VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

VEuPathDB Annual Workshop

Pre-workshop module

June 2021

Note: the exercises in this pre-workshop module cover some of the basic functionality in VEuPathDB.

AmoebaDB Amoeba Informatics Resources
CryptoDB Cryptosporidium Informatics Resources
FungiDB Fungal & Oomycete Informatics Resources
GiardiaDB Giardia Informatics Resources
HostDB Pathogen Host Informatics Resources
MicrosporidiaDB Microsporidia Informatics Resources
OrthoMCL DB Ortholog Groups of Protein Sequences
PiroplasmaDB Piroplasma Informatics Resources
PlasmoDB Plasmodium Informatics Resources
ToxoDB Toxoplasma Informatics Resources
TrichDB Trichomonas Informatics Resources
TriTrypDB Kinetoplastid Informatics Resources
VectorBase Bioinformatics Resources for Invertebrate Vectors of Human Pathogens

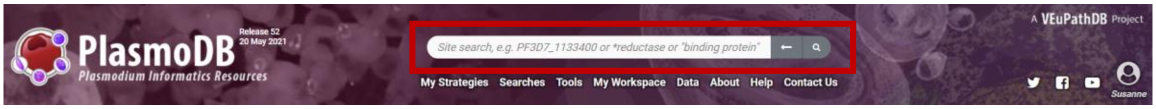
Site Search

Note: *this exercise uses PlasmoDB.org as an example database, but the same functionality is available on all VEuPathDB resources.*

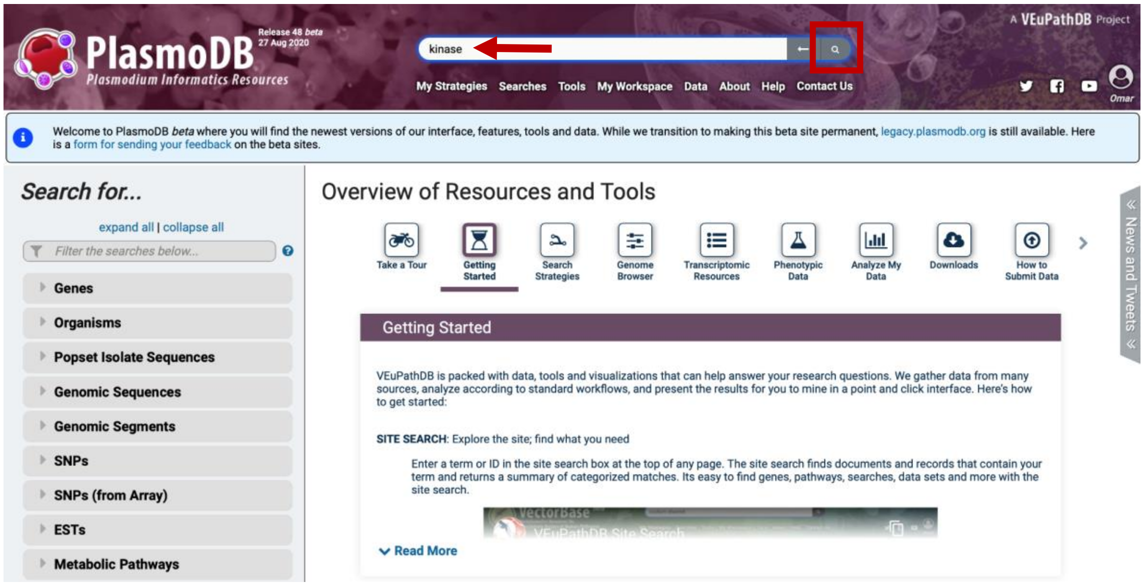
Learning objectives:

- Use keywords in site search
- Explore site search results
- Filter site search results by categories
- Filter site search results by organisms
- Filter site search results by category fields
- Export results to a search strategy
- Find a specific gene using its ID in site search

The site search is located in the header of any VEuPathDB site and is available from every page. The site search queries the databases for your term or ID and returns a list of pages and documents that contain your query term.



1. Enter the word *kinase* in the site search window (arrow in the image below). Then click enter on your keyboard or click on the search icon (square in the image below).



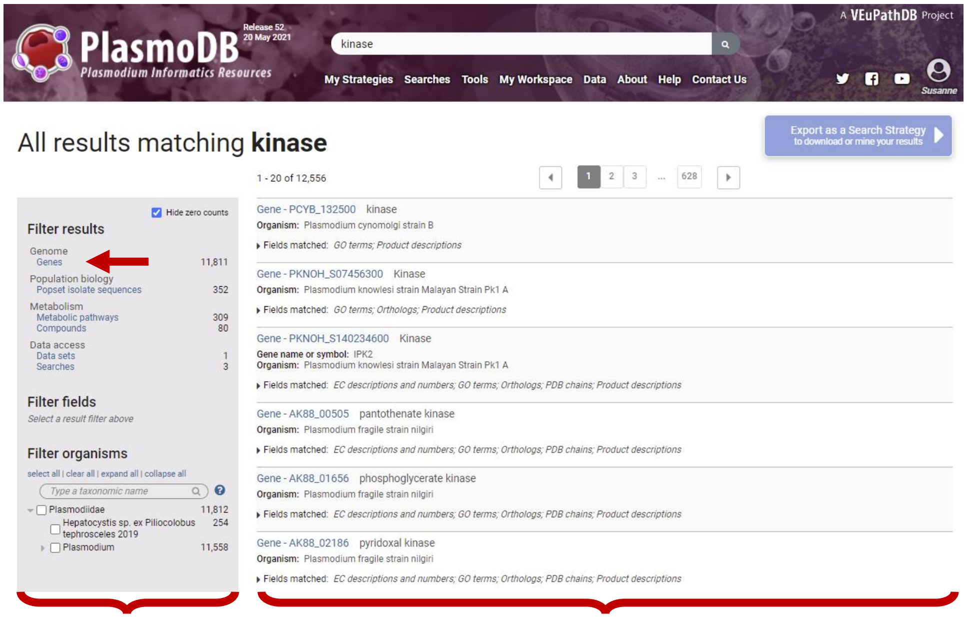
Welcome to PlasmoDB *beta* where you will find the newest versions of our interface, features, tools and data. While we transition to making this beta site permanent, legacy.plasmodb.org is still available. Here is a form for sending your feedback on the beta sites.

VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

2. The site search returns a categorized list of pages and documents that contain your term. Site search results are summarized by category, with a details panel on the right. Changing the panel on the left will populate the details panel with that list. What is the total number of results with the word kinase? Are all the results genes? Explore the filter panel on the left side of the webpage.



All results matching **kinase**

1 - 20 of 12,556

☑ Hide zero counts

Filter results

Genome	
Genes	11,811
Population biology	
Popset isolate sequences	352
Metabolism	
Metabolic pathways	309
Compounds	80
Data access	
Data sets	1
Searches	3

Filter fields

Select a result filter above

Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

□ Plasmodiidae	11,812
□ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	254
□ Plasmodium	11,558

Gene - PCYB_132500   kinase
Organism: Plasmodium cynomolgi strain B
Fields matched: *GO terms; Product descriptions*

Gene - PKNOH_S07456300   Kinase
Organism: Plasmodium knowlesi strain Malayan Strain Pk1 A
Fields matched: *GO terms; Orthologs; Product descriptions*

Gene - PKNOH_S140234600   Kinase
Gene name or symbol: IPK2
Organism: Plasmodium knowlesi strain Malayan Strain Pk1 A
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Gene - AK88_00505   pantothenate kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Gene - AK88_01656   phosphoglycerate kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

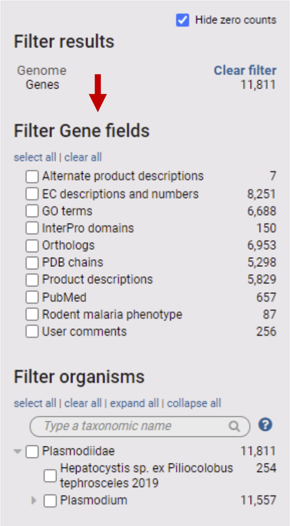
Gene - AK88_02186   pyridoxal kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Results summarized in categories

Details panel with information about each item returned

3. Filter the results so that you only view gene results (hint: click on the word *genes* in the *Filter results* section; arrow in image above). How many of the genes included the word kinase in their product descriptions?

   Notice that once you filter the result by genes (click on the *Genes* filter), the Filter Fields section expands to reveal additional filtering options. Select the *Product descriptions* field and Choose *Apply* this filter or cancel it (box middle panel below). Once a filter is applied it can be cleared by clicking on *Clear filter* (box left panel below).



☑ Hide zero counts

Filter results

Genome Genes	Clear filter 11,811

Filter Gene fields

select all | clear all

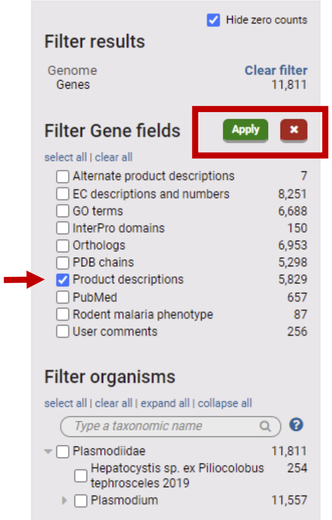
☐ Alternate product descriptions	7
☐ EC descriptions and numbers	8,251
☐ GO terms	6,688
☐ InterPro domains	150
☐ Orthologs	6,953
☐ PDB chains	5,298
☐ Product descriptions	5,829
☐ PubMed	657
☐ Rodent malaria phenotype	87
☐ User comments	256

Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

☐ Plasmodiidae	11,811
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	254
☐ Plasmodium	11,557



☑ Hide zero counts

Filter results

Genome Genes	Clear filter 11,811

Filter Gene fields

select all | clear all

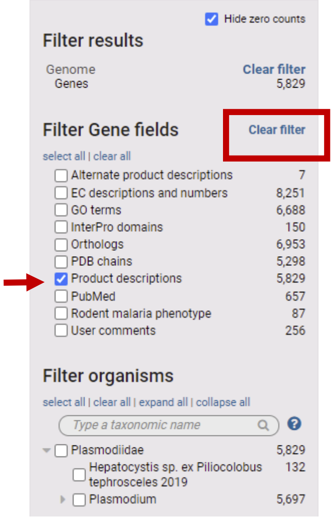
☐ Alternate product descriptions	7
☐ EC descriptions and numbers	8,251
☐ GO terms	6,688
☐ InterPro domains	150
☐ Orthologs	6,953
☐ PDB chains	5,298
☑ Product descriptions	5,829
☐ PubMed	657
☐ Rodent malaria phenotype	87
☐ User comments	256

Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

☐ Plasmodiidae	11,811
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	254
☐ Plasmodium	11,557



☑ Hide zero counts

Filter results

Genome Genes	Clear filter 5,829

Filter Gene fields

select all | clear all

☐ Alternate product descriptions	7
☐ EC descriptions and numbers	8,251
☐ GO terms	6,688
☐ InterPro domains	150
☐ Orthologs	6,953
☐ PDB chains	5,298
☑ Product descriptions	5,829
☐ PubMed	657
☐ Rodent malaria phenotype	87
☐ User comments	256

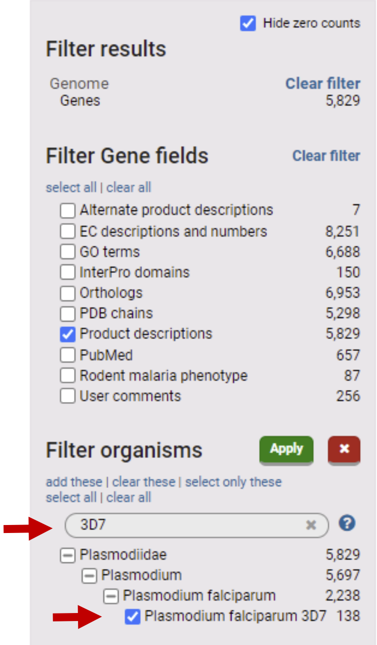
Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

☐ Plasmodiidae	5,829
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	132
☐ Plasmodium	5,697

4. How many of the above genes are found in *Plasmodium falciparum* 3D7? How did you find this number? Hint: explore the *Filter organisms* section of the results filter. There is a search option to aid navigation through the organism tree (left) or the tree can be expanded to find the organism of interest (right). Select the correct organism and apply the filter.



☑ Hide zero counts

Filter results

Genome Genes	Clear filter 5,829

Filter Gene fields

select all | clear all

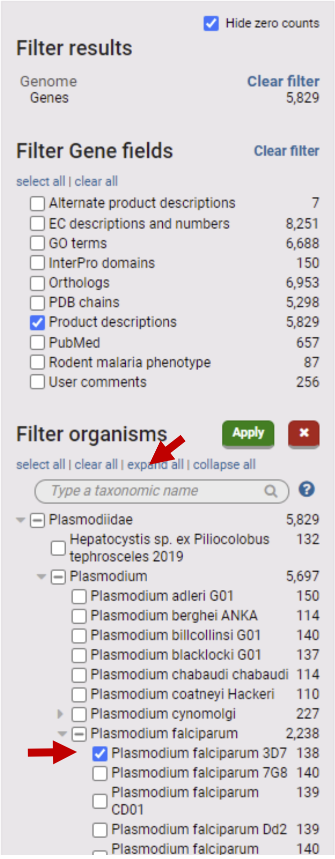
☐ Alternate product descriptions	7
☐ EC descriptions and numbers	8,251
☐ GO terms	6,688
☐ InterPro domains	150
☐ Orthologs	6,953
☐ PDB chains	5,298
☑ Product descriptions	5,829
☐ PubMed	657
☐ Rodent malaria phenotype	87
☐ User comments	256

Filter organisms

add these | clear these | select only these
select all | clear all

3D7

⊟ Plasmodiidae	5,829
⊟ Plasmodium	5,697
⊟ Plasmodium falciparum	2,238
☑ Plasmodium falciparum 3D7	138



☑ Hide zero counts

Filter results

Genome Genes	Clear filter 5,829

Filter Gene fields

select all | clear all

☐ Alternate product descriptions	7
☐ EC descriptions and numbers	8,251
☐ GO terms	6,688
☐ InterPro domains	150
☐ Orthologs	6,953
☐ PDB chains	5,298
☑ Product descriptions	5,829
☐ PubMed	657
☐ Rodent malaria phenotype	87
☐ User comments	256

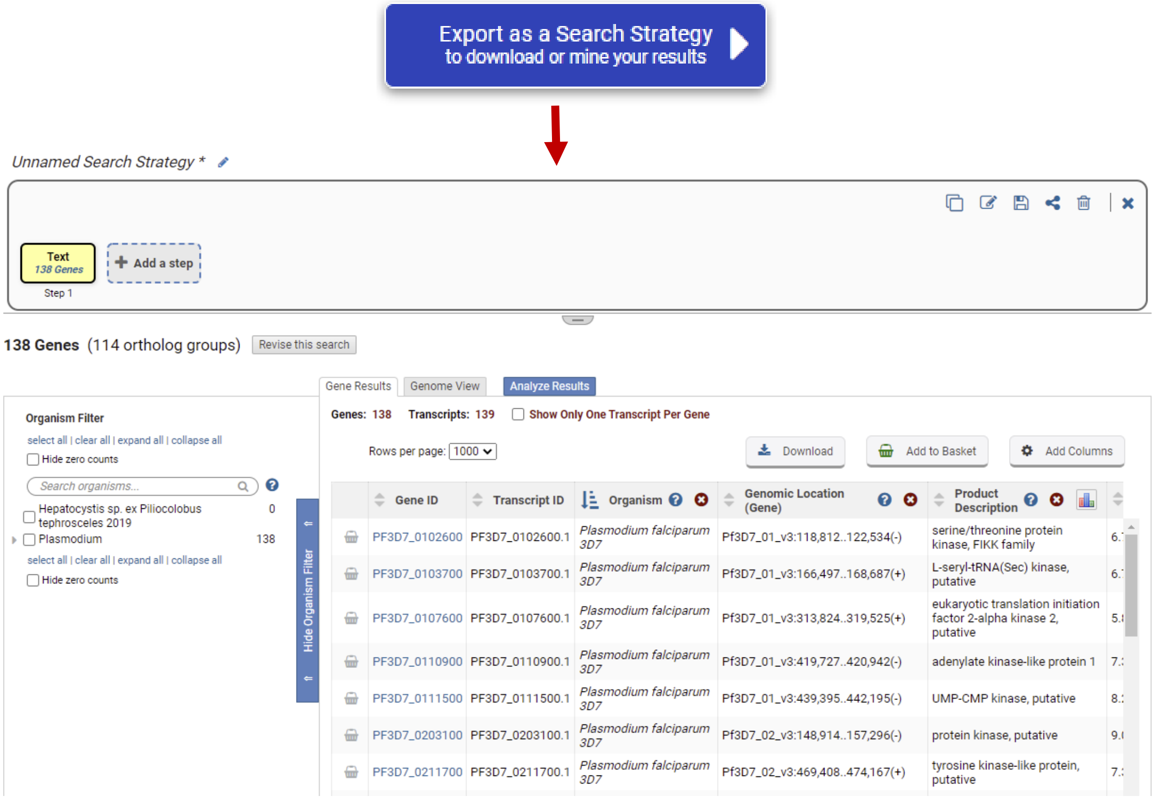
Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

⊟ Plasmodiidae	5,829
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	132
⊟ Plasmodium	5,697
☐ Plasmodium adleri G01	150
☐ Plasmodium berghei ANKA	114
☐ Plasmodium billcollinsi G01	140
☐ Plasmodium blacklocki G01	137
☐ Plasmodium chabaudi chabaudi	114
☐ Plasmodium coatneyi Hackeri	110
☐ Plasmodium cynomolgi	227
⊟ Plasmodium falciparum	2,238
☑ Plasmodium falciparum 3D7	138
☐ Plasmodium falciparum 7G8	140
☐ Plasmodium falciparum CD01	139
☐ Plasmodium falciparum Dd2	139
Plasmodium falciparum	140

5. Export the results to a search strategy. Use the blue *Export as a search strategy* button at the top right-hand side of the results.



Export as a Search Strategy
to download or mine your results
Unnamed Search Strategy *
Text
138 Genes
Step 1
+ Add a step

138 Genes (114 ortholog groups) Revise this search

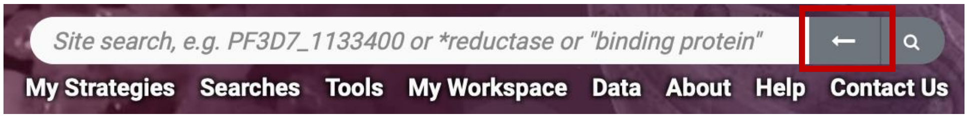
□ Hepatocystis sp. ex Piliocolobus tephrosceles 2019 0
□ Plasmodium 138

Genes: 138 Transcripts: 139 □ Show Only One Transcript Per Gene

Rows per page: 1000

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	
PF3D7_0102600	PF3D7_0102600.1	*Plasmodium falciparum 3D7*	Pf3D7_01_v3:118,812..122,534(-)	serine/threonine protein kinase, FIKK family	6.
PF3D7_0103700	PF3D7_0103700.1	*Plasmodium falciparum 3D7*	Pf3D7_01_v3:166,497..168,687(+)	L-seryl-tRNA(Sec) kinase, putative	6.
PF3D7_0107600	PF3D7_0107600.1	*Plasmodium falciparum 3D7*	Pf3D7_01_v3:313,824..319,525(+)	eukaryotic translation initiation factor 2-alpha kinase 2, putative	5.
PF3D7_0110900	PF3D7_0110900.1	*Plasmodium falciparum 3D7*	Pf3D7_01_v3:419,727..420,942(-)	adenylate kinase-like protein 1	7.
PF3D7_0111500	PF3D7_0111500.1	*Plasmodium falciparum 3D7*	Pf3D7_01_v3:439,395..442,195(-)	UMP-CMP kinase, putative	8.
PF3D7_0203100	PF3D7_0203100.1	*Plasmodium falciparum 3D7*	Pf3D7_02_v3:148,914..157,296(-)	protein kinase, putative	9.
PF3D7_0211700	PF3D7_0211700.1	*Plasmodium falciparum 3D7*	Pf3D7_02_v3:469,408..474,167(+)	tyrosine kinase-like protein, putative	7.

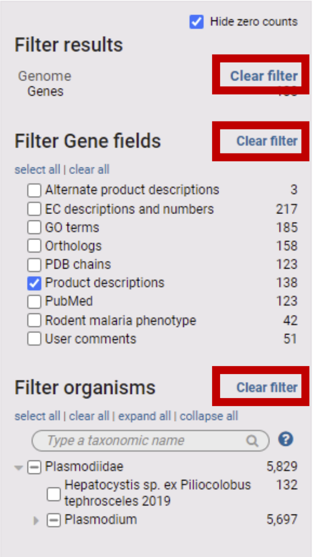
6. Return to the site search results page. You can achieve this in two ways: 1. Your previous results and filter settings were preserved and can be accessed by clicking on the ‘back to results’ arrow in the site search window. 2. Click on your browser’s back arrow. Notice that



Site search, e.g. PF3D7_1133400 or *reductase or "binding protein"
My Strategies
Searches
Tools
My Workspace
Data
About
Help
Contact Us

7. Clear all filters. You can achieve this in two ways: 1. You can click on each of the clear filter options in the filter results panel (boxes below). 2. You can click on the *clear filters option* in the site search window, which serves to Clear All filters.

