GENE PAGE EXERCISES
FINDING GENES, BUILDING SEARCH STRATEGIES AND VISITING A GENE PAGE

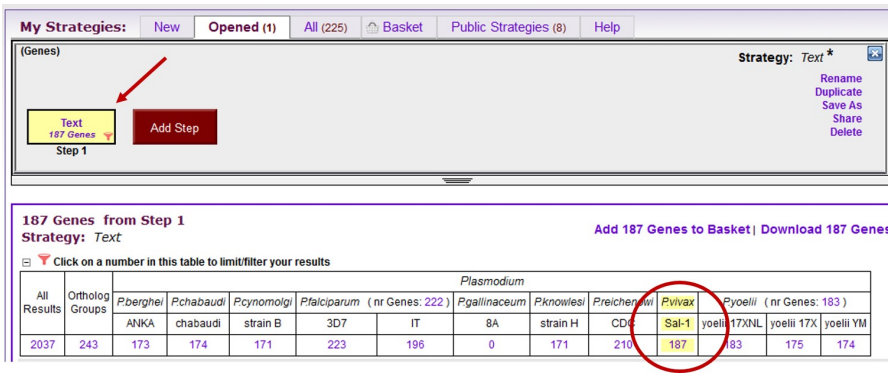
1. Finding a gene using text search.
For this exercise use http://www.plasmodb.org

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. vivax*? 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 genomes that you searched. Click on a number to display on that species' portion of the results.



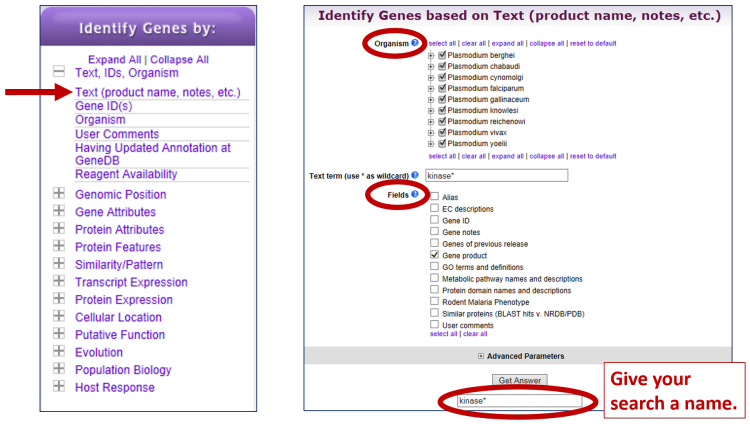
- What happens if you search using the term **kinases** in the Gene Text Search box? How many results are returned?

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

The search you ran in step 1.1a 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 in your search term will broaden your search (wild card = asterisk = any character). Use the full text search, the specific page where you can define the fields to be searched (Fields = Gene Product).

Try kinase* *kinase*

- 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?

