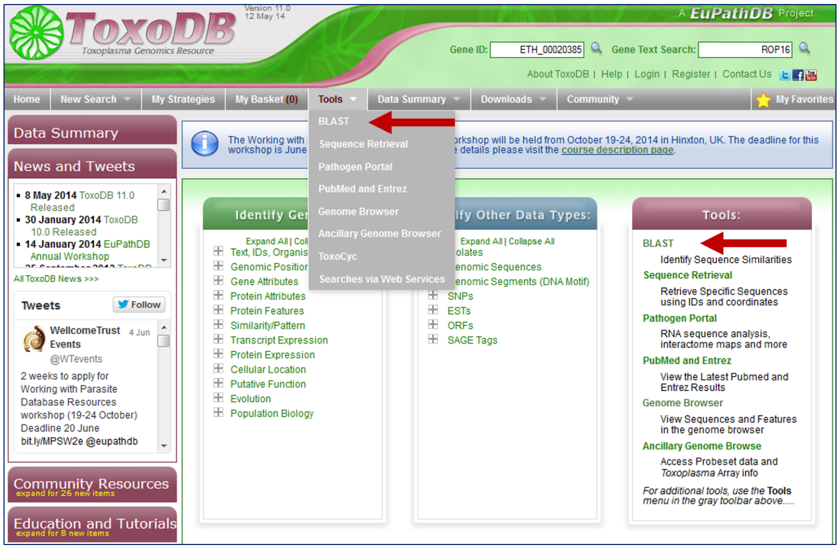
1.3 Finding a gene by BLAST Similarity.

Note: For this exercise use 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**

- 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).



The Working with workshop is June ... orkshop will be held from October 19-24, 2014 in Hinxton, UK. The deadline for this ... e details please visit the course description page.

- **8 May 2014** ToxoDB 11.0 Released
- **30 January 2014** ToxoDB 10.0 Released
- **14 January 2014** EuPathDB Annual Workshop

WellcomeTrust Events @WTevents 4 Jun
2 weeks to apply for Working with Parasite Database Resources workshop (19-24 October) Deadline 20 June bit.ly/MPSW2e @eupathdb

BLAST
Identify Sequence Similarities
Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates
Pathogen Portal
RNA sequence analysis, interactome maps and more
PubMed and Entrez
View the Latest Pubmed and Entrez Results
Genome Browser
View Sequences and Features in the genome browser
Ancillary Genome Browse
Access Probeset data and *Toxoplasma* Array info

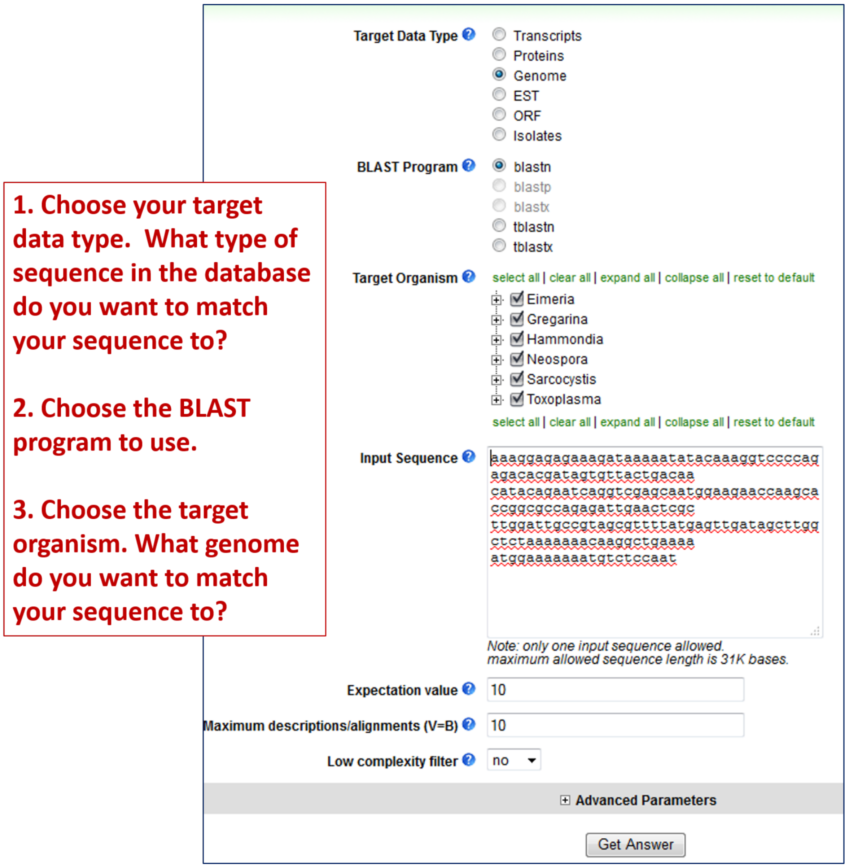
For additional tools, use the **Tools** *menu in the gray toolbar above....*

- 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
☑ Hammondia
☑ Neospora
☑ Sarcocystis
☑ Toxoplasma

Input Sequence:

```
aaaggagagaaagataaaaatatacaaaggtccccag
agacacgatagtgttactgacaa
catacagaatcaggtcgagcaatggaagaaccaagca
ccggcgccagagattgaactcgc
ttggattgccgtagcgttttatgagttgatagcttgg
ctctaaaaaaacaaggctgaaaa
atggaaaaaaatgtctccaat
```

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

Expectation value: 10

Maximum descriptions/alignments (V=B): 10

Low complexity filter: no

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?

- 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).

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 ID 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: 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).
   Expectation value : 10
   Maximum descriptions/alignments (V=B) : 100
j. Select the target organism. 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* 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*.
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* 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.