1



☑ Hide zero counts

Filter results — Clear filter

Genome
Genes

Filter Gene fields — Clear filter

select all | clear all

☐ Alternate product descriptions	3
☐ EC descriptions and numbers	217
☐ GO terms	185
☐ Orthologs	158
☐ PDB chains	123
☑ Product descriptions	138
☐ PubMed	123
☐ Rodent malaria phenotype	42
☐ User comments	51

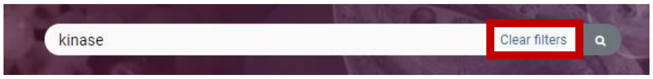
Filter organisms — Clear filter

select all | clear all | expand all | collapse all

Type a taxonomic name

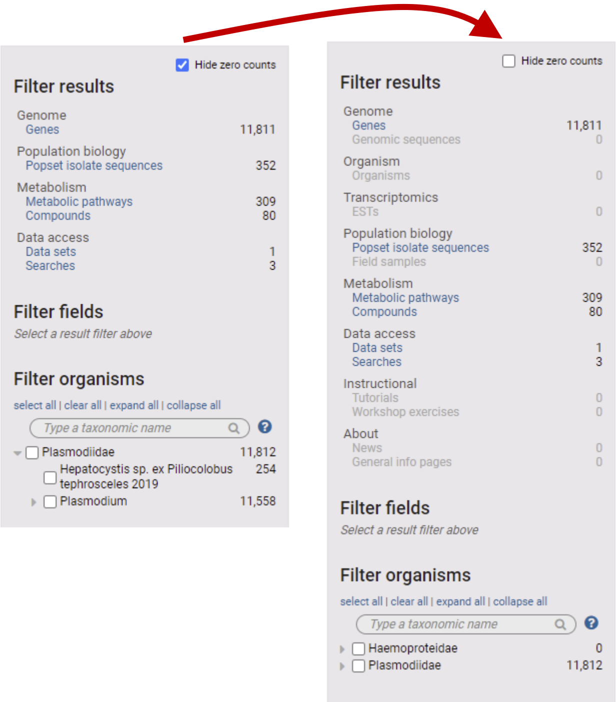
⊟ Plasmodiidae	5,829
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	132
⊟ Plasmodium	5,697

2



kinase — Clear filters

8. Click the *Hide zero counts* check box in the *Filter results* panel. What does this do?



☑ Hide zero counts

Filter results

Genome	
Genes	11,811
Population biology	
Popset isolate sequences	352
Metabolism	
Metabolic pathways	309
Compounds	80
Data access	
Data sets	1
Searches	3

Filter fields

Select a result filter above

Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

☐ Plasmodiidae	11,812
☐ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	254
☐ Plasmodium	11,558

☐ Hide zero counts

Filter results

Genome	
Genes	11,811
Genomic sequences	0
Organism	
Organisms	0
Transcriptomics	
ESTs	0
Population biology	
Popset isolate sequences	352
Field samples	0
Metabolism	
Metabolic pathways	309
Compounds	80
Data access	
Data sets	1
Searches	3
Instructional	
Tutorials	0
Workshop exercises	0
About	
News	0
General info pages	0

Filter fields

Select a result filter above

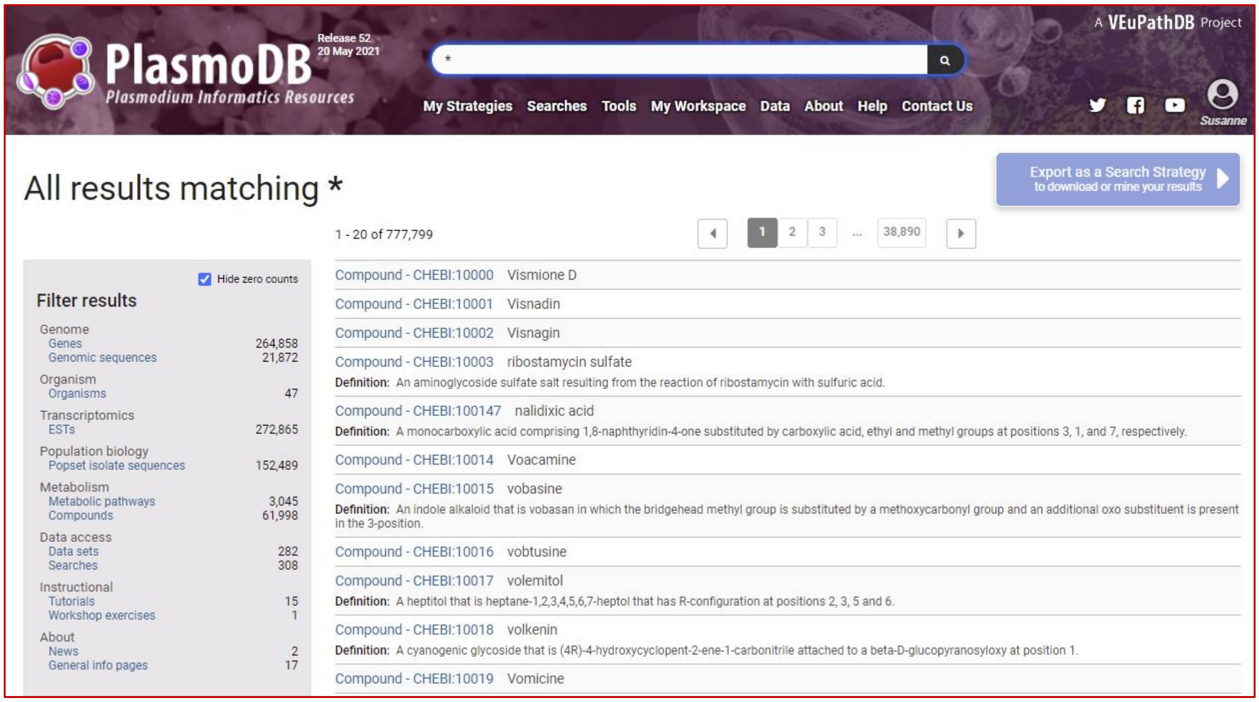
Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

☐ Haemoproteidae	0
☐ Plasmodiidae	11,812

9. Try running a search with a wild card. The wild card is denoted by an asterisk *. The wild card can be used alone to retrieve all results available to the site search or combined with a word such as **kinase* to retrieve compound words ending with the word kinase like phosphofructokinase. As usual results can then be explored using the filters in the *Results filter* on the left side of the website.



All results matching *

1 - 20 of 777,799

Filter results

Genome	
Genes	264,858
Genomic sequences	21,872
Organism	
Organisms	47
Transcriptomics	
ESTs	272,865
Population biology	
Popset isolate sequences	152,489
Metabolism	
Metabolic pathways	3,045
Compounds	61,998
Data access	
Data sets	282
Searches	308
Instructional	
Tutorials	15
Workshop exercises	1
About	
News	2
General info pages	17

Compound - CHEBI:10000 Vismione D

Compound - CHEBI:10001 Visnadin

Compound - CHEBI:10002 Visnagin

Compound - CHEBI:10003 ribostamycin sulfate
Definition: An aminoglycoside sulfate salt resulting from the reaction of ribostamycin with sulfuric acid.

Compound - CHEBI:100147 nalidixic acid
Definition: A monocarboxylic acid comprising 1,8-naphthyridin-4-one substituted by carboxylic acid, ethyl and methyl groups at positions 3, 1, and 7, respectively.

Compound - CHEBI:10014 Voacamine

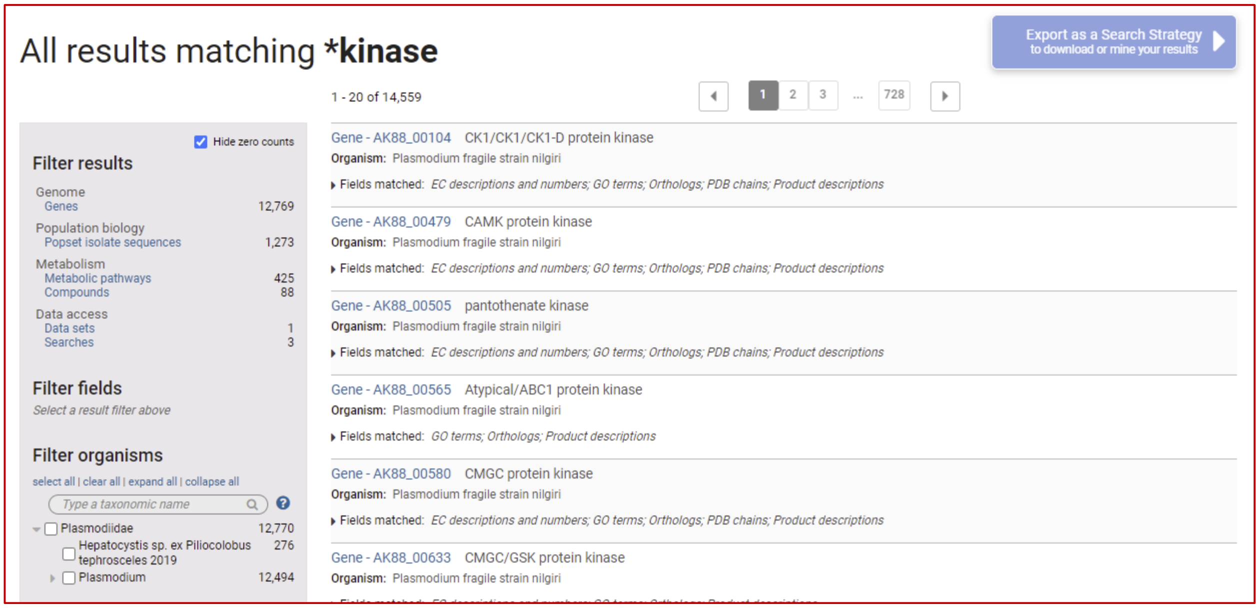
Compound - CHEBI:10015 vobasine
Definition: An indole alkaloid that is vobasan in which the bridgehead methyl group is substituted by a methoxycarbonyl group and an additional oxo substituent is present in the 3-position.

Compound - CHEBI:10016 vobtusine

Compound - CHEBI:10017 volemitol
Definition: A heptitol that is heptane-1,2,3,4,5,6,7-heptol that has R-configuration at positions 2, 3, 5 and 6.

Compound - CHEBI:10018 volkenin
Definition: A cyanogenic glycoside that is (4R)-4-hydroxycyclopent-2-ene-1-carbonitrile attached to a beta-D-glucopyranosyloxy at position 1.

Compound - CHEBI:10019 Vomicine


Export as a Search Strategy
to download or mine your results
1
2
3
...
728
☑ Hide zero counts
select all | clear all | expand all | collapse all
Type a taxonomic name

All results matching ***kinase**

1 - 20 of 14,559

Filter results

Genome	
Genes	12,769
Population biology	
Popset isolate sequences	1,273
Metabolism	
Metabolic pathways	425
Compounds	88
Data access	
Data sets	1
Searches	3

Filter fields

Select a result filter above

Filter organisms

□ Plasmodiidae	12,770
□ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	276
□ Plasmodium	12,494

Gene - AK88_00104 CK1/CK1/CK1-D protein kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Gene - AK88_00479 CAMK protein kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

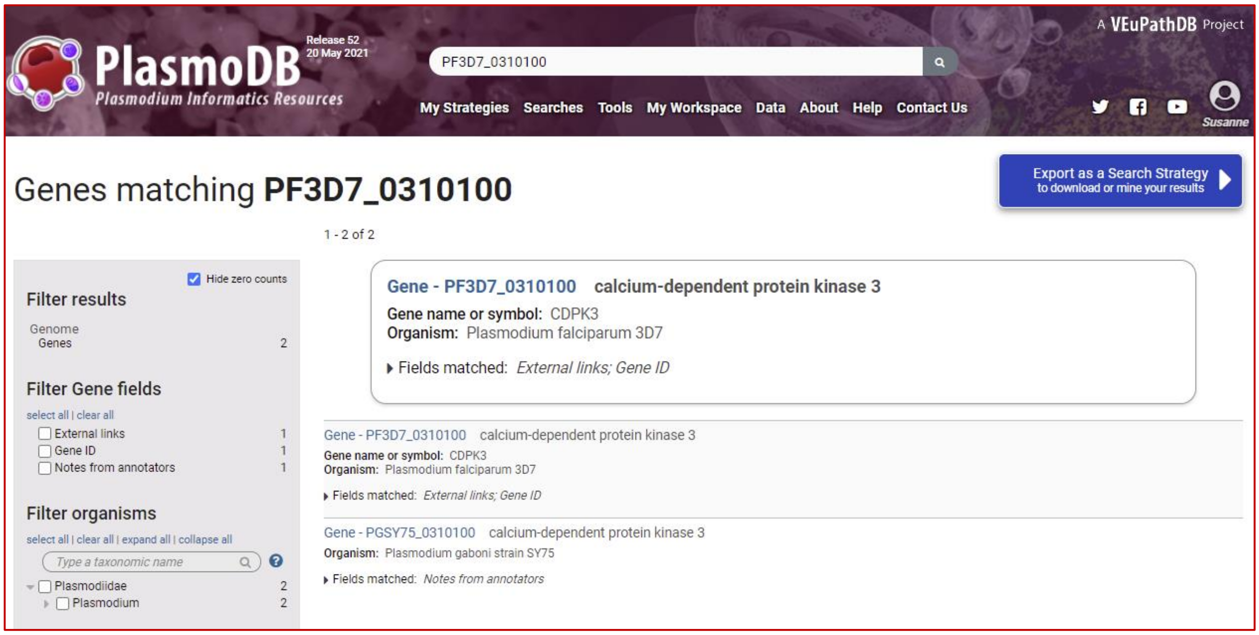
Gene - AK88_00505 pantothenate kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Gene - AK88_00565 Atypical/ABC1 protein kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *GO terms; Orthologs; Product descriptions*

Gene - AK88_00580 CMGC protein kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; PDB chains; Product descriptions*

Gene - AK88_00633 CMGC/GSK protein kinase
Organism: Plasmodium fragile strain nilgiri
Fields matched: *EC descriptions and numbers; GO terms; Orthologs; Product descriptions*

10. Try searching for a specific gene ID. Enter the gene ID below in the site search window: *PF3D7_0310100*



Genes matching **PF3D7_0310100**

1 - 2 of 2

☑ Hide zero counts

Filter results

Genome
Genes 2

Filter Gene fields

select all | clear all

□ External links 1
□ Gene ID 1
□ Notes from annotators 1

Filter organisms

select all | clear all | expand all | collapse all

Type a taxonomic name

□ Plasmodiidae 2
□ Plasmodium 2

Gene - PF3D7_0310100 calcium-dependent protein kinase 3
Gene name or symbol: CDPK3
Organism: Plasmodium falciparum 3D7
▸ Fields matched: *External links; Gene ID*

Gene - PF3D7_0310100 calcium-dependent protein kinase 3
Gene name or symbol: CDPK3
Organism: Plasmodium falciparum 3D7
▸ Fields matched: *External links; Gene ID*

Gene - PGSY75_0310100 calcium-dependent protein kinase 3
Organism: Plasmodium gaboni strain SY75
▸ Fields matched: *Notes from annotators*

When the query ID has an exact match in the database, the site search returns a card at the top of the details panel for easy access to the gene page. The site search also returns other pages that contain the query ID. Click on the Gene ID to go the gene page.

Exploring the Gene Page

Note: *this exercise uses ToxoDB (https://ToxoDB.org) as an example database, but the same functionality is available on all VEuPathDB resources.*

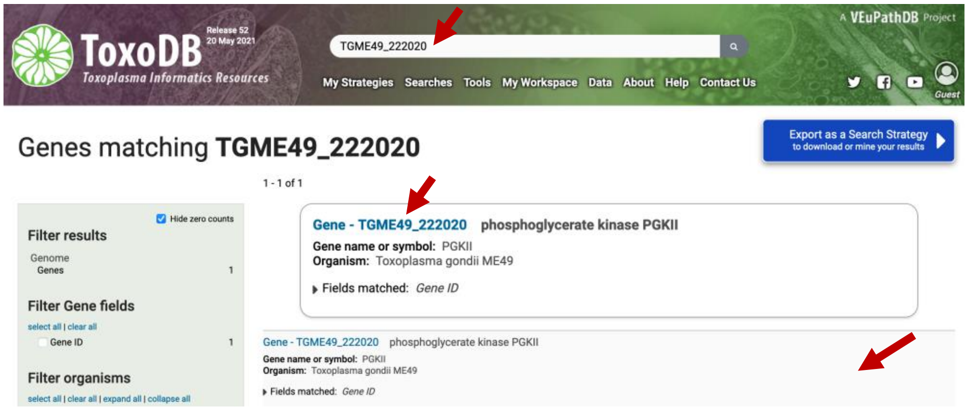
Learning objectives

Gene pages:

- Become familiar with the information in gene pages
- Navigate to and from the gene pages
- Use the contents section of the gene page
- Interact with gene page subsections

1. **Navigation to the Gene pages**

   For this exercise visit the gene page for TGME49_222020 (phosphoglycerate kinase PGKII). How did you get to this gene? (hint: copy and paste the ID in the site search, then click on the gene ID in the results.



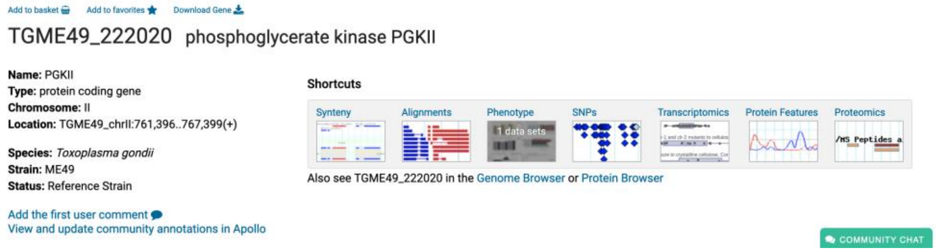
Genes matching **TGME49_222020**

1 - 1 of 1

Gene - TGME49_222020 **phosphoglycerate kinase PGKII**
Gene name or symbol: PGKII
Organism: Toxoplasma gondii ME49
▸ Fields matched: *Gene ID*

Gene - TGME49_222020 phosphoglycerate kinase PGKII
Gene name or symbol: PGKII
Organism: Toxoplasma gondii ME49
▸ Fields matched: *Gene ID*

2. **Explore the top section of the gene page**
   - What information is in this section?
   - Can you easily find which chromosome this gene is located on?
   - Is this gene protein coding?
   - What do the shortcuts do?



TGME49_222020 phosphoglycerate kinase PGKII

Name: PGKII
Type: protein coding gene
Chromosome: II
Location: TGME49_chrII:761,396..767,399(+)

Species: *Toxoplasma gondii*
Strain: ME49
Status: Reference Strain

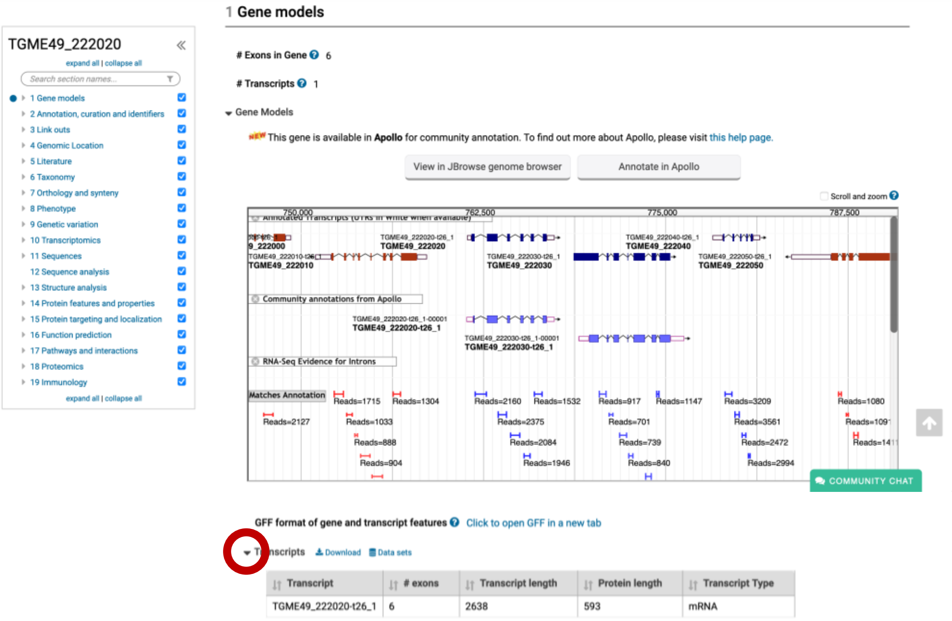
Add the first user comment
View and update community annotations in Apollo

Also see TGME49_222020 in the Genome Browser or Protein Browser

3. **Explore the gene model section.**

   Scroll down to the gene model section of the gene page.
   - What direction is the transcript relative to the chromosome?