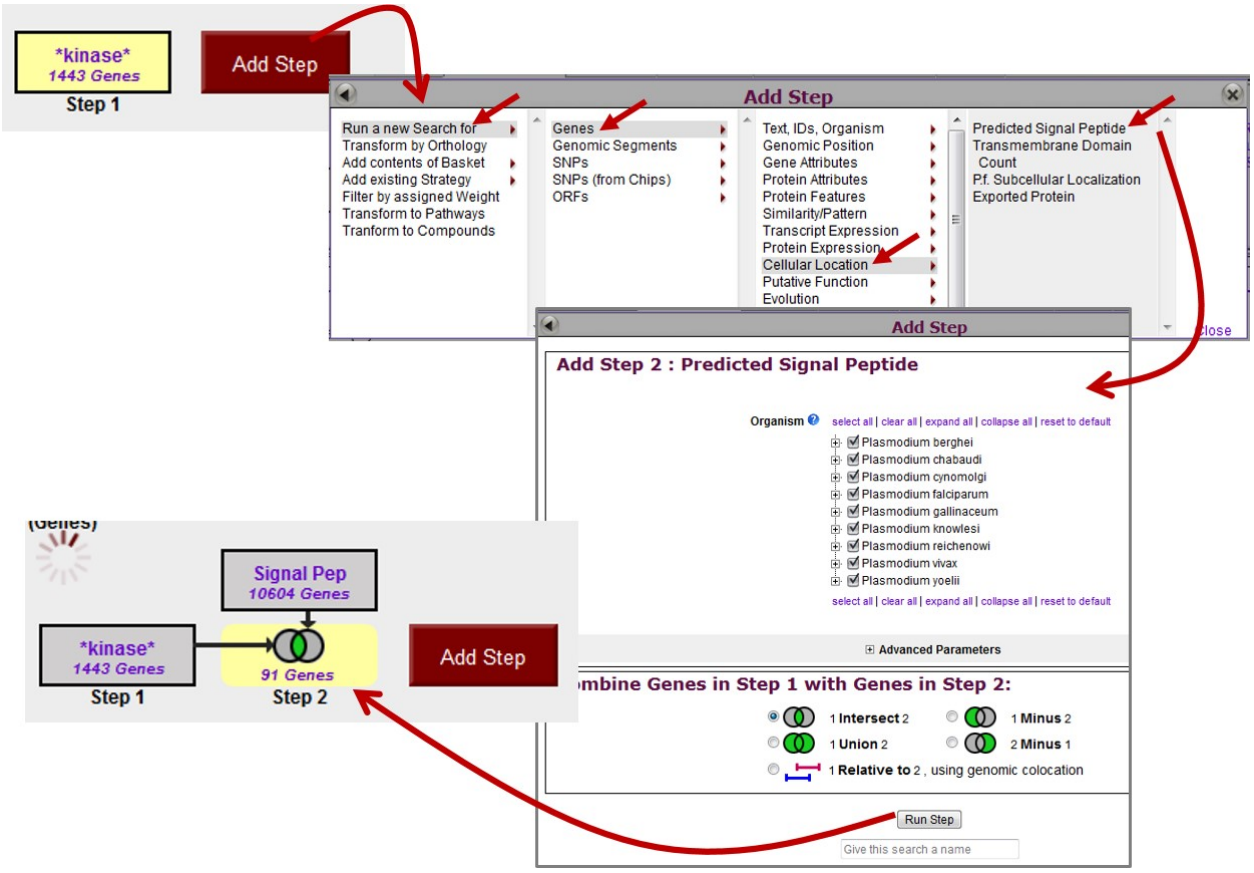
2. Combing text search results with results from other searches

a. Find kinase genes that are likely secreted.

In exercise 1.1b 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 *P. vivax* kinases are predicted to have a signal peptide? (use the filter table)



3. Visiting a specific gene page.

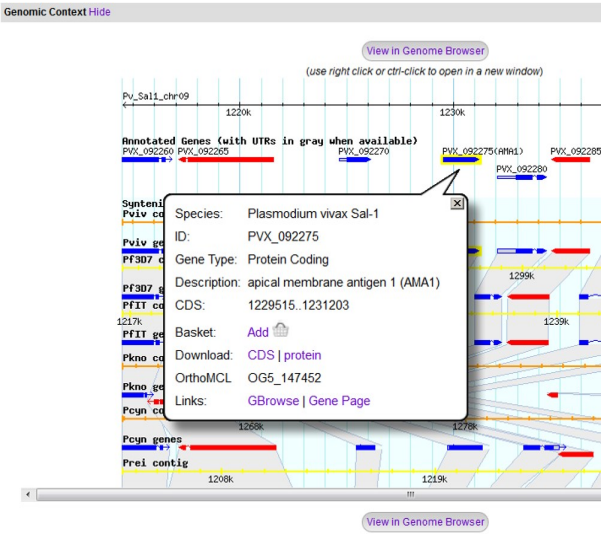
a. Find the gene page for one of the following *P. vivax* genes and explore the information there to answer these questions.