- Does the gene have UTRs?
- How many exons does the gene have?
- Does this gene have an available community annotation?
- How long is the transcript? You can determine transcript length by expanding the Transcripts section.



1 Gene models

Exons in Gene 6

Transcripts 1

Gene Models

NEW This gene is available in **Apollo** for community annotation. To find out more about Apollo, please visit this help page.

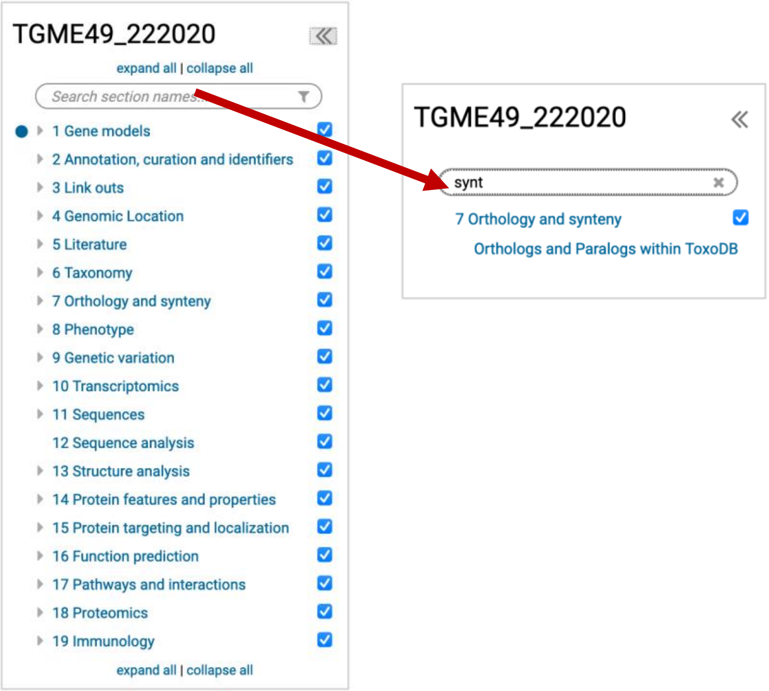
GFF format of gene and transcript features Click to open GFF in a new tab

Transcripts

Transcript	# exons	Transcript length	Protein length	Transcript Type
TGME49_222020-t26_1	6	2638	593	mRNA

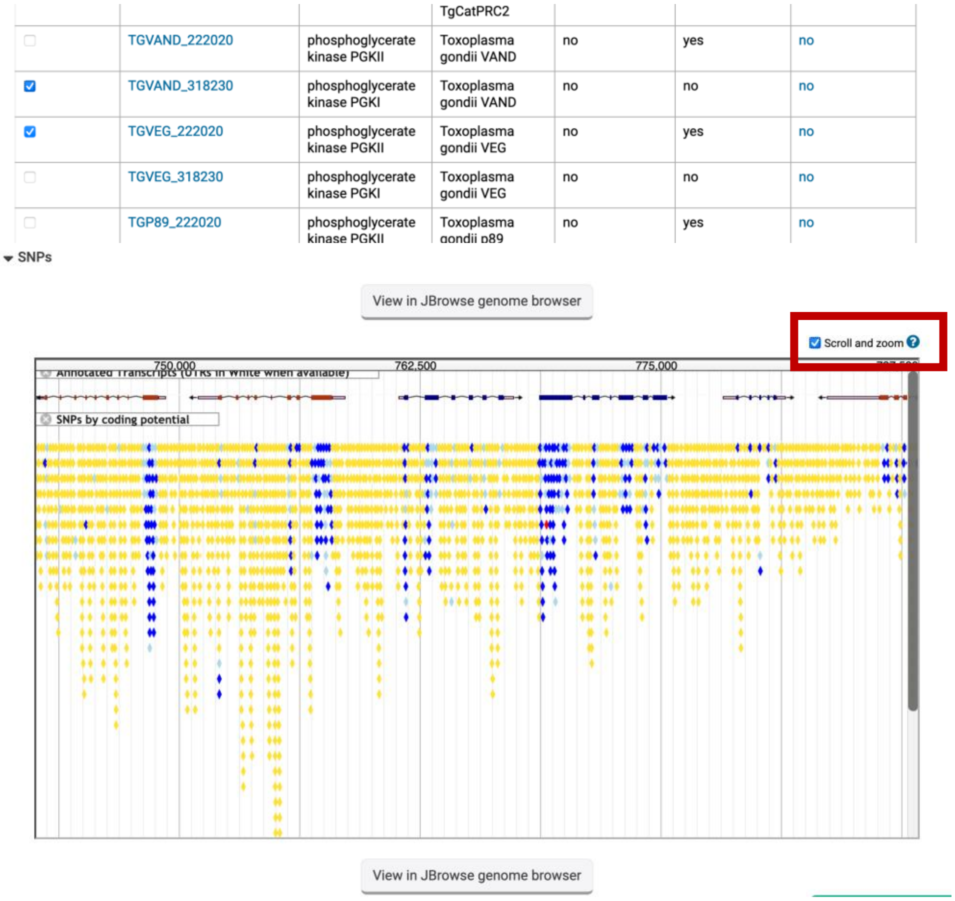
4. Content navigation.

How do you find/navigate to the different sections of the page? Use the "Contents" menu on the left side, type a keyword and cl ick on the menu, click on the work to



navigate to it on the page. In the example below the word “synteny” is used. You can also click on the images in the Shortcuts section in the top of the page.

5. **Running an alignment of selected sequences**
   a. Expand the “Orthologs and Paralogs in ToxoDB” section.
   b. Select a few genes from the table using the checkbox.
   c. Scroll to the bottom of the table and click on the Run Clustal Omega button.
6. **Exploring the genetic variation section**



			TgCatPRC2			
□	TGVAND_222020	phosphoglycerate kinase PGKII	Toxoplasma gondii VAND	no	yes	no
☑	TGVAND_318230	phosphoglycerate kinase PGKI	Toxoplasma gondii VAND	no	no	no
☑	TGVEG_222020	phosphoglycerate kinase PGKII	Toxoplasma gondii VEG	no	yes	no
□	TGVEG_318230	phosphoglycerate kinase PGKI	Toxoplasma gondii VEG	no	no	no
□	TGP89_222020	phosphoglycerate kinase PGKII	Toxoplasma gondii p89	no	yes	no

SNPs
View in JBrowse genome browser
☑ Scroll and zoom
750,000
762,500
775,000
787,500
Annotated Transcripts (UTRs in white when available)
SNPs by coding potential
View in JBrowse genome browser

Go to the Genetic variation section of the gene page and expand the SNP section. Notice that by default you cannot scroll within the embedded browser window. To enable scrolling, select the “Scroll and Zoom” check box in the upper right-hand side of the browser window. To scroll down within the browser window, you click and drag or use two-finger scrolling. You can also double click in an area to zoom in.
SNP color code: Dark blue (non-synonymous), light blue (synonymous), Yellow (non-coding), Red (nonsense). What kind of SNPs are in this gene? Can you see any non-synonymous SNPs? How does this compare to the neighboring genes?

7. **Explore other sections of the gene page.**
   Feel free to scroll around the gene page and ask questions for clarification. Here are some questions you may want to ask about this gene:
   a. Is there evidence that this protein is phosphorylated? (hint: go to the proteomics section and expand the Post Translational Modification section).

b. Where is the protein localized? (hint: go to the Protein Targeting and Localization section and expand the cellular localization section).
c. Is there any phenotypic data available for this gene? (hint: go to the Phenotype section and expand its subsections).
d. Is there any RNA-Seq data available for this gene? (hint: go to the Transcriptomics section and expand the subsections called RNA-Seq transcription summary and Transcript Expression).

JBrowse Basics

Note: *this exercise uses TriTrypDB (https://TriTrypdb.org) as an example database, but the same functionality is available on all VEuPathDB resources.*

Learning objectives:

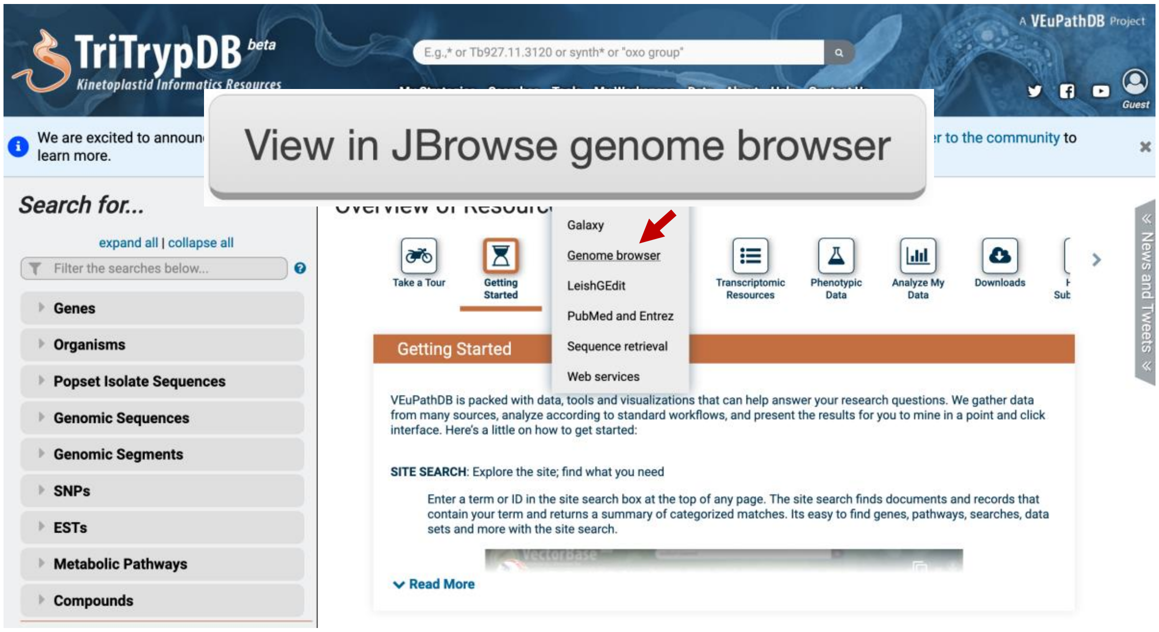
- Navigate to the genome browser
- Use the menu and navigation bars
- Run searches
- Add pre-loaded data tracks
- Upload your own data tracks
- Configure tracks
- Download track data

1. Navigating to the Genome Browser (JBrowse)

JBrowse is a fast and full-featured genome browser built with JavaScript and HTML5. You can read more about JBrowse and its features here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4830012/

Links to the genome browser are available from multiple locations:

a. The tools menu in the banner of any page.



VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's a little on how to get started:

SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

b. From record pages such as gene, SNP or genomic sequence pages – these links are usually to a specific JBrowse configuration that includes data relevant to the section on that record page. For example, a JBrowse link from an RNAseq dataset on the gene page would display the gene of interest along with the RNAseq data as density or coverage plots. These links are usually indicated by "View in JBrowse genome browser" button.

2. Getting around JBrowse.

a. Use any of the above described JBrowse linking strategies to get to the genome browser.

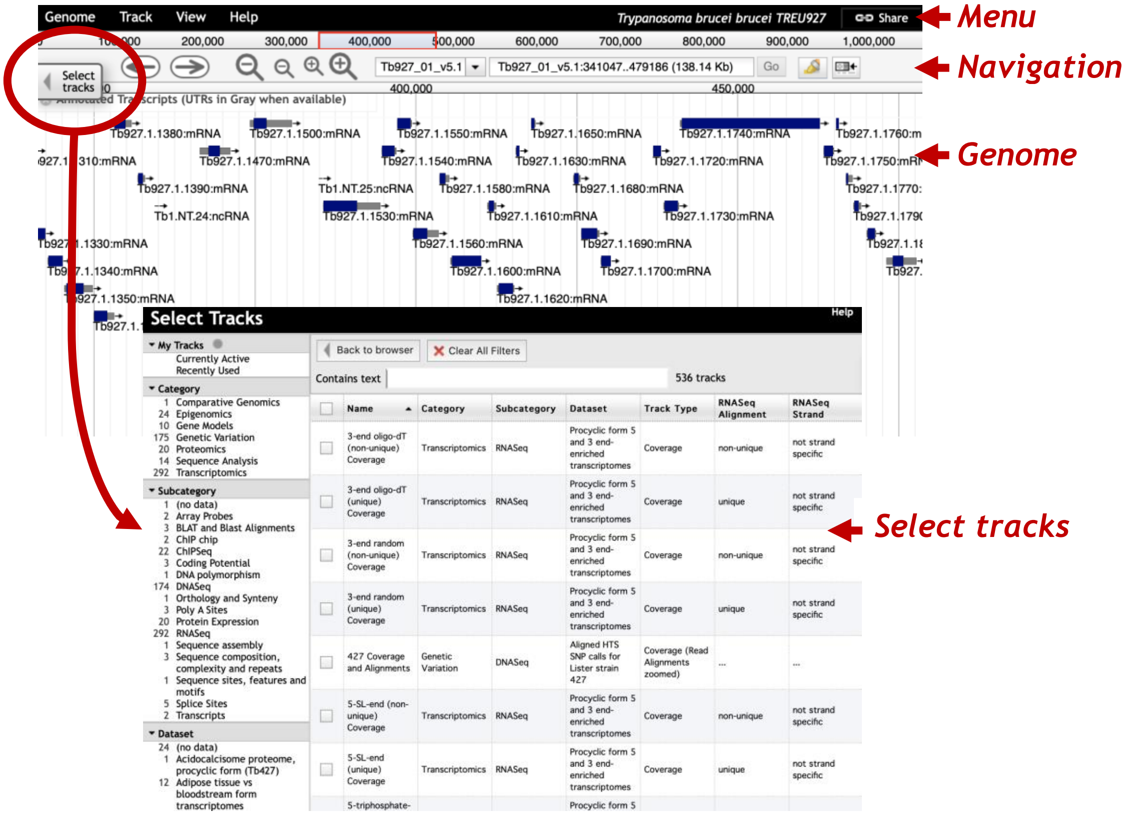
b. Once in JBrowse examine the following features:

   i. The **menu bar**: located at the top of the JBrowse frame. This includes the Genome menu, Track menu, View menu, Help menu and the Sharing link. What do each of these do/provide?

   ii. The **navigation bar**: located below the menu bar. This contains zooming (magnifying glass icons), panning (left/right arrows) and highlighting (yellow highlighter) buttons, reference sequence selector (drop down with sequences from the selected genome sorted by length), a text box to search for features such as gene IDs and overview bar which shows the location of the region in view.

   iii. The **genome view**: this is where the data tracks are displayed.

c. Selecting tracks: click on the "select track" button (top left). You can search/filter for tracks and then select them for display by checking the check box next to the track name.



Select Tracks

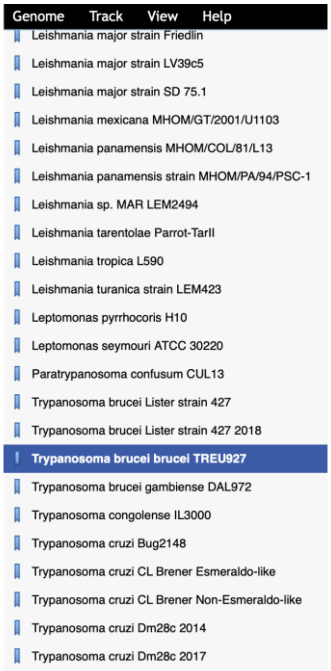
- My Tracks
  - Currently Active
  - Recently Used
- Category
  - 1 Comparative Genomics
  - 24 Epigenomics
  - 10 Gene Models
  - 175 Genetic Variation
  - 20 Proteomics
  - 14 Sequence Analysis
  - 292 Transcriptomics
- Subcategory
  - 1 (no data)
  - 2 Array Probes
  - 3 BLAT and Blast Alignments
  - 2 ChIP chip
  - 22 ChIPSeq
  - 3 Coding Potential
  - 1 DNA polymorphism
  - 174 DNASeq
  - 1 Orthology and Synteny
  - 3 Poly A Sites
  - 20 Protein Expression
  - 292 RNASeq
  - 1 Sequence assembly
  - 3 Sequence composition, complexity and repeats
  - 1 Sequence sites, features and motifs
  - 5 Splice Sites
  - 2 Transcripts
- Dataset
  - 24 (no data)
  - 1 Acidocalcisome proteome, procyclic form (Tb427)
  - 12 Adipose tissue vs bloodstream form transcriptomes

Contains text 536 tracks

□	Name	Category	Subcategory	Dataset	Track Type	RNASeq Alignment	RNASeq Strand
□	3-end oligo-dT (non-unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	non-unique	not strand specific
□	3-end oligo-dT (unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	unique	not strand specific
□	3-end random (non-unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	non-unique	not strand specific
□	3-end random (unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	unique	not strand specific
□	427 Coverage and Alignments	Genetic Variation	DNASeq	Aligned HTS SNP calls for Lister strain 427	Coverage (Read Alignments zoomed)	...	...
□	5-SL-end (non-unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	non-unique	not strand specific
□	5-SL-end (unique) Coverage	Transcriptomics	RNASeq	Procyclic form 5 and 3 end-enriched transcriptomes	Coverage	unique	not strand specific
	5-triphosphate-			Procyclic form 5			

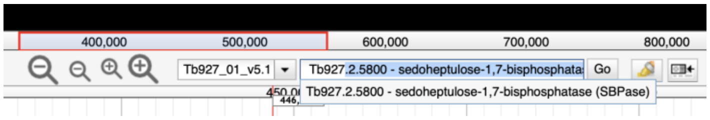
3. Navigating to a specific gene in JBrowse.

The goal of this step is to navigate to the sedoheptulose-1,7-bisphosphatase (SBPase) gene of *T. brucei* 927.

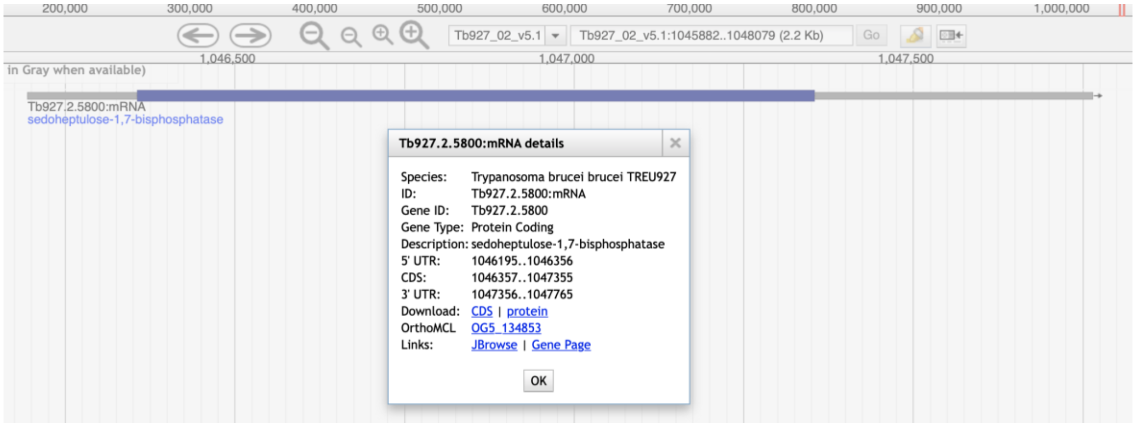

Genome
Track
View
Help
Leishmania major strain Friedlin
Leishmania major strain LV39c5
Leishmania major strain SD 75.1
Leishmania mexicana MHOM/GT/2001/U1103
Leishmania panamensis MHOM/COL/81/L13
Leishmania panamensis strain MHOM/PA/94/PSC-1
Leishmania sp. MAR LEM2494
Leishmania tarentolae Parrot-TarII
Leishmania tropica L590
Leishmania turanica strain LEM423
Leptomonas pyrrhocoris H10
Leptomonas seymouri ATCC 30220
Paratrypanosoma confusum CUL13
Trypanosoma brucei Lister strain 427
Trypanosoma brucei Lister strain 427 2018
Trypanosoma brucei brucei TREU927
Trypanosoma brucei gambiense DAL972
Trypanosoma congolense IL3000
Trypanosoma cruzi Bug2148
Trypanosoma cruzi CL Brener Esmeraldo-like
Trypanosoma cruzi CL Brener Non-Esmeraldo-like
Trypanosoma cruzi Dm28c 2014
Trypanosoma cruzi Dm28c 2017

a. Make sure the *Trypanosoma brucei brucei* TREU927 genome is selected from the genome menu.

b. Start typing the word sedoheptulose in the search box. After a few seconds you should see the result of the search (do not hit enter). Select the gene from the search dropdown. This will take you to Tb927.2.5800.


400,000
500,000
600,000
700,000
800,000
Tb927_01_v5.1
Tb927.2.5800 - sedoheptulose-1,7-bisphospha
Go
Tb927.2.5800 - sedoheptulose-1,7-bisphosphatase (SBPase)

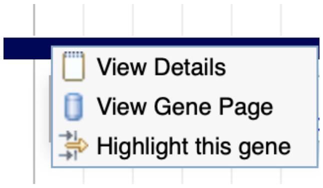
c. You can get information about any feature in the genome view window by clicking on it. Click on the gene feature. What information is available in the popup?



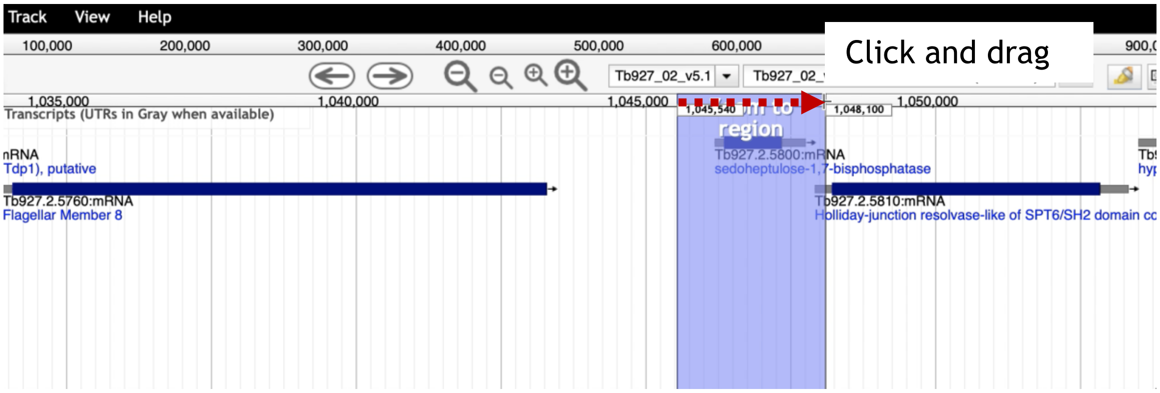
Tb927.2.5800:mRNA details

Species: Trypanosoma brucei brucei TREU927
ID: Tb927.2.5800:mRNA
Gene ID: Tb927.2.5800
Gene Type: Protein Coding
Description: sedoheptulose-1,7-bisphosphatase
5' UTR: 1046195..1046356
CDS: 1046357..1047355
3' UTR: 1047356..1047765
Download: CDS | protein
OrthoMCL OG5_134853
Links: JBrowse | Gene Page

d. You can also right click (or control click) on a feature to display the context menu which provides quick links to highlight a feature, go to the feature page (like the gene page) or get the info popup (the same one you get when you click on the feature).

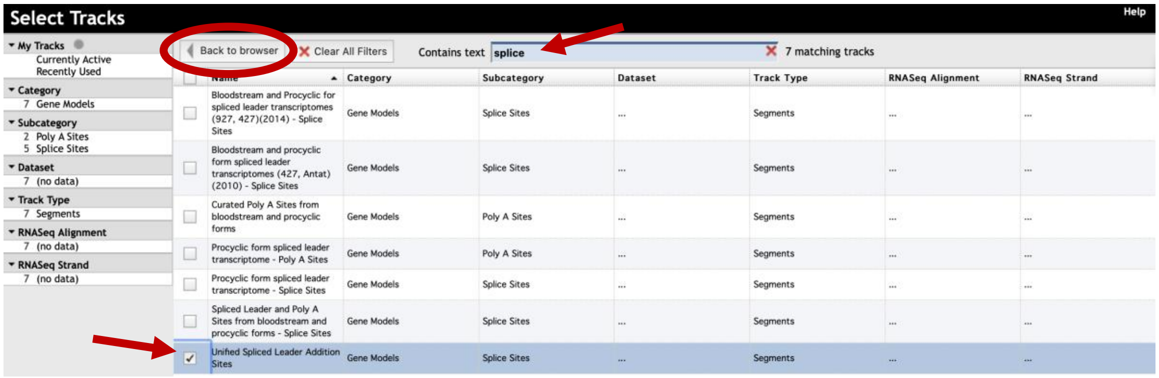


e. What genes are immediately upstream and downstream of SBP? (Hint: use the zoom out button in the navigation bar). What is the difference between the small and large zoom buttons? (*Tip1:* another way to zoom in and out is by clicking on shift and the up or down arrows. What happens if you click shift and left or right arrows? *Tip2:* you can also zoom in by clicking and dragging your cursor in the location ruler in the navigation bar).



4. Exploring transcription start sites.

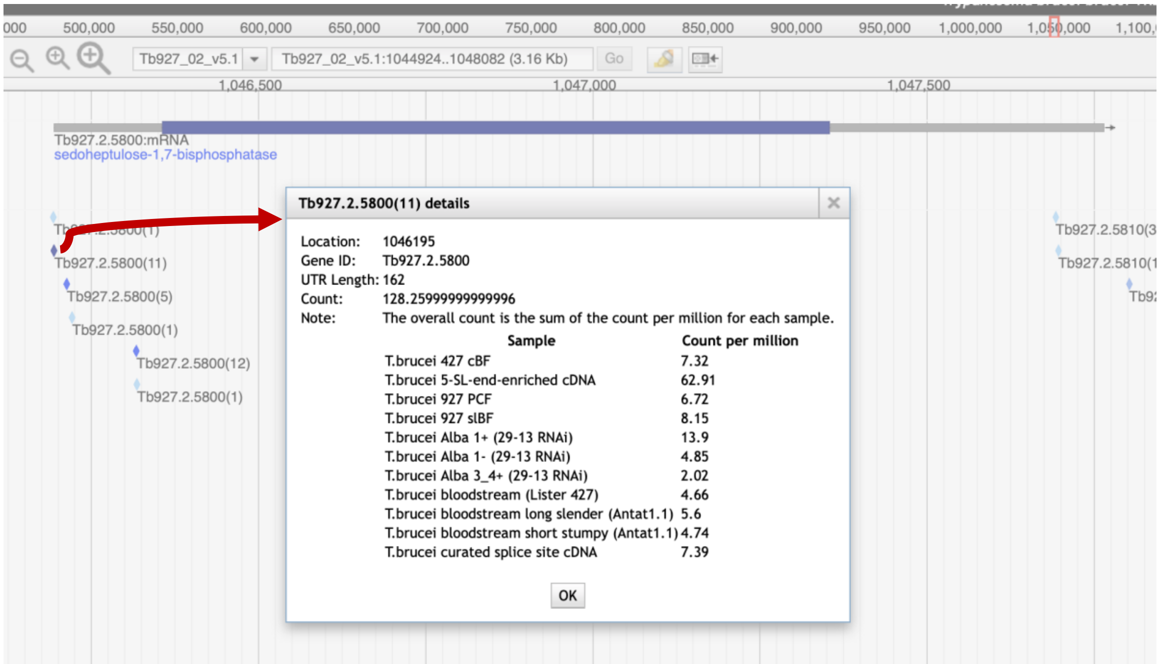
Are you confident about the gene transcription start? (Note: gene features are in blue (left to right) or red (right to left) with untranslated regions (UTRs) in grey).



Contains text splice

7 matching tracks

	Name	Category	Subcategory	Dataset	Track Type	RNASeq Alignment	RNASeq Strand
☐	Bloodstream and Procyclic for spliced leader transcriptomes (927, 427)(2014) - Splice Sites	Gene Models	Splice Sites	...	Segments	...	...
☐	Bloodstream and procyclic form spliced leader transcriptomes (427, Antat) (2010) - Splice Sites	Gene Models	Splice Sites	...	Segments	...	...
☐	Curated Poly A Sites from bloodstream and procyclic forms	Gene Models	Poly A Sites	...	Segments	...	...
☐	Procyclic form spliced leader transcriptome - Poly A Sites	Gene Models	Poly A Sites	...	Segments	...	...
☐	Procyclic form spliced leader transcriptome - Splice Sites	Gene Models	Splice Sites	...	Segments	...	...
☐	Spliced Leader and Poly A Sites from bloodstream and procyclic forms - Splice Sites	Gene Models	Splice Sites	...	Segments	...	...
☑	Unified Spliced Leader Addition Sites	Gene Models	Splice Sites	...	Segments	...	...



Tb927.2.5800(11) details

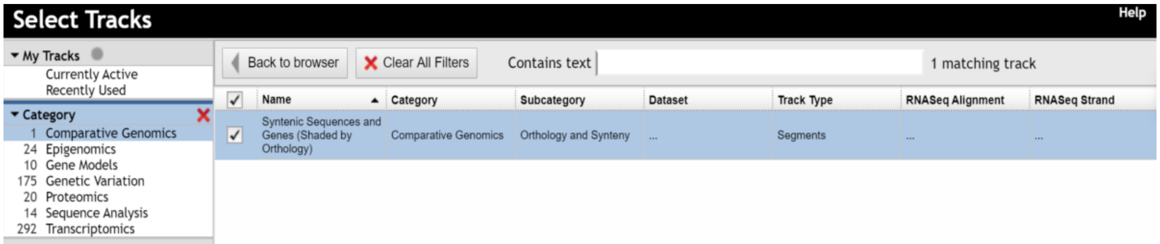
Location: 1046195
Gene ID: Tb927.2.5800
UTR Length: 162
Count: 128.25999999999996
Note: The overall count is the sum of the count per million for each sample.

Sample	Count per million
T.brucei 427 cBF	7.32
T.brucei 5-SL-end-enriched cDNA	62.91
T.brucei 927 PCF	6.72
T.brucei 927 slBF	8.15
T.brucei Alba 1+ (29-13 RNAi)	13.9
T.brucei Alba 1- (29-13 RNAi)	4.85
T.brucei Alba 3_4+ (29-13 RNAi)	2.02
T.brucei bloodstream (Lister 427)	4.66
T.brucei bloodstream long slender (Antat1.1)	5.6
T.brucei bloodstream short stumpy (Antat1.1)	4.74
T.brucei curated splice site cDNA	7.39

What additional data track would be useful for you to assess this? (hint: Click on the “Select Tracks” button to reveal all available tracks. Now type the word “splice” in the “contains text” box. This will filter all tracks that contain the word splice. Find the one called “Unified Splice Leader Addition Sites” and select it. Click on the “Back to browser” button). What do the different diamond colors mean? Click on them and see if you can figure this out from the popups? Which color provides the most evidence for a splice junction?

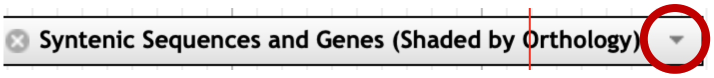
5. **Exploring synteny between genomes.**
   Synteny helps define conservation of homologous genes and gene order between genomes.

- Go to the “Select Tracks” tab on the left of the page and turn on the track called “Syntenic Sequences and Genes”. How did you find this track? One option is to click on the “Comparative Genomics” category on the left side to filter the tracks.

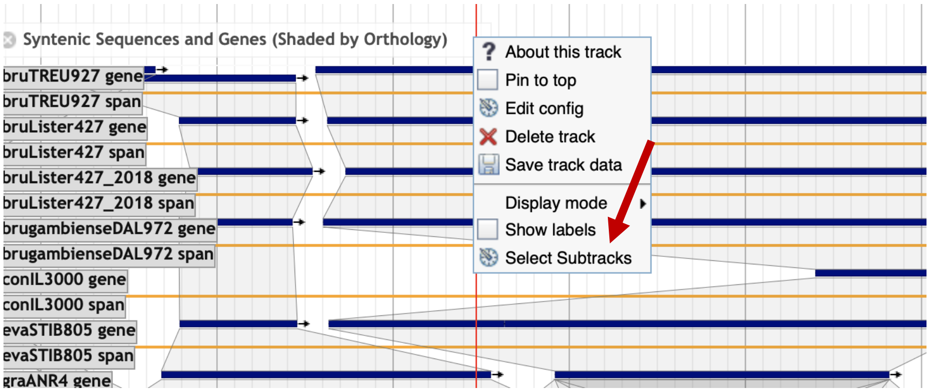


✓	Name	Category	Subcategory	Dataset	Track Type	RNASeq Alignment	RNASeq Strand
✓	Syntenic Sequences and Genes (Shaded by Orthology)	Comparative Genomics	Orthology and Synteny	...	Segments	...	...

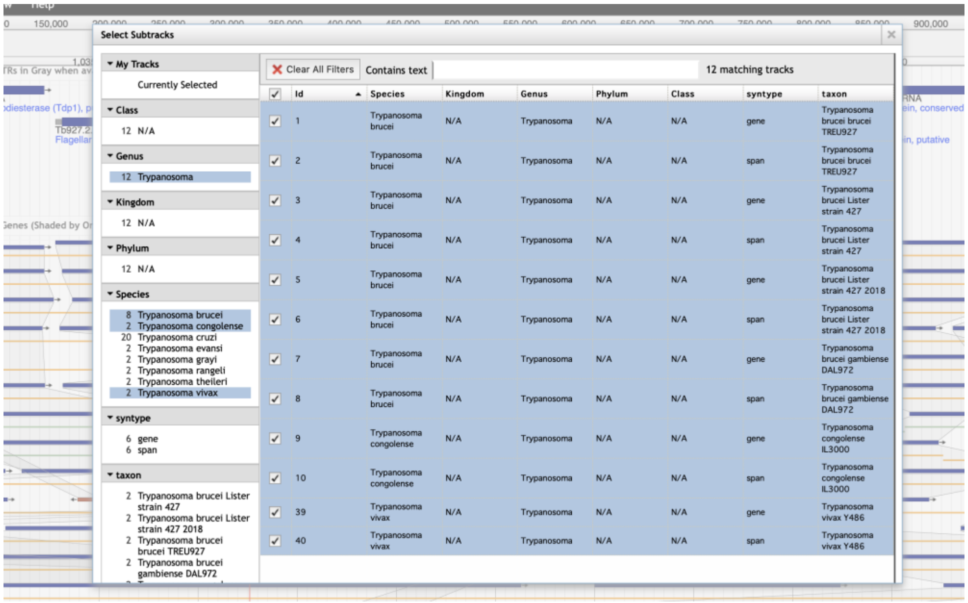
- Return to the browser by clicking “Back to Browser” and zoom out so you can see a couple of genes on either side of SBP (does not have to be exact)
- Configure the synteny track to include the following species subtracks: *Trypanosoma brucei 927*, *T. brucei 427*, *T. brucei gambiense*, *T. congolense*, *T. evansi*, *T. grayi*, *T. theileri* and *T. vivax*.
  - To configure the subtracks:
    - Click on the down arrow in the track name



    - Select the option called “Select Subtracks” from the menu



- In the next popup first uncheck all organisms, second use the filters on the left to select Trypanosoma, third select the species of interest (note that you should select both the gene and span subtracks for each species), fourth click on the save button at the bottom of the popup.



Select Subtracks

▾ My Tracks

Currently Selected

▾ Class

12 N/A

▾ Genus

12 Trypanosoma

▾ Kingdom

12 N/A

▾ Phylum

12 N/A

▾ Species

8	Trypanosoma brucei
2	Trypanosoma congolense
20	Trypanosoma cruzi
2	Trypanosoma evansi
2	Trypanosoma grayi
2	Trypanosoma rangeli
2	Trypanosoma theileri
2	Trypanosoma vivax

▾ syntype

6	gene
6	span

▾ taxon

2	Trypanosoma brucei Lister strain 427
2	Trypanosoma brucei Lister strain 427 2018
2	Trypanosoma brucei brucei TREU927
2	Trypanosoma brucei gambiense DAL972

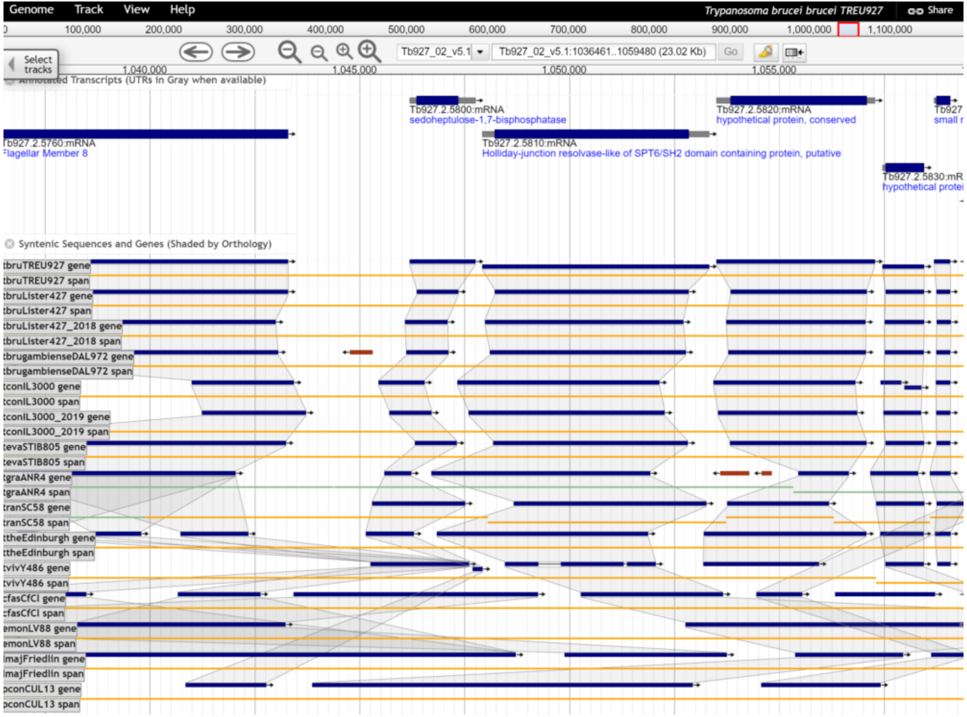
Contains text

12 matching tracks

☑	Id	Species	Kingdom	Genus	Phylum	Class	syntype	taxon
☑	1	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma brucei brucei TREU927
☑	2	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma brucei brucei TREU927
☑	3	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma brucei Lister strain 427
☑	4	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma brucei Lister strain 427
☑	5	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma brucei Lister strain 427 2018
☑	6	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma brucei Lister strain 427 2018
☑	7	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma brucei gambiense DAL972
☑	8	Trypanosoma brucei	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma brucei gambiense DAL972
☑	9	Trypanosoma congolense	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma congolense IL3000
☑	10	Trypanosoma congolense	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma congolense IL3000
☑	39	Trypanosoma vivax	N/A	Trypanosoma	N/A	N/A	gene	Trypanosoma vivax Y486
☑	40	Trypanosoma vivax	N/A	Trypanosoma	N/A	N/A	span	Trypanosoma vivax Y486

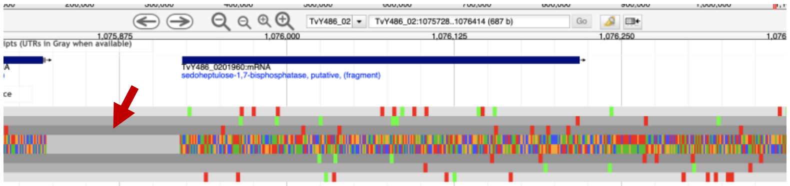
- What does the synteny track in this region look like? Feel free to zoom out some more. Are genes (in general) similarly organized between these species? What does the shading between genes mean?
- What direction is the SBPase gene relative to the chromosome?
- What genes are upstream and downstream of the SBPase? Are these genes syntenic?
- What does synteny look like if you add more distantly related species? Does SBPase appear to have orthologs in *Leishmania*? *Endotrypanum*? *Crithidia*?

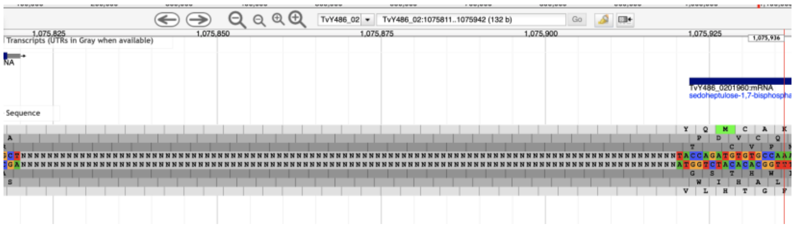
- Examine the gene corresponding to the *T. vivax* SBPase in the synteny track.



Hover over the gene image to find the gene name in the popup. Does this gene appear to be a fragment? What could be some possible reasons for this?

- Do you think all the genomes in the database are fully sequenced? Is it possible that gaps in sequence exist in the available genomes? Let's find out if there is a gap next to the SBPase gene in T. vivax:
  - Select T. vivax from the list of genomes in the menu bar.
  - Turn on the **annotated transcripts** and the **Reference sequence** tracks.
  - Search for the SBPase gene by typing "sedoheptulose" in the search box then select the gene.
  - Zoom to about 600bps. Do you see something missing on the left side of the gene?
  - Zoom in to this area (click and drag). What do you see? What do all of these Ns mean?