1. apical membrane antigen 1 gene (AMA1, PVX_092275)
2. merozoite surface protein 1 (MSP1, PVX_099980)

- 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)
- Does this gene have a user comment?

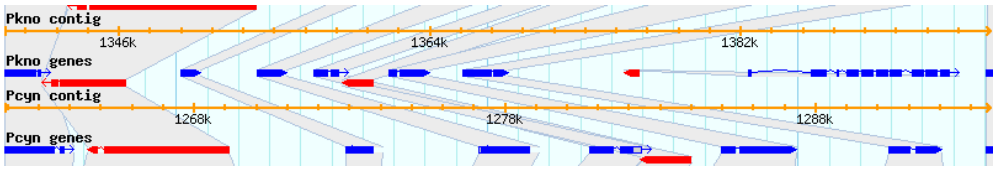
Look at the information on the gene page.

- 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?

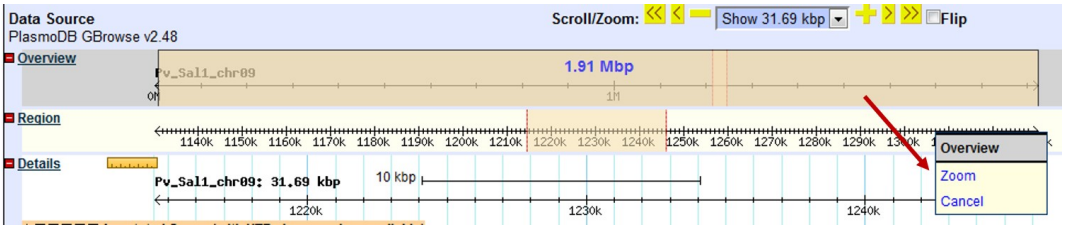


b. What genes are located upstream & downstream of AMA1 (MSP1) in *P. vivax*?

- 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. Do the contigs contain gaps or truncations?



- What does synteny look like across the entire chromosome? To do this:
  - Click on the **"View in GBrowse"** 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 Section and then select "zoom" from the popup menu.



- Click on the tab called "**Select tracks**". Select the track:
  Gene Models

B. Synteny

"**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 back by looking at synteny over large genomic regions?

c. Does the *P. vivax* AMA1 (or MSP1) gene contain any Single Nucleotide Polymorphisms (SNPs)?

SNPs are represented in a table called "SNP Overview".