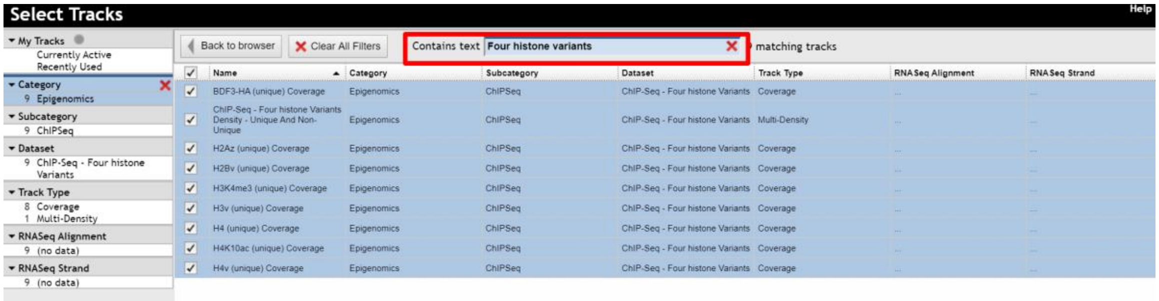

TvY486_02
TvY486_02:1075728..1076414 (687 b)
Go
TvY486_0201960:mRNA
sedoheptulose-1,7-bisphosphatase, putative, (fragment)


TvY486_02
TvY486_02:1075811..1075942 (132 b)
Go
Transcripts (UTRs in Gray when available)
Sequence
TvY486_0201960:mRNA
sedoheptulose-1,7-bisphospha
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

6. Exploring other data tracks in JBrowse.

For this example, we will view *T. brucei* data, so the data tracks you turn on will display data only if the data is aligned to the *T. brucei* genome. Return to the SBPase gene in *T. brucei* by searching for the gene ID in the (Tb927.2.5800) in 'Landmark or Region' to redirect the browser. Then zoom to the area between 0.7M and the end of the chromosome.

Turn on the ChIP-seq coverage plots and turn off the syntenic gene and region tracks. The data tracks are from an experiment called: **ChIP-Seq - Four histone Variants ChIP-Seq Coverage aligned to T brucei TREU927 (Cross) (linear plot)**. For this experiment, chromatin was immunoprecipitated using several different histone antibodies. The DNA that precipitated with the histone was sequenced and aligned to the *T. brucei* TREU927 genome. Peaks in the sequence coverage plots represent areas of histone binding. Different histone variants can be associated with start and termination sites for transcription ([http://www.ncbi.nlm.nih.gov/pubmed/19369410)](http://www.ncbi.nlm.nih.gov/pubmed/19369410)



Contains text: Four histone variants
matching tracks

▾ My Tracks
Currently Active
Recently Used

▾ Category
9 Epigenomics

▾ Subcategory
9 ChIPSeq

▾ Dataset
9 ChIP-Seq - Four histone Variants

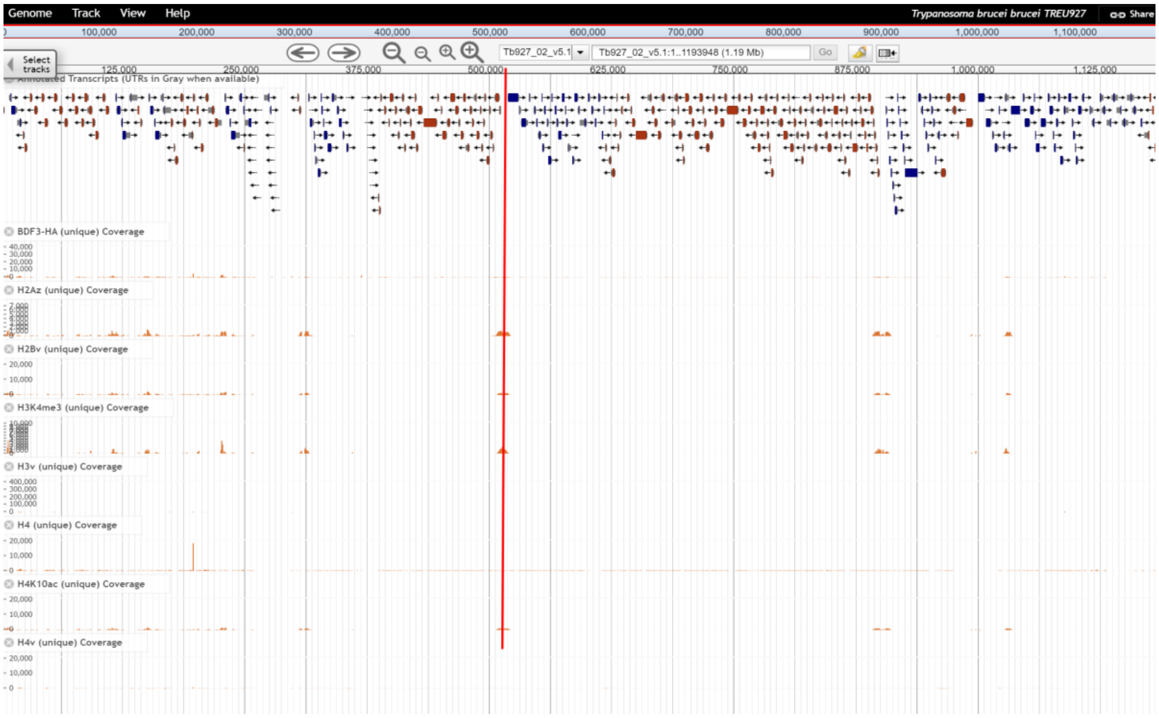
▾ Track Type
8 Coverage
1 Multi-Density

▾ RNASeq Alignment
9 (no data)

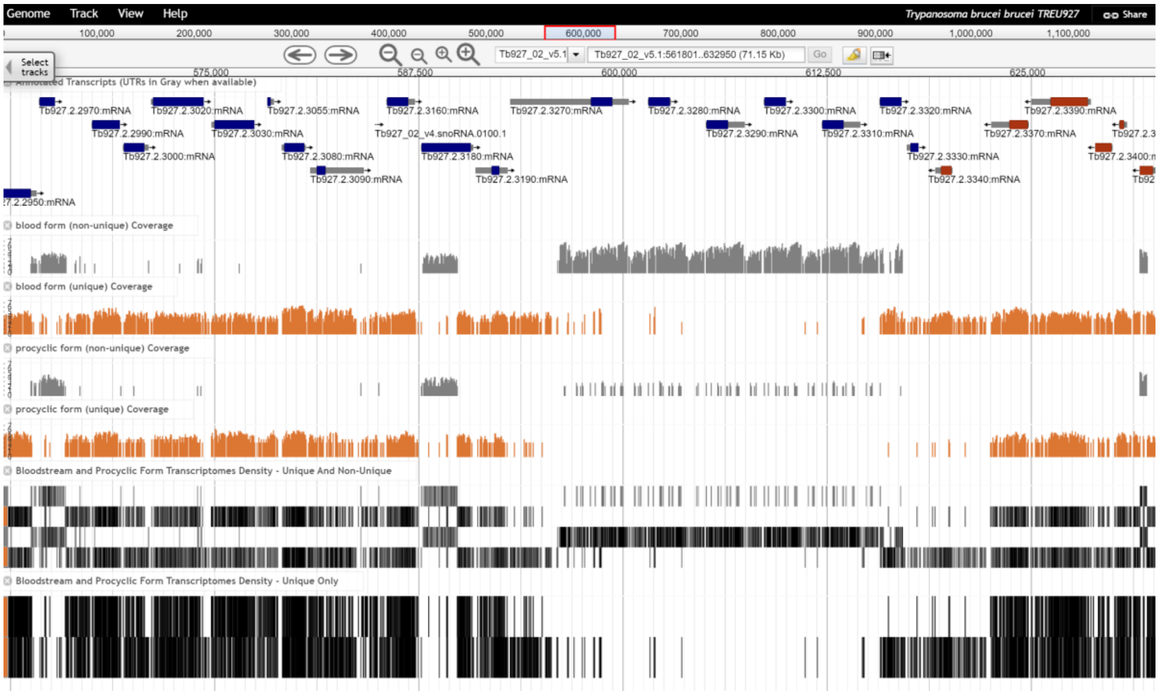
▾ RNASeq Strand
9 (no data)

✓	Name	Category	Subcategory	Dataset	Track Type	RNASeq Alignment	RNASeq Strand
✓	BDF3-HA (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	ChIP-Seq - Four histone Variants Density - Unique And Non-Unique	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Multi-Density	...	...
✓	H2Az (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H2Bv (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H3K4me3 (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H3v (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H4 (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H4K10ac (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...
✓	H4v (unique) Coverage	Epigenomics	ChIPSeq	ChIP-Seq - Four histone Variants	Coverage	...	...

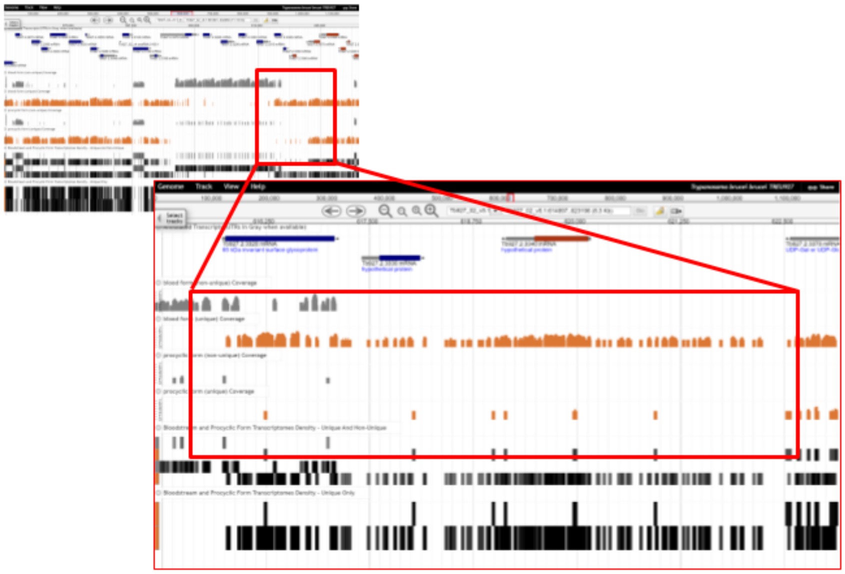
- You may need to adjust the y-axis scaling to bring the tracks into proper view (try setting the score range to "global" by mousing over the track name, clicking the dropdown arrow and selecting "Change Score Range").
- What does this data show you?
- Roughly how many polycistronic units does this chromosome have? Zoom out to the entire chromosome.



- Do the ChIP-seq peaks correlate with the direction of gene transcription (blue vs. red)?
- Now zoom back to around 50Kb. Turn off the ChIP-Seq tracks and turn on the RNASeq Coverage track called: **Bloodstream and Procyclic Form Transcriptomes mRNAseq Coverage aligned to T brucei TREU927**.



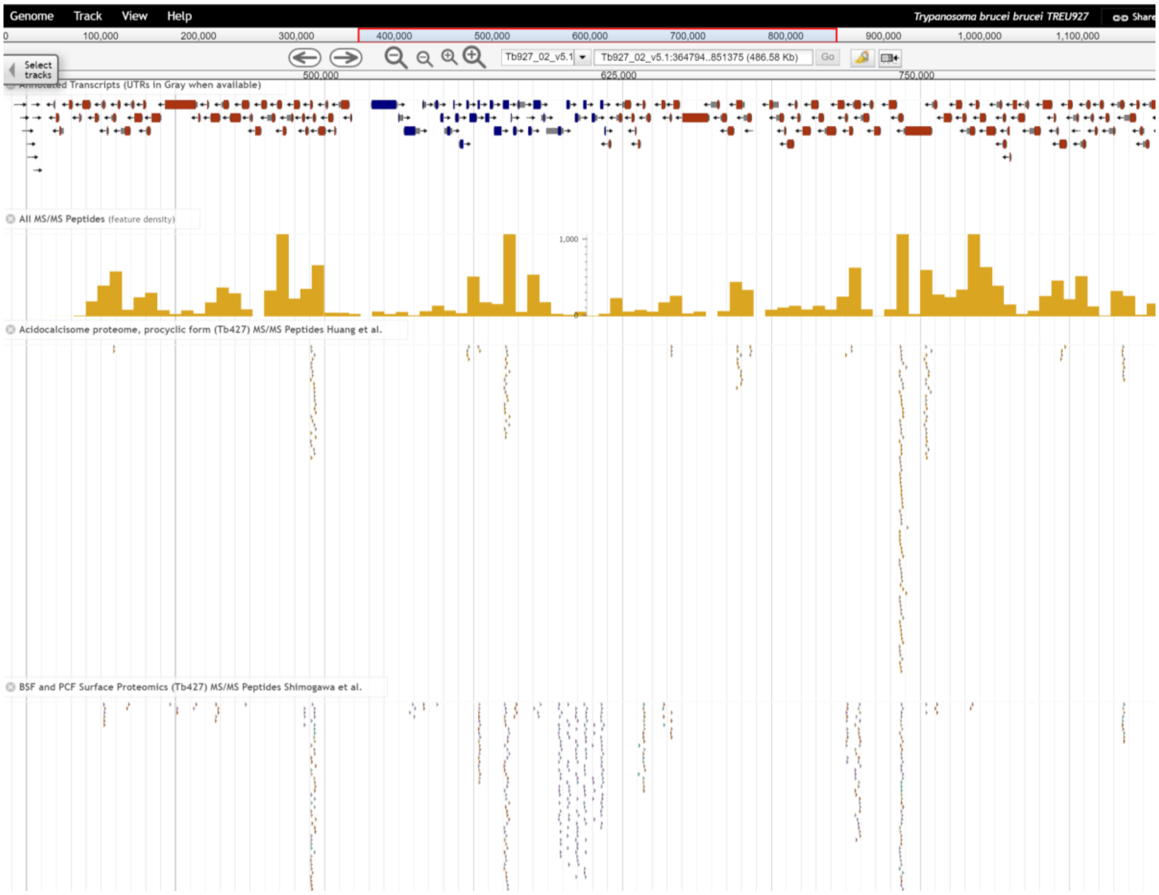
- Move to the **region around 0.6Mbs of the chromosome** (you should be on chromosome 2) and turn on all four subtracks. Take note of the orange and grey bars in the coverage plots. What do you think the grey bars indicate?
- Now zoom out to 100Kb – do you see a difference between the blood and procyclic forms?



- Zoom in to a gene that looks like it is differentially expressed. What are your conclusions? Are the reads supported by unique or non-unique reads?

- Can you turn on additional tracks that may give some more support to your conclusions?
  Hint: turn on the EST and *T. brucei* protein expression evidence tracks.
  - Is there any proteomics evidence for this region?
  - How about EST evidence? Click on an EST graphic (glyph) to get additional information.

- Turn off the RNA-seq graphs and make sure the *T. brucei* protein expression evidence tracks are on. **Zoom out to 500Kb**. Explore the evidence for gene expression based on mapped peptides from proteomics experiments – which gene in this view has the highest number of peptide hits? Try looking at the “All MS/MS peptides (feature density)” track for an overview.


Genome
Track
View
Help
Trypanosoma brucei brucei TREU927
Share
0
100,000
200,000
300,000
400,000
500,000
600,000
700,000
800,000
900,000
1,000,000
1,100,000
Select tracks
Tb927_02_v5.1
Tb927_02_v5.1:364794..851375 (486.58 Kb)
Go
500,000
625,000
750,000
Annotated Transcripts (UTRs in Gray when available)
All MS/MS Peptides (feature density)
1,000
Acidocalcisome proteome, procyclic form (Tb427) MS/MS Peptides Huang et al.
BSF and PCF Surface Proteomics (Tb427) MS/MS Peptides Shimogawa et al.

7. Retrieving data from and uploading your own tracks to JBrowse

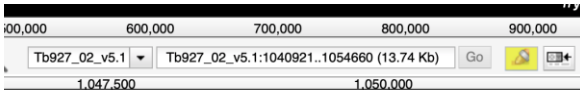
a. Downloading sequence in FASTA format from a region of interest:

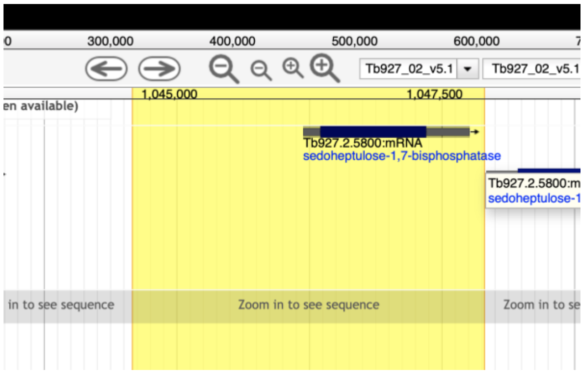
i. Make sure the “annotated transcripts” and the “reference sequence” tracks are turned on.

ii. Click on the “highlight a region” button in the navigation bar. It should turn yellow when activated.

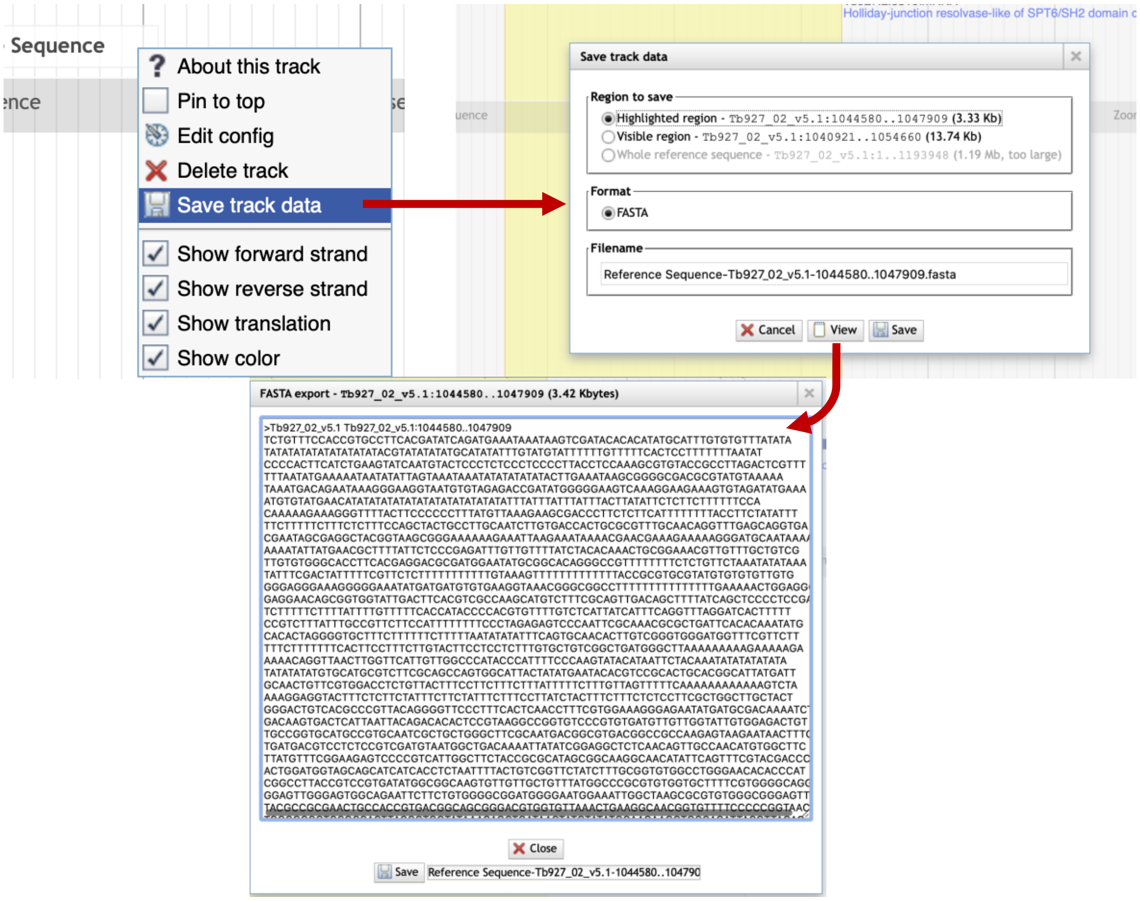
iii. Click and drag in the genome view region and select the area you would like to highlight.

iv. Click on the down arrow on the reference sequence track and select “Save track data”.





v. In the next popup window you can keep everything as the default and either save or view the sequence.



Save track data

Region to save
● Highlighted region - Tb927_02_v5.1:1044580..1047909 (3.33 Kb)
○ Visible region - Tb927_02_v5.1:1040921..1054660 (13.74 Kb)
○ Whole reference sequence - Tb927_02_v5.1:1..1193948 (1.19 Mb, too large)

Format
● FASTA

Filename
Reference Sequence-Tb927_02_v5.1-1044580..1047909.fasta

Cancel View Save

FASTA export - Tb927_02_v5.1:1044580..1047909 (3.42 Kbytes)

```
>Tb927_02_v5.1 Tb927_02_v5.1:1044580..1047909
TCTGTTTCCACCGTGCCTTCACGATATCAGATGAAATAAATAAGTCGATACACACATATGCATTTGTGTGTTTATATA
TATATATATATATATATACGTATATATATGCATATATTTGTATGTATTTTTTGTTTTTCACTCCTTTTTTTAATAT
CCCCACTTCATCTGAAGTATCAATGTACTCCCTCTCCCTCCCCTTACCTCCAAAGCGTGTACCGCCTTAGACTCGTTT
TTTAATATGAAAAATAATATATTAGTAAATAAATATATATATATACTTGAAATAAGCGGGGCGACGCGTATGTAAAAA
TAAATGACAGAATAAAGGGAAGGTAATGTGTAGAGACCGATATGGGGGAAGTCAAAGGAAGAAAGTGTAGATATGAAA
ATGTGTATGAACATATATATATATATATATATATATATATTTATTTATTTATTTACTTATATTCTCTTCTTTTTTCCA
CAAAAAGAAAGGGTTTTACTTCCCCCCTTTATGTTAAAGAAGCGACCCTTCTCTTCATTTTTTTTACCTTCTATATTT
TTCTTTTTCTTTCTCTTTCCAGCTACTGCCTTGCAATCTTGTGACCACTGCGCGTTTGCAACAGGTTTGAGCAGGTGA
CGAATAGCGAGGCTACGGTAAGCGGGAAAAAAGAAATTAAGAAATAAAACGAACGAAAGAAAAAGGGATGCAATAAAA
AAAATATTATGAACGCTTTTATTCTCCCGAGATTTGTTGTTTTATCTACACAAACTGCGGAAACGTTGTTTGCTGTCG
TTGTGTGGGCACCTTCACGAGGACGCGATGGAATATGCGGCACAGGGCCGTTTTTTTTCTCTGTTCTAAATATATAAA
TATTTCGACTATTTTTCGTTCTCTTTTTTTTTTTGTAAAGTTTTTTTTTTTTTACCGCGTGCGTATGTGTGTGTTGTG
GGGAGGGAAAGGGGGAAATATGATGATGTGTGAAGGTAAACGGGCGGCCTTTTTTTTTTTTTTTGAAAAACTGGAGG
GAGGAACAGCGGTGGTATTGACTTCACGTCGCCAAGCATGTCTTTCGCAGTTGACAGCTTTTATCAGCTCCCCTCCGA
TCTTTTTCTTTTATTTTGTTTTTCACCATACCCCACGTGTTTTGTCTCATTATCATTTCAGGTTTAGGATCACTTTTT
CCGTCTTTATTTGCCGTTCTTCCATTTTTTTTCCCTAGAGAGTCCCAATTCGCAAACGCGCTGATTCACACAAATATG
CACACTAGGGGTGCTTTCTTTTTTCTTTTTAATATATATTTCAGTGCAACACTTGTCGGGTGGGATGGTTTCGTTCTT
TTTCTTTTTTTCACTTCCTTTCTTGTACTTCCTCCTCTTTGTGCTGTCGGCTGATGGGCTTAAAAAAAAAGAAAAAGA
AAAACAGGTTAACTTGGTTCATTGTTGGCCCATACCCATTTTCCCAAGTATACATAATTCTACAAATATATATATATA
TATATATATGTGCATGCGTCTTCGCAGCCAGTGGCATTACTATATGAATACACGTCCGCACTGCACGGCATTATGATT
GCAACTGTTCGTGGACCTCTGTTACTTTCCTTCTTTCTTTATTTTTCTTTGTTAGTTTTTCAAAAAAAAAAAAGTCTA
AAAGGAGGTACTTTCTCTTCTATTTCTTCTATTTCTTTCCTTATCTACTTTCTTTCTCTCCTTCGCTGGCTTGCTACT
GGGACTGTCACGCCCGTTACAGGGGTTCCCTTTCACTCAACCTTTCGTGGAAAGGGAGAATATGATGCGACAAAATCT
GACAAGTGACTCATTAATTACAGACACACTCCGTAAGGCCGGTGTCCCGTGTGATGTTGTTGGTATTGTGGAGACTGT
TGCCGGTGCATGCCGTGCAATCGCTGCTGGGCTTCGCAATGACGGCGTGACGGCCGCCAAGAGTAAGAATAACTTTC
TGATGACGTCCTCTCCGTCGATGTAATGGCTGACAAAATTATATCGGAGGCTCTCAACAGTTGCCAACATGTGGCTTC
TTATGTTTCGGAAGAGTCCCCGTCATTGGCTTCTACCGCGCATAGCGGCAAGGCAACATATTCAGTTTCGTACGACCC
ACTGGATGGTAGCAGCATCATCACCTCTAATTTTACTGTCGGTTCTATCTTTGCGGTGTGGCCTGGGAACACACCCAT
CGGCCTTACCGTCCGTGATATGGCGGCAAGTGTTGTTGCTGTTTATGGCCCGCGTGTGGTGCTTTTCGTGGGGCAGG
GGAGTTGGGAGTGGCAGAATTCTTCTGTGGGGCGGATGGGGAATGGAAATTGGCTAAGCGCGTGTGGGCGGGAGTT
TACGCCGCGAACTGCCACCGTGACGGCAGCGGGACGTGGTGTTAAACTGAAGGCAACGGTGTTTTCCCCCGGTAAC
```

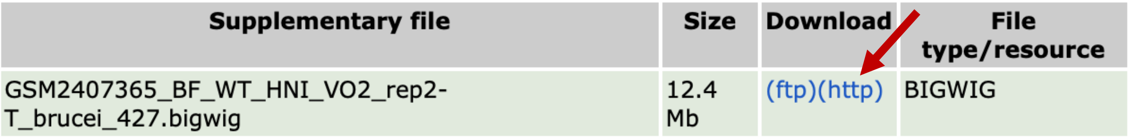
Close
Save Reference Sequence-Tb927_02_v5.1-1044580..104790

b. Uploading data to JBrowse:

JBrowse can accept several standard-format data files by direct upload or through a URL if the data is stored remotely. Some file formats like BAM and VCF require indexing before uploading. In this exercise we will download a bigwig file from GEO and then upload it to JBrowse:

i. Go to this GEO sample record: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2407365

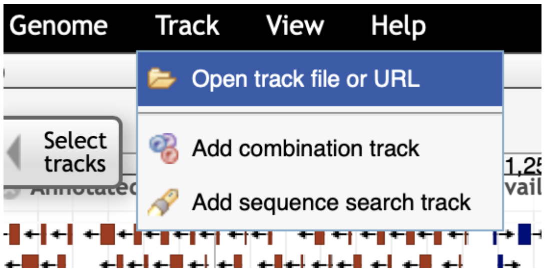
ii. Scroll down to the bottom of the page and download the bigwig file with the http link.



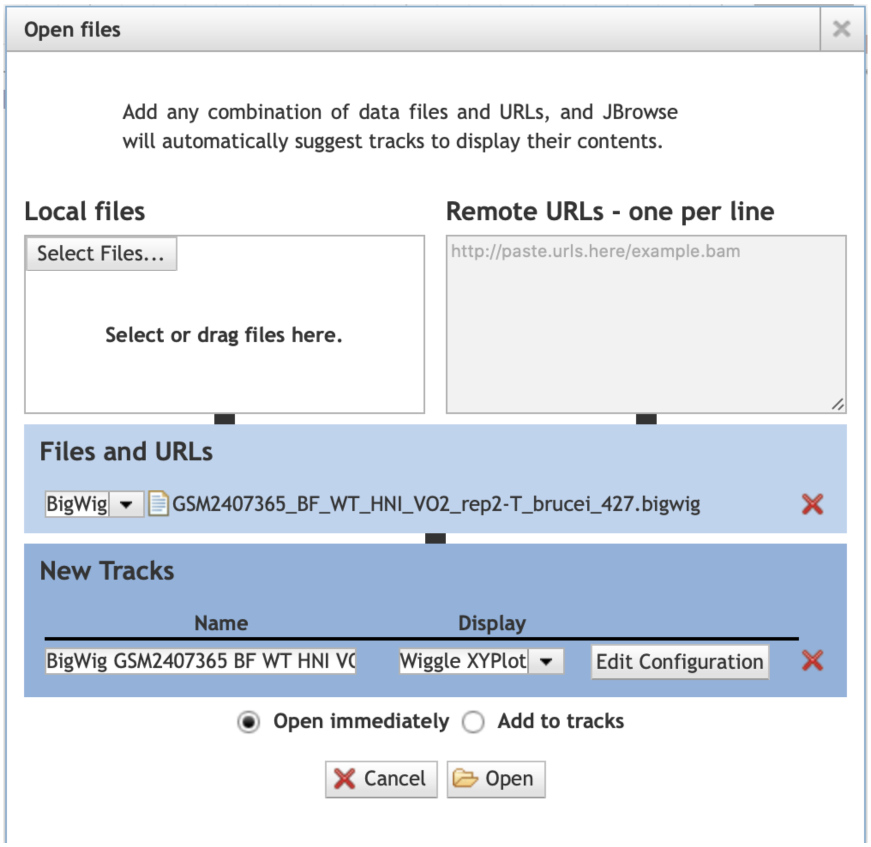
Supplementary file	Size	Download	File type/resource
GSM2407365_BF_WT_HNI_VO2_rep2-T_brucei_427.bigwig	12.4 Mb	(ftp)(http)	BIGWIG

iii. Once the file is downloaded go to JBrowse and select *Trypanosoma brucei brucei* Lister 427 as the reference genome (hint: use the Genome link in the menu panel, top left).

iv. Turn on the track for annotated transcripts if it is not on already.



v. Click on the Tracks menu item and select “Open track file or URL”.
vi. In the popup click on select file then select the file you just downloaded. JBrowse should automatically recognize that the file is in bigwig format.
vii. Click on the Open button.The bigWig output should appear very quickly in your browser.



Open files ✕

Add any combination of data files and URLs, and JBrowse will automatically suggest tracks to display their contents.

Local files

Select Files…

Select or drag files here.

Remote URLs - one per line

http://paste.urls.here/example.bam

Files and URLs

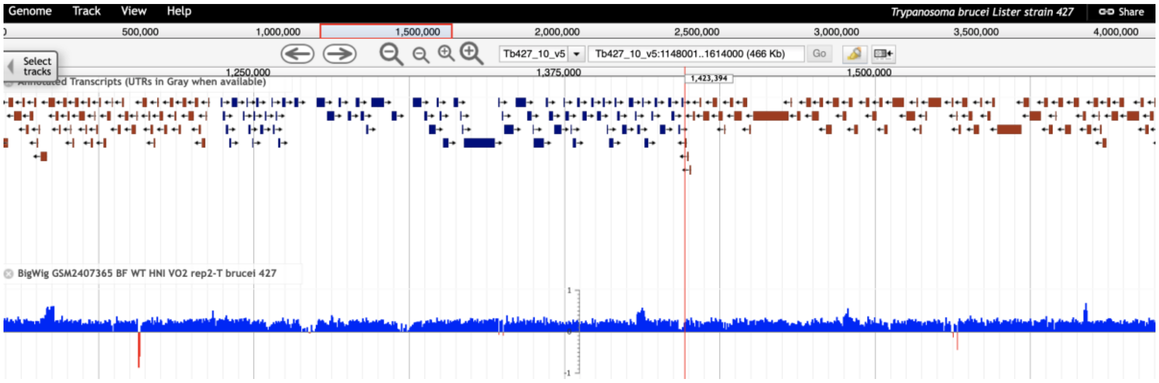
BigWig ▾ GSM2407365_BF_WT_HNI_VO2_rep2-T_brucei_427.bigwig ✕

New Tracks

Name	Display		
BigWig GSM2407365 BF WT HNI V(	Wiggle XYPlot ▾	Edit Configuration	✕

● Open immediately ○ Add to tracks

✕ Cancel Open


Genome
Track
View
Help
Trypanosoma brucei Lister strain 427
Share
500,000
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
3,500,000
4,000,000
Tb427_10_v5
Tb427_10_v5:1148001..1614000 (466 Kb)
Go
Select tracks
1,250,000
1,375,000
1,423,394
1,500,000
Annotated Transcripts (UTRs in Gray when available)
BigWig GSM2407365 BF WT HNI VO2 rep2-T brucei 427
1
0
-1

Strategies Tutorial

Note: This exercise uses PlasmoDB.org as an example, but the same functionality is available on a VEuPathDB resources.

Learning objectives:

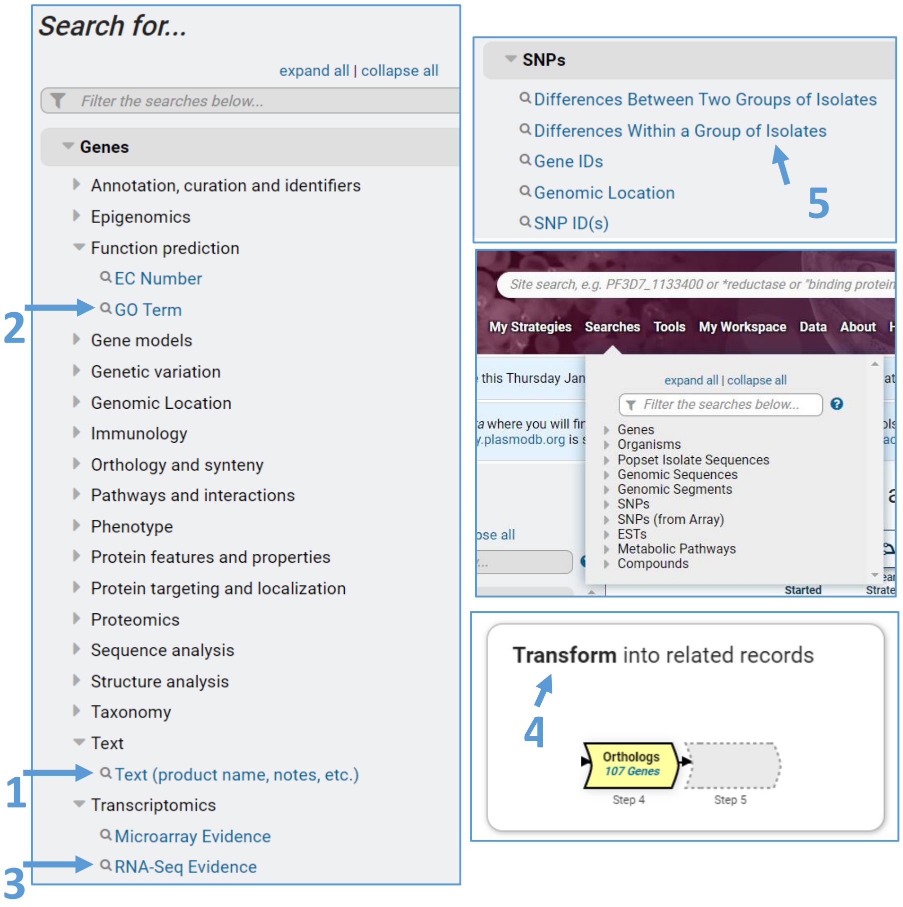
- Build a multistep strategy
- Use the Text, GO Term, RNA-Seq, and SNP searches
- Combine search results using Boolean operators and the colocation tool
- Transform genes of one organism into their orthologs in another organism
- Infer expression timing from a well-studied organism onto another organism that lacks data.

In this tutorial you will find genes expressed in gametocytes that are likely proteases and have variation in their upstream regions, possibly affecting promoter regions and other gene features. The strategy you build will combine three different searches that query *P. falciparum* data, then transform the *P. falciparum* genes returned by those searches into their *P. vivax* orthologs and look for SNPs in the upstream regions of the *P. vivax* genes. The ortholog transform enables you make inferences about genes in *P. vivax*, an organism with limited functional data, based on existing data in the closely related and well-studied *P. falciparum*. The *P. vivax* genes returned by the search are likely to share two biological properties, proteolytic activity and expression in gametocytes, and show variation in their upstream regions, possibly affecting promoter regions or other gene features.

Strategies Overview:

The strategy system offers over 100 structured searches that can be combined to produce multi-step strategies. Each search queries a specific data set and **returns a list of IDs** that share the biological characteristic defined by the data.

Searches are accessible from the ‘Search For…’ menu on the home page and from the ‘Searches’ dropdown menu in the header of every page. Searches listed under Genes will return a list of gene IDs, while searches listed under ‘SNPs’ or ‘Metabolic Pathways’ will return record IDs representing SNPs, or metabolic pathways.



The 5 searches you will use in this tutorial are:

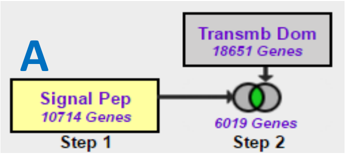
1. Identify Genes by Text (product name, notes, etc.) – The search compares your term against the text in the fields that you specify, returning the IDs of gene records that have a match.
2. Identify Genes by GO Term – Returns genes that have your specified Gene Ontology (GO) Term(s) or ID(s) assigned to them.
3. Identify Genes based on RNA Seq Evidence – PlasmoDB integrates raw RNA sequencing data from many different experiments and analyzes all data according to the same workflow to produce expression values. This search returns genes based on their transcript expression as measure by RNA sequencing.
4. Transform by Orthology – PlasmoDB integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at 150 genomes that span the tree of life. The transform we perform here will convert a list of genes in one organism to their orthologs in a different organism. In this case, we will transform a list of *P. falciparum* genes into their *P. vivax* orthologs.
5. Identify SNPs based on Differences within a Group of Isolates – PlasmoDB integrates whole genome resequencing of isolates and analyzes each isolate for single nucleotide polymorphisms compared to a reference genome. This search returns SNPs that are shared between all the *P. vivax* isolates that are integrated in PlasmoDB.

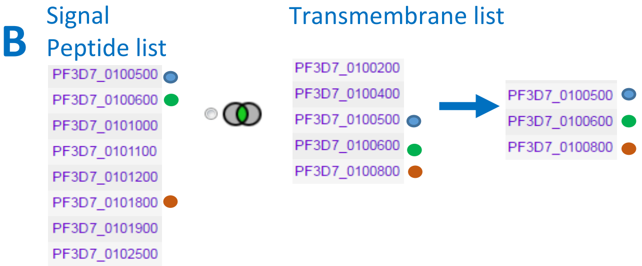
Before we get started… a few words about combining search results:

Each search returns a list of IDs. When two searches are combined, the two result sets (list of IDs) are merged. The table shows the 5 options for combining search results.

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 (collocated) in the genome

If the searches return the same type of genomic feature they can be combined using any of the 5 operators (i.e. search 1 returns genes, search 2 returns genes as in screenshot group A below).





Signal Peptide list

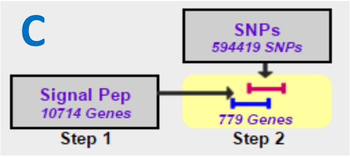
PF3D7_0100500 🔵
PF3D7_0100600 🟢
PF3D7_0101000
PF3D7_0101100
PF3D7_0101200
PF3D7_0101800 🟠
PF3D7_0101900
PF3D7_0102500

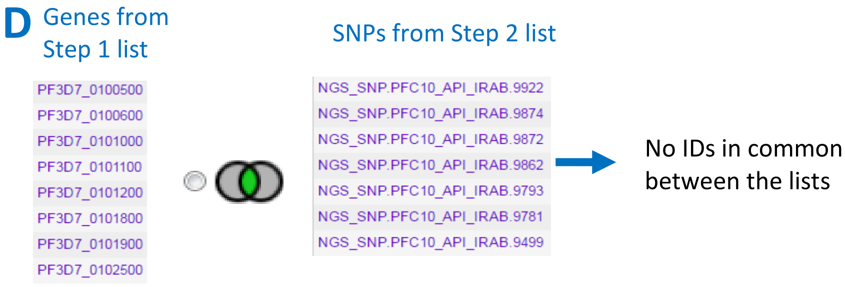
Transmembrane list

PF3D7_0100200
PF3D7_0100400
PF3D7_0100500 🔵
PF3D7_0100600 🟢
PF3D7_0100800 🟠

PF3D7_0100500 🔵
PF3D7_0100600 🟢
PF3D7_0100800 🟠

However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. This is illustrated in screenshot groupings C and D below. Because genes and SNPs are different genomic features, there are no IDs in the list of genes (Step 1) that are present in the list of SNPs (Step 2). To combine a search that returns genes with a search that returns SNPs, you must use the collocation option (1 relative to 2). We know the genomic location of each gene and each SNP and the colocation option is designed to return features based on their relative genomic location, i.e. SNPs that are near or within genes.