- What is the total number of SNPs in the gene?
- How many impact the predicted protein sequence?
- (optional) Compare the SNP characteristics of this gene to upstream and downstream genes. How do these results compare with SNP distribution in other genes?
- (optional) You can create an alignment between isolates using the "Isolate Alignments in this Gene Region" that will highlight SNPs in pink. Try creating an alignment between China_LZCH-4 and Columbia_30102100437.

```
       Pv_Sal1_chr09 1229515 ATGAATAAAA TATACTACAT AATCTTTTTA AGCGCCCAGT GCCTTGTGCA CATTGGGAAG
        China_LZCH-4 1229515 .TGAATAAAA TATACTACAT AATCTTTTTA AGCGCCCAGT GCCTTGTGCA CATTGGGAAG
Columbia_30102100437 1229515 .TGAATAAAA TATACTACAT AATCTTTTTA AGCGCCCAGT GCCTTGTGCA CATTGGGAAG

       Pv_Sal1_chr09 1229595 GAGCAGGCTG ACCCGTAGCG CCAACAACGT TCTACTGGAA AAGGGGCCTA CCGTTGAGAG
        China_LZCH-4 1229594 GAGCAGGCTG ACCCGTAGCG CCAACAACGT TCTACTGGAA AAGGGGCCTA CCGTTGAGAG
Columbia_30102100437 1229594 GAGCAGGCTG ACCCGTAGCG CCAACAACGT TCTACTGGAA AAGGGGCCTA CCGTTGAGAG

       Pv_Sal1_chr09 1229675 CCTGGAAAGC GTTCATGGAA AAATACGACA TCGAAAGAAC ACACAGTTCT GGGGTTCGAG
        China_LZCH-4 1229674 CCTGGAAAGC GTTCATGGAA AAATACGACA TCGAAAGAAC ACACAGTTCT GGGGTTCGAG
Columbia_30102100437 1229674 CCTGGAAAGC GTTCATGGAA AAATACGACA TCGAAAGAAC ACACAGTTCT GGGGTTCGAG

       Pv_Sal1_chr09 1229755 GAAGTGGAAA ATGCAAAGTA CAGAATTCCA GCTGGAAGAT GTCCTGTTTT TGGAAAGGGT
        China_LZCH-4 1229754 GAAGTGGAAA ATGCAAAGTA CAGAATTCCA GCTGGAAGAT GTCCTGTTTT TGGAAAGGGT
Columbia_30102100437 1229754 GAAGTGGAAA ATGCAAAGTA CAGAATTCCA GCTGGAAGAT GTCCTGTTTT TGGAAAGGGT

       Pv_Sal1_chr09 1229835 CGTTAGCTTC TTAAGACCTG TGGCTACAGG AGATCAGAAG CTGAAGGATG GAGGTTTCGC
        China_LZCH-4 1229834 CGTTAGCTTC TTAAAACCTG TGGCTACAGG AGATCAGAGG CTGAAGGATG GAGGTTTCGC
Columbia_30102100437 1229834 TGTTAGCTTC TTAAAACCTG TGGCTACAGG AGATCAGAGG CTGAAGGATG GAGGTTTCGC

       Pv_Sal1_chr09 1229915 ATATCTCCCC AATGACATTA GCGAACCTTA AGGAAAGGTA TAAAGACAAT GTAGAGATGA
        China_LZCH-4 1229914 ATATCTCCCC CATGACAATA GCGAACCTTA AGGCAAGGTA TAAAGACAAT GTAGAGATGA
Columbia_30102100437 1229914 ATATCTCCCC CATGACAATA GCGAACCTTA AGGCAAGGTA TAAAGACAAT GTAGAGATGA

       Pv_Sal1_chr09 1229995 TTGTGCAGAA CCCACGCAGC TAGCTTTGTC ATGGCAGGGG ATCAAAATTC GTCCTACAGA
        China_LZCH-4 1229994 TTGTGCAGAA CCCACGCAGC TAGCTTTGTC ATGGCAGGGG ATCAAAATTC GTCCTACAGA
Columbia_30102100437 1229994 TTGTGCAGAA CCCACGCAGC TAGCTTTGTC ATGGCAGGGG ATCAAAATTC GTCCTACAGA

       Pv_Sal1_chr09 1230075 AAAGGAAAAA ACATGCCACA TGTTGTATTT ATCAGCGCAG GAAAATATGG GTCCGAGGTA
        China_LZCH-4 1230074 AAAGGAAAAA ACATGCCACA TGTTGTATTT ATCAGCGCAG GAAAATATGG GTCCGAGGTA
Columbia_30102100437 1230074 AAAGGAAAAA ACATGCCACA TGTTGTATTT ATCAGCGCAG GAAAATATGG GTCCGAGGTA

       Pv_Sal1_chr09 1230155 ATAGAGATGC CGTGTTCTGC TTCAAGCCAG ATAAAAATGA AAGCTTTGAA AACCTGGTGT
        China_LZCH-4 1230154 ATAGAGATGC CGTGTTCTGC TTCAAGCCAG ATAAAAATGT AGATTTTGAA AACCTGGTGT
Columbia_30102100437 1230154 ATAGAGATGC CTTGTTCTGC TTCAAGCCAG ATAAAAATGA AAGCTTTGAA AACCTGGTGT
```

d. Is the AMA1 (or MSP1) gene expressed?

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 expression?
- Is this gene expressed at the protein level in "Schizont proteome from human blood"? Look at the Protein Features graphic and the table of Mass Spec.-based Expression Evidence.
- Look at the Expression data track labeled **Intraerythrocytic developmental cycle of three isolates**. At what time point is AMA1 (MSP1) most abundant?