Genes from Step 1 list

PF3D7_0100500
PF3D7_0100600
PF3D7_0101000
PF3D7_0101100
PF3D7_0101200
PF3D7_0101800
PF3D7_0101900
PF3D7_0102500

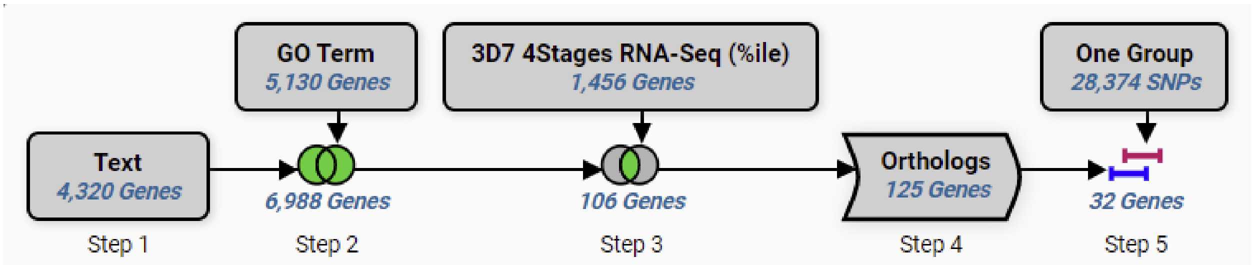
SNPs from Step 2 list

NGS_SNP.PFC10_API_IRAB.9922
NGS_SNP.PFC10_API_IRAB.9874
NGS_SNP.PFC10_API_IRAB.9872
NGS_SNP.PFC10_API_IRAB.9862
NGS_SNP.PFC10_API_IRAB.9793
NGS_SNP.PFC10_API_IRAB.9781
NGS_SNP.PFC10_API_IRAB.9499

No IDs in common between the lists

Build the Strategy:

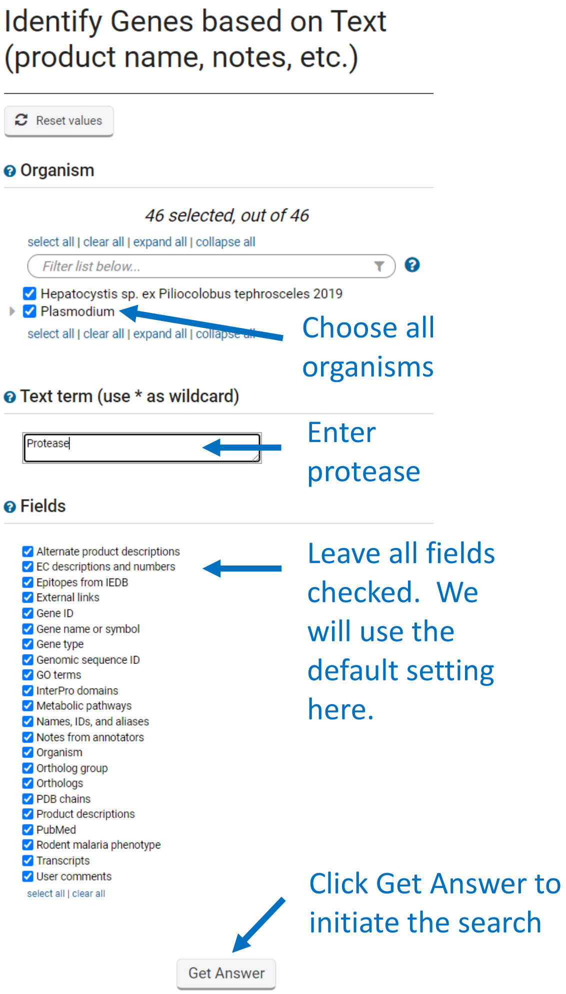
Find *P. vivax* genes that are possible proteases, likely expressed during the gametocyte stages and contain SNPs in their upstream regions. This search strategy employs 4 searches, an ortholog transform and the colocation tool to integrate SNP information. Steps 1 and 2 return *P. falciparum* proteases using two different lines of evidence – a text search in step 1 and a Gene Ontology (GO) term search in step 2. These searches are combined with a union to obtain a more comprehensive list of possible proteases. Step 3 returns genes with evidence for expression during the gametocyte stages based on RNA sequencing data collected in *P. falciparum*. Steps 2 and 3 are combined using the intersect operator to produce a list of genes that have BOTH biological properties: these genes are likely proteases with evidence for expression during gametocyte stages. In the next step, the *P. falciparum* genes returned in the step 3 result are transformed into their *P. vivax* orthologs. This results in a set of 125 *P. vivax* genes with suspected protease activity and expression in gametocytes based on annotation and experimental evidence from *P. falciparum*, an organism for which more complete annotation and functional genomics data is available. In Step 5 we look for single nucleotide polymorphisms (SNPs) among isolates of *P. vivax* and collocate these SNPs to the upstream regions of the *P. vivax* genes. The final result is a set of 32 *P. vivax* genes that are likely proteases expressed in the gametocyte stage and that have SNPs in their upstream regions. Your strategy should look like this when you are done:



Step by Step Instructions

1. **Run a text search using protease as the text term.**
   Identify Genes by Text (product name, notes, etc.): Using the Text Search, find genes whose records contain the term 'protease'. To reach the text search, click on the link in the home page 'Search For…' menu. The page opens showing a list of parameters that are needed to query the data. Every search is loaded with default parameters so that you can click Get Answer and run the search. Change the Text term to 'protease' and click Get Answer to initiate the search. The search results are displayed in the My Strategies section which consists of a strategy panel followed by a filter table and a result table.

   Navigation: >PlasmoDB >Search for Genes >Text > Text (product name, notes, etc.)



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

Reset values

Organism

46 selected, out of 46

select all | clear all | expand all | collapse all

Filter list below...

☑ Hepatocystis sp. ex Piliocolobus tephrosceles 2019

☑ Plasmodium

select all | clear all | expand all | collapse all

Choose all organisms

Text term (use * as wildcard)

Protease

Enter protease

Fields

- ☑ Alternate product descriptions
- ☑ EC descriptions and numbers
- ☑ Epitopes from IEDB
- ☑ External links
- ☑ Gene ID
- ☑ Gene name or symbol
- ☑ Gene type
- ☑ Genomic sequence ID
- ☑ GO terms
- ☑ InterPro domains
- ☑ Metabolic pathways
- ☑ Names, IDs, and aliases
- ☑ Notes from annotators
- ☑ Organism
- ☑ Ortholog group
- ☑ Orthologs
- ☑ PDB chains
- ☑ Product descriptions
- ☑ PubMed
- ☑ Rodent malaria phenotype
- ☑ Transcripts
- ☑ User comments

select all | clear all

Leave all fields checked. We will use the default setting here.

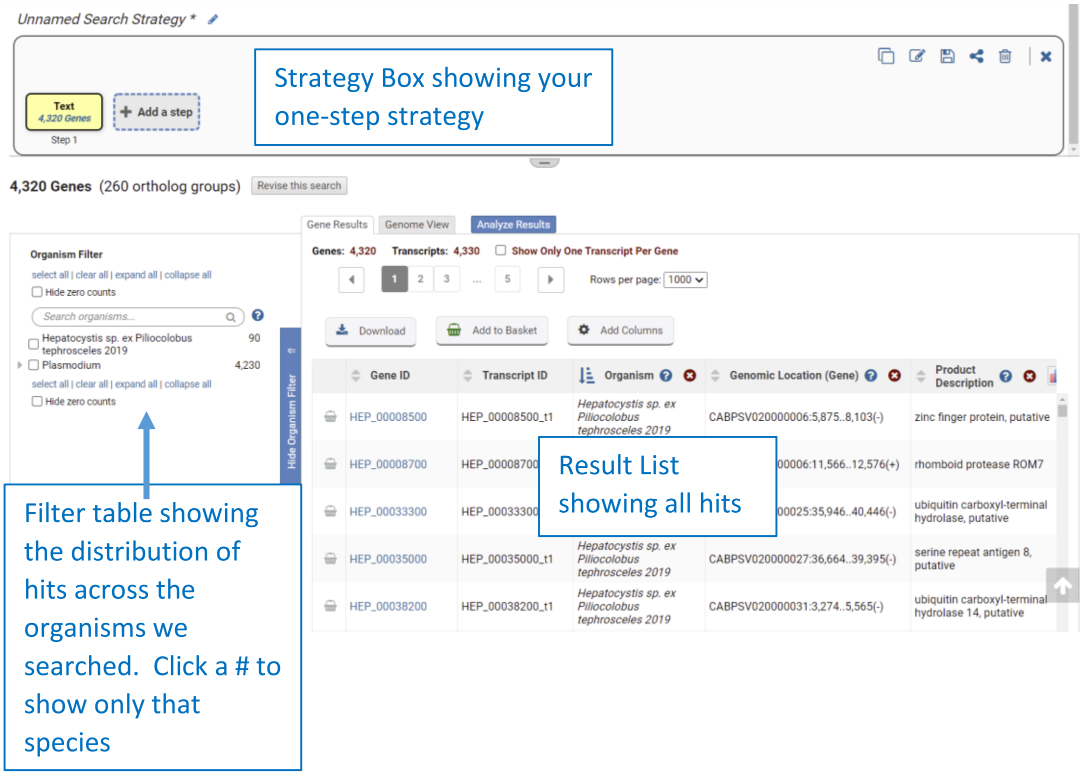
Click Get Answer to initiate the search

Get Answer

Parameters:

Organism	:	Default - all
Text term (use * as wildcard)	:	protease
Fields	:	Default - all

Results and strategy: You created a one-step strategy by running the text search. The strategy returns 4320 genes that are annotated with the word ‘protease’. This annotation could appear in any field that you searched. You can analyze this result by exploring the hits. Look at the data in the columns of the result table. You can add more data with the Add Columns button. Clicking a gene ID in the first column will take you to that gene’s record page. Please explore your results to see if they make sense. For example, gene product names might contain the word ‘protease’. Functional data assigned to the genes (GO terms and EC numbers) may indicate protease activity.



Unnamed Search Strategy *

Strategy Box showing your one-step strategy

4,320 Genes (260 ortholog groups) Revise this search

Organism Filter

select all | clear all | expand all | collapse all

□ Hide zero counts

□ Hepatocystis sp. ex Piliocolobus tephrosceles 2019	90
□ Plasmodium	4,230

select all | clear all | expand all | collapse all

□ Hide zero counts

Filter table showing the distribution of hits across the organisms we searched. Click a # to show only that species

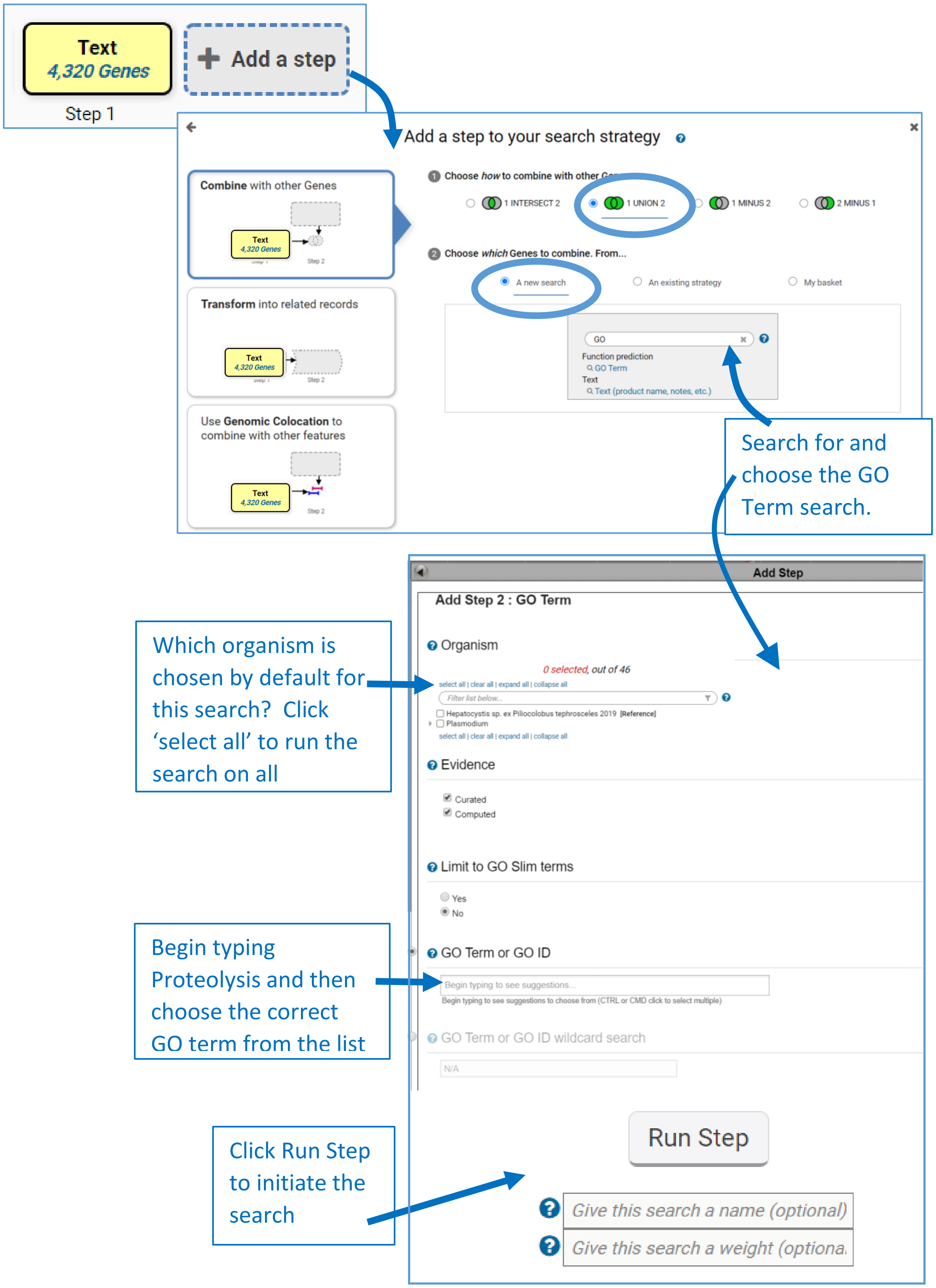
Genes: 4,320 Transcripts: 4,330 □ Show Only One Transcript Per Gene

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description
HEP_00008500	HEP_00008500_t1	*Hepatocystis sp. ex Piliocolobus tephrosceles 2019*	CABPSV020000006:5,875..8,103(-)	zinc finger protein, putative
HEP_00008700	HEP_00008700		00006:11,566..12,576(+)	rhomboid protease ROM7
HEP_00033300	HEP_00033300		00025:35,946..40,446(-)	ubiquitin carboxyl-terminal hydrolase, putative
HEP_00035000	HEP_00035000_t1	*Hepatocystis sp. ex Piliocolobus tephrosceles 2019*	CABPSV020000027:36,664..39,395(-)	serine repeat antigen 8, putative
HEP_00038200	HEP_00038200_t1	*Hepatocystis sp. ex Piliocolobus tephrosceles 2019*	CABPSV020000031:3,274..5,565(-)	ubiquitin carboxyl-terminal hydrolase 14, putative

Result List showing all hits

Add a step choosing to run a search for genes annotated with the biological process gene ontology term – GO:0006508: proteolysis. Gene Ontology annotations offer a second line of evidence for finding proteases. The ontologies are a controlled vocabulary for describing the molecular function, biological process and subcellular location of a gene product. GO annotations in PlasmoDB were either provided by the sequencing and annotation centers or inferred based on a gene's similarity to protein domains from the InterPro databases. The GO Term search returns a gene if it is annotated with the GO term that you are looking for. Let's use that search to look for genes annotated with GO:0006508: proteolysis. We will union the text search results with our GO term results when we combine the results of the two searches.

Navigation: Add Step >Combine with other Genes >1 union 2 > A new search >GO Term



Text
4,320 Genes
Add a step
Step 1

Add a step to your search strategy

Combine with other Genes
Transform into related records
Use **Genomic Colocation** to combine with other features
Text
4,320 Genes
Step 2

1 Choose *how* to combine with other Ge

○ 1 INTERSECT 2 ● 1 UNION 2 ○ 1 MINUS 2 ○ 2 MINUS 1

2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

GO

Function prediction
GO Term
Text
Text (product name, notes, etc.)

Search for and choose the GO Term search.

Add Step 2 : GO Term

Organism

0 selected, out of 46

select all | clear all | expand all | collapse all

Filter list below...

□ Hepatocystis sp. ex Piliocolobus tephrosceles 2019 **[Reference]**
□ Plasmodium

select all | clear all | expand all | collapse all

Which organism is chosen by default for this search? Click 'select all' to run the search on all

Evidence

☑ Curated
☑ Computed

Limit to GO Slim terms

○ Yes
● No

GO Term or GO ID

Begin typing to see suggestions...

Begin typing to see suggestions to choose from (CTRL or CMD click to select multiple)

GO Term or GO ID wildcard search

N/A

Begin typing Proteolysis and then choose the correct GO term from the list

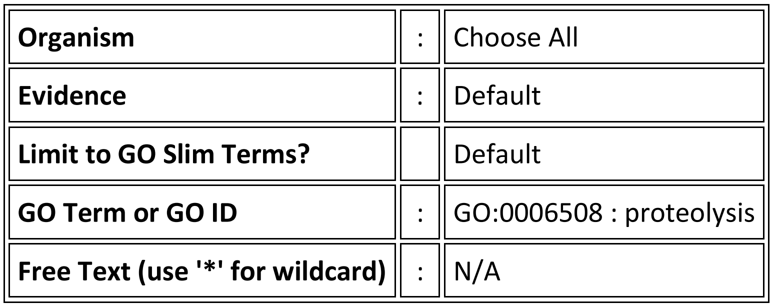
Run Step

Give this search a name (optional)

Give this search a weight (optiona

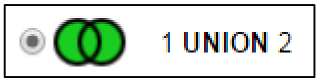
Click Run Step to initiate the search

Parameters:



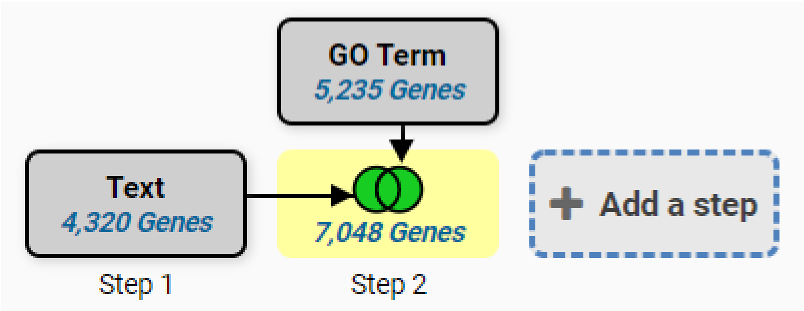
Organism	:	Choose All
Evidence	:	Default
Limit to GO Slim Terms?		Default
GO Term or GO ID	:	GO:0006508 : proteolysis
Free Text (use '*' for wildcard)	:	N/A

Combine:



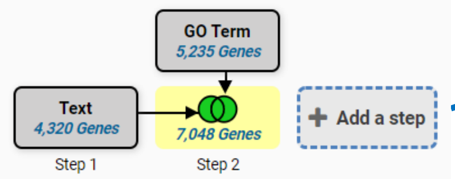
● 1 UNION 2

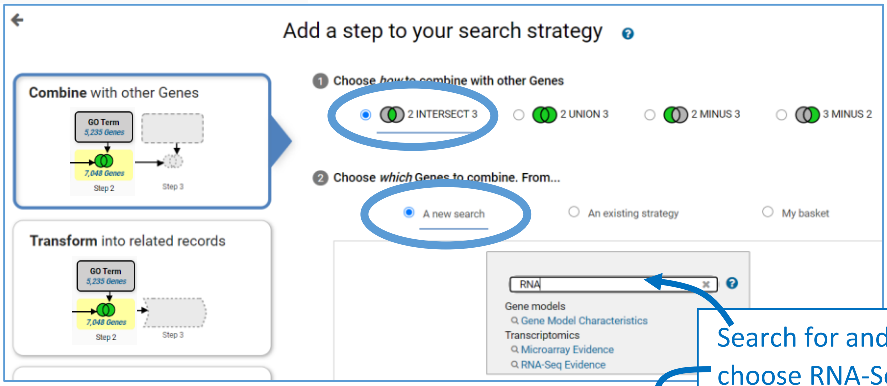
Strategy Result: The GO term search returned 5,235 genes annotated with the proteolysis GO term. The union of the text and GO search returns 7,048 genes that are suspected to have proteolytic activity.



2. **Add a step choosing to run a search for genes based on Transcript Expression using RNA-Seq Evidence.** Since PlasmoDB has integrated several RNA sequencing data sets you must first choose what data set (experiment) to search before you are taken to the search form to choose parameters. Use the Filter Data set tool to choose the Percentile search (P) for 'Strand specific Transcriptomes of 4 life cycle stages (Lopez-Barragan et al)'. This data set contains the RNA sequencing analysis of two gametocyte samples. Running the percentile search using the default parameters will return the genes whose expression levels are in the top 20% for those samples.

   Navigation: Add Step >Combine with other Genes >2 intersect 3 >A new search >RNA Seq Evidence




Add a step to your search strategy

1 Choose *how* to combine with other Genes

● 2 INTERSECT 3 ○ 2 UNION 3 ○ 2 MINUS 3 ○ 3 MINUS 2

2 Choose *which* Genes to combine. From...

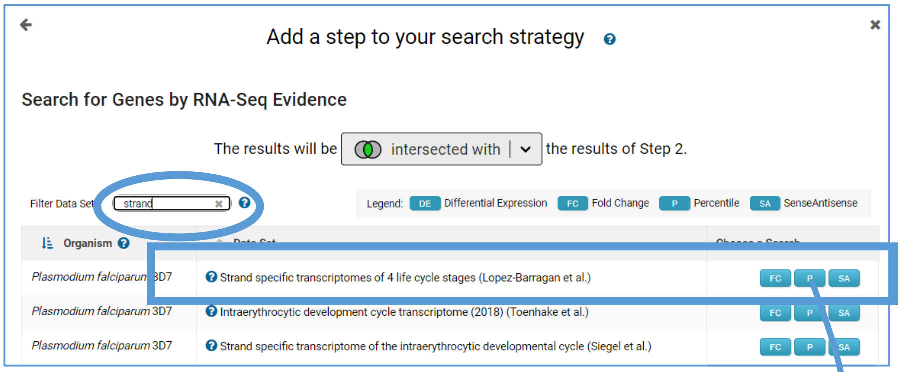
● A new search ○ An existing strategy ○ My basket

RNA

Gene models
- Gene Model Characteristics

Transcriptomics
- Microarray Evidence
- RNA-Seq Evidence

Search for and choose RNA-Seq evidence.


Add a step to your search strategy

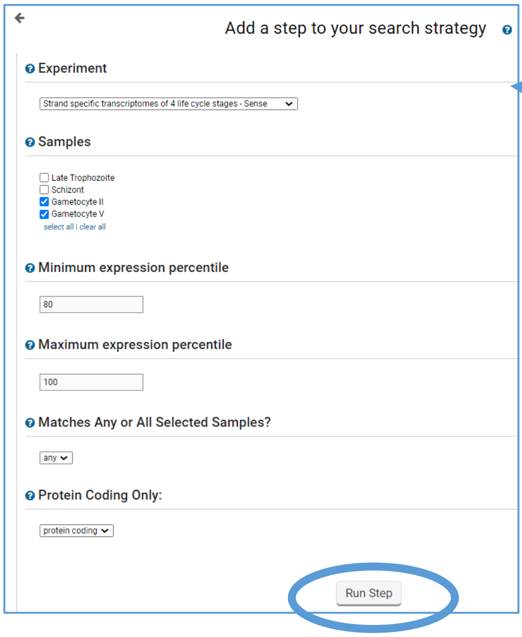
Search for Genes by RNA-Seq Evidence

The results will be intersected with the results of Step 2.

Filter Data Set: strand

Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense

Organism	Data Set	Choose a Search
Plasmodium falciparum 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)	FC P SA
Plasmodium falciparum 3D7	Intraerythrocytic development cycle transcriptome (2018) (Toenhake et al.)	FC P SA
Plasmodium falciparum 3D7	Strand specific transcriptome of the intraerythrocytic developmental cycle (Siegel et al.)	FC P SA


Add a step to your search strategy

Experiment

Strand specific transcriptomes of 4 life cycle stages - Sense

Samples

- ☐ Late Trophozoite
- ☐ Schizont
- ☑ Gametocyte II
- ☑ Gametocyte V

select all | clear all

Minimum expression percentile

80

Maximum expression percentile

100

Matches Any or All Selected Samples?

any

Protein Coding Only:

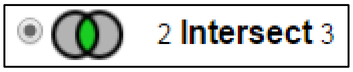
protein coding

Run Step

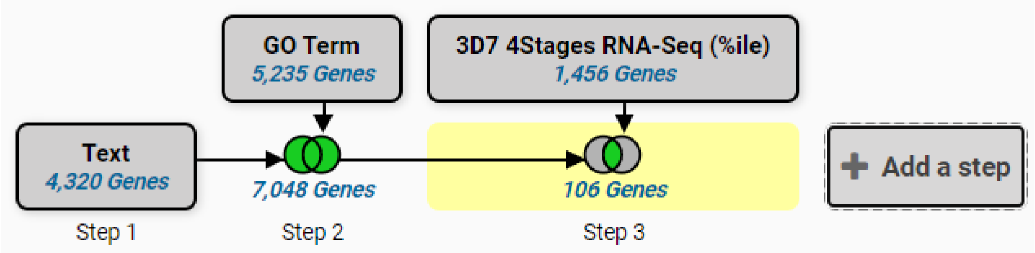
Parameters:

Experiment	:	Strand specific transcriptomes of 4 life cycle stages sense strand
Samples	:	Gametocyte II, Gametocyte V
Minimum expression percentile	:	default
Maximum expression percentile	:	default
Matches Any or All Selected Samples?	:	default
Protein Coding Only:	:	default

Combine: Intersecting this search with the previous result will produce a list of genes that are common to both result lists.

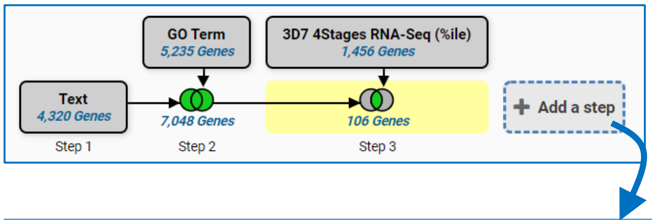


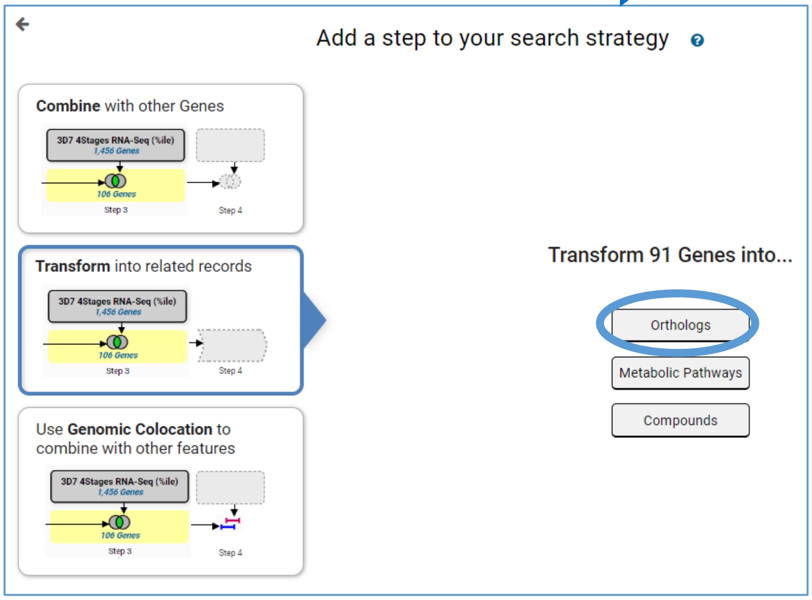
Strategy result: We have a three-step strategy that returns 106 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!



3. **Add a step to the strategy that transforms the 106 *P. falciparum* genes into *P. vivax* genes.**
P. falciparum is a well-studied organism with active curatorial efforts and large amounts of functional data. For example, PlasmoDB has 18 RNA sequencing and 11 microarray data sets integrated for *P. falciparum*, but only 4 RNA-Seq and 2 microarray for *P. vivax*. A researcher interested in *P. vivax* can take advantage of the *P. falciparum* data by creating a strategy based on *P. falciparum* data to retrieve genes with the biological properties they are interested in, and then transforming the results to their *P. vivax* orthologs.

Navigation: >Add Step >Transform into related records >Orthologs



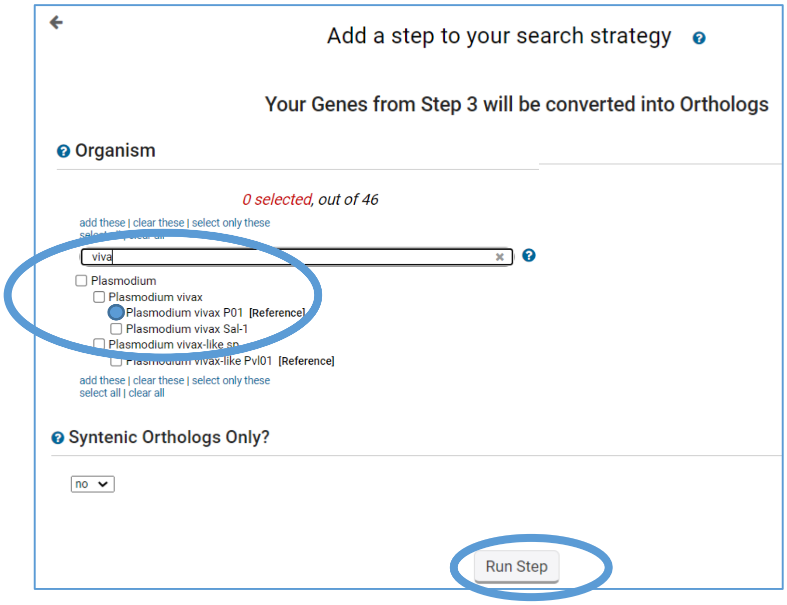


Add a step to your search strategy

Combine with other Genes
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
106 Genes
Step 3
Step 4
Transform into related records
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
106 Genes
Step 3
Step 4
Use **Genomic Colocation** to combine with other features
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
106 Genes
Step 3
Step 4

Transform 91 Genes into...

Orthologs
Metabolic Pathways
Compounds



Add a step to your search strategy

Your Genes from Step 3 will be converted into Orthologs

Organism

0 selected, out of 46

add these | clear these | select only these
select all | clear all

viva

□ Plasmodium
□ Plasmodium vivax
● Plasmodium vivax P01 **[Reference]**
□ Plasmodium vivax Sal-1
□ Plasmodium vivax-like sp.
□ Plasmodium vivax-like Pvl01 **[Reference]**

add these | clear these | select only these
select all | clear all

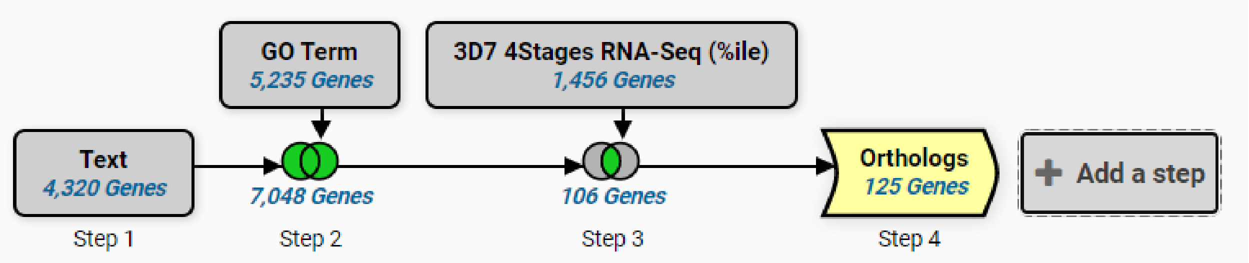
Syntenic Orthologs Only?

no

Parameters: Choose only *P. vivax* P01 in the Organism parameter of the Add Step Popup.

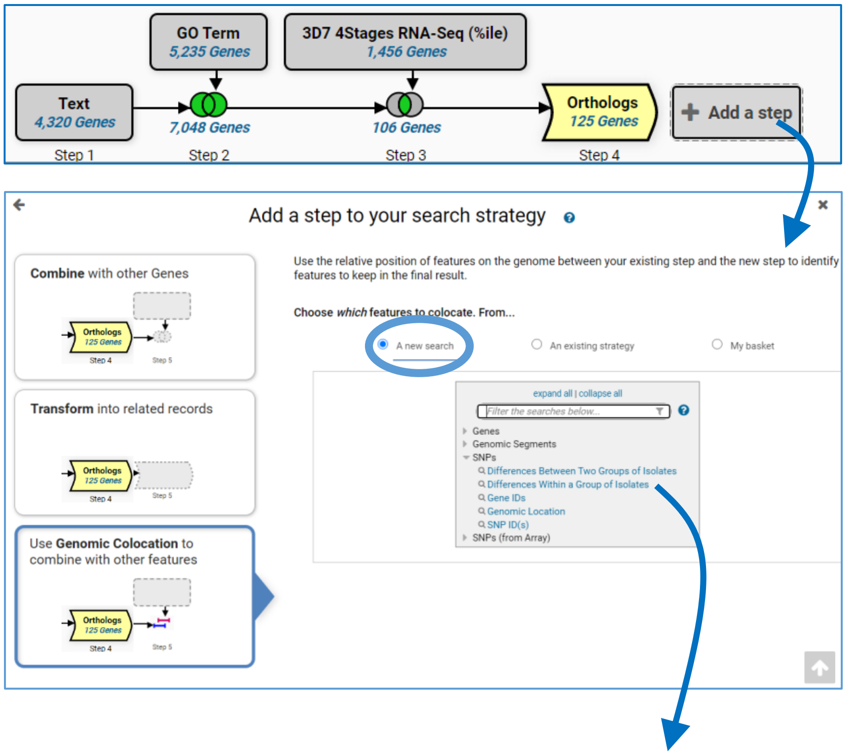
Combine: The ortholog transform function does not combine lists, but instead transforms the results into orthologs from a different species.

Strategy Result: We have a four-step strategy that returns 125 *P. vivax* genes that are suspected proteases expressed in gametocytes based on *P. falciparum* RNA Sequencing data.



4. **Add a step to the strategy that returns *P vivax* SNPs and collocate those SNPs to the upstream 1000bp of the *P. vivax* genes in step 4.** We can look for variation (SNPs) associated with the genes from Step 4. PlasmoDB integrates whole genome resequencing data from many isolates, and PlasmoDB contains 195 data sets from whole-genome sequencing of *P. vivax* isolates. PlasmoDB analyzes the whole genome sequencing reads by aligning them to the reference genome and then examines the genome one base at a time to find bases in the isolate that do not match the reference sequence. The SNPs are loaded in the database along with other information such as how many sequencing reads supported the SNP call and the genomic location of the SNP. The search we will use analyzes whole genome resequencing data from all *P. vivax* isolates to find SNPs shared between all isolates. You will notice that initiating the search does not immediately bring up the result, but instead leads you to the colocation tool.

Navigation: >Add Step >Use Genomic Colocation >A new search >Differences Within a Group of Isolates

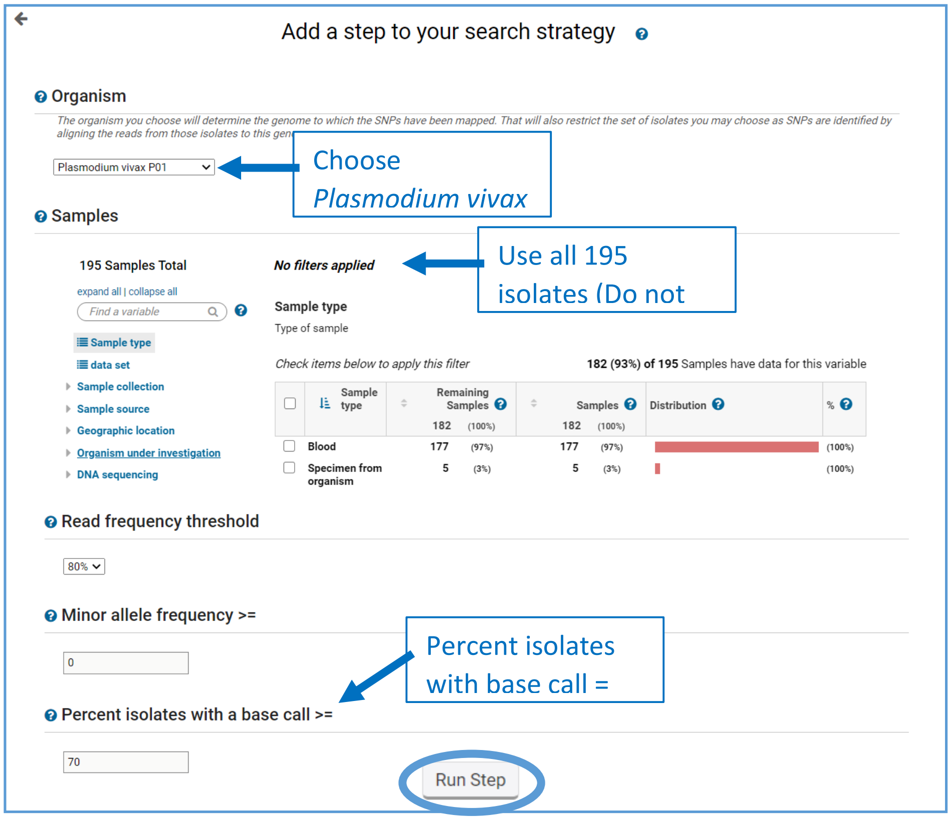


Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.

Choose *which* features to colocate. From...

● A new search ○ An existing strategy ○ My basket

expand all | collapse all
Filter the searches below...
Genes
Genomic Segments
SNPs
Differences Between Two Groups of Isolates
Differences Within a Group of Isolates
Gene IDs
Genomic Location
SNP ID(s)
SNPs (from Array)



Add a step to your search strategy

Organism

The organism you choose will determine the genome to which the SNPs have been mapped. That will also restrict the set of isolates you may choose as SNPs are identified by aligning the reads from those isolates to this gen

Plasmodium vivax P01

Choose *Plasmodium vivax*

Samples

195 Samples Total

No filters applied

Use all 195 isolates (Do not

Sample type

Type of sample

Check items below to apply this filter

182 (93%) of 195 Samples have data for this variable

□	Sample type	Remaining Samples	Samples	Distribution	%
		182 (100%)	182 (100%)		
□	Blood	177 (97%)	177 (97%)		(100%)
□	Specimen from organism	5 (3%)	5 (3%)		(100%)

Read frequency threshold

80%

Minor allele frequency >=

0

Percent isolates with base call =

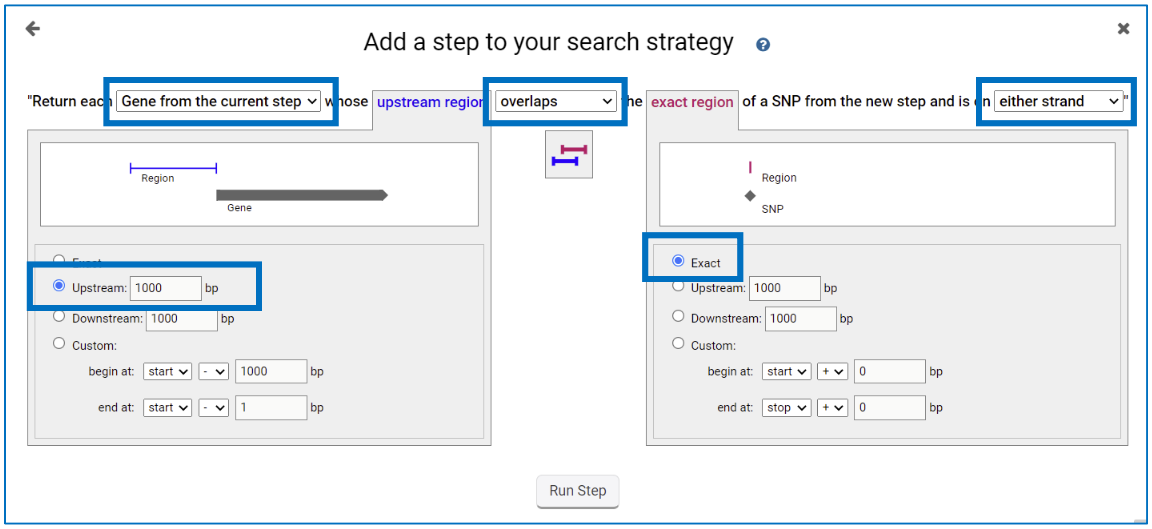
Percent isolates with a base call >=

70

Parameters:

Organism	:	*P. vivax* P01
Isolates	:	Default = All Isolates (195)
Read frequency threshold	:	Default - 80%
Minor allele frequency >=	:	Default - 0
Percent isolates with a base call >=	:	Default - 70

Colocation: Because this search returns SNPs and not genes, the only option for combining the two result lists is by relative genomic location. Arrange the statement in the Colocation popup to read: **Return each Gene from step 4 whose upstream 1000bp region overlaps the exact region of a SNP in Step 5 and is on either strand**. Remember to indicate that you want to locate the SNPs in the upstream region of the gene.



Add a step to your search strategy

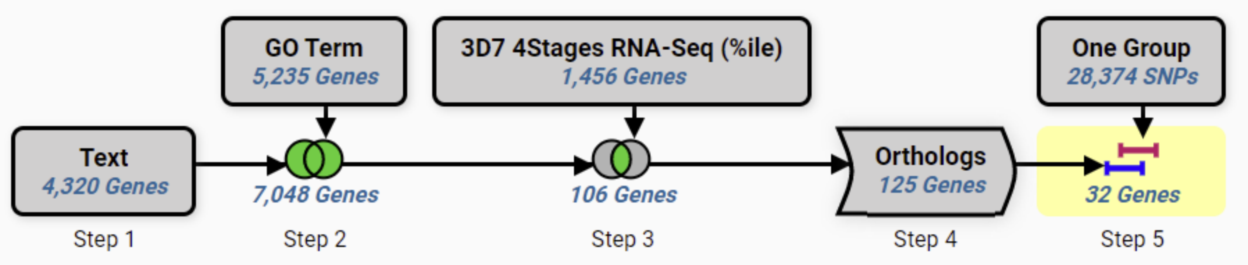
"Return each Gene from the current step whose upstream region overlaps the exact region of a SNP from the new step and is on either strand"

○ Exact
● Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start - 1000 bp
end at: start - 1 bp

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

Strategy: Congratulations! You have completed the strategy and have a list of 32 *P. vivax* genes that are possible proteases, are likely expressed in gametocytes and have upstream SNPs.

This link will retrieve the completed strategy: https://plasmodb.org/plasmo/app/workspace/strategies/import/76a3cff6f01535ea



GO Term
5,235 Genes
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
One Group
28,374 SNPs
Text
4,320 Genes
7,048 Genes
106 Genes
Orthologs
125 Genes
32 Genes
Step 1
Step 2
Step 3
Step 4
Step 5