VEuPathDB User Documentation

Contract Name: VEuPathDB
Contract #: 75N93019C00077
Contractor: University of Pennsylvania

Release 50 - December 2020

Prepared By: Jeremy DeBarry

Reviewed By: Brian Brunk, Jessica Kissinger, Omar Harb, other staff as appropriate

Submitted By: Jeremy DeBarry

Date	Version/release	Description/Change log
December 17, 2020	Release 50	Changes relative to last submission • Added 'About OrthoMCL' section (page 5) • Training exercises added or updated: OrthoMCL tutorial (page 19) • Updated Analyses Methods section for CHIP Seq, Copy Number Variation, Genetic Variation and SNP Calling, Microarray Data, Protein Array Data, and Metabolic Pathways (page 176)

Table of Contents

INTRODUCTION **4**

ABOUT VEUPATHDB.ORG **4**

About OrthoMCL 5

DATA ACCESS POLICY **8**

CITING VEUPATHDB IN PUBLICATIONS AND PRESENTATIONS **8**

COMMUNITY INTERACTIONS AND DATA SUBMISSION POLICIES **9**

HOW TO SUBMIT DATA TO VEUPATHDB **9**

How to submit data for integration in VEuPathDB 10

Genome Sequence and/or Annotation *11*

High Throughput or Next Generation Sequencing *11*

Microarray *11*

Proteomics *12*

Quantitative Proteomics *12*

ChIP-chip *12*

Isolates typed by sequencing limited genetic loci *13*

Isolates or Strains typed by High Throughput Sequencing *13*

General Data Submission *13*

Data submission and release on VEuPathDB databases 13

General principles: *14*

Why submit my data to VEuPathDB? *14*

How do I submit data to VEuPathDB? *14*

What data types are supported by VEuPathDB? *15*

THE DATA PRODUCTION CYCLE **15**

Data Management SOPs (Standard Operating Procedures) for VEuPathDB databases 16

HOW TO USE OUR SITES **18**

Online instructional material 18

Print-based instructional material 19

OrthoMCL Tutorial *19*

Site Search *26*

Search Strategies *32*

Advanced Search Strategies *38*

Public Strategies *48*

Exploring the Gene Page *49*

JBrowse Basics *54*

Interpreting RNA-seq data (Browser Exercise II) *67*

Gene Ontology (GO) Enrichment *69*

Regular Expressions & Genomic Colocation *75*

Variant calling in VEuPathDB galaxy (Part 1) *84*

Genome Annotation with Companion (Part 1) *90*

Genome Annotation with Companion (Part 2) ... *97*
Genetic Variation Exercises ... *100*
Host Response ... *112*
Metabolic Pathways - Exploring pathways and compounds ... *118*
Orthology and Phyletic Patterns ... *127*
Exploring Transcriptomic data ... *136*

RELATED SITES OF INTEREST TO OUR COMMUNITIES ... **162**

VEUPATHDB PUBLICATIONS AND CITATIONS ... **163**

PUBLICATIONS THAT USE OUR RESOURCE ... 164

RELEASE NOTES ... **164**

EXAMPLE RELEASE NOTES - VECTORBASE 48 RELEASED ... 164

ANALYSES METHODS ... **167**

GENOME ANALYSES ... 168
Supplements to the EBI Pipelines ... *172*
In-house genome analyses in lieu of the EBI Pipelines ... *173*
PROTEOMICS ... 175
RNA-SEQUENCE ... 175
CHIP-SEQUENCE ... 176
COPY NUMBER VARIATION ... 176
GENETIC VARIATION AND SNP CALLING ... 177
MICROARRAY DATA ... 177
PROTEIN ARRAY DATA ... 178
METABOLIC PATHWAYS ... 178

DATASET DESCRIPTIONS ... **178**

ORGANISMS - GENOME INFO AND STATS ... **179**

TECHNICAL INFRASTRUCTURE AND SOFTWARE DOCUMENTATION: ... **180**

BROWSER COMPATIBILITY STATEMENT ... 180
DATA LOADING AND DATABASE SCHEMA ... 180
CODE AVAILABILITY ... 181
WEB PRESENTATION SYSTEM AND USER INTERFACES ... 181
SOFTWARE CODE REPOSITORY ... 181
SYSTEM HARDWARE AND THIRD-PARTY SOFTWARE ... 181
OVERVIEW OF THE VEUPATHDB DATA PRODUCTION WORKFLOW AND ARCHITECTURE ... 182

VEUPATHDB WEBSITE PRIVACY POLICY ... **184**

INTRODUCTION ... 184
Information Automatically Collected ... *184*
Information You Directly Provide ... *185*
"Contact Us" Form ... *185*
How VEuPathDB Uses Cookies ... *186*
Google Analytics ... *186*

Third-Party Websites and Applications *186*
Your Rights based on the General Data Protection Regulation (GDPR) *188*

VEUPATHDB ACCESSIBILITY CONFORMANCE **189**

VEUPATHDB PERSONNEL **189**

VEuPathDB Management 189
Current VEuPathDB Team members 189
VEuPathDB acknowledgements 191
VEuPathDB Community Representatives *192*
Previous Scientific Working Group *192*

WEBSITE USAGE STATISTICS **193**

Website usage links: *193*
Sample awstats report from FungiDB.org *194*

VEUPATHDB GLOSSARY **196**

Introduction

This material is made available to our users and NIH administrators to help them use VEuPathDB resources and understand the underlying tools, experiments and analyses provided by this Bioinformatics Resource Center funded in part by the US National Institute of Allergy and Infectious Diseases (Contract HHSN75N93019C00077). Note that the content of this document is also provided through the website, often in a context dependent manner. The web links are provided in the appropriate places throughout.

This report summarizes the categories of user documentation and provides a link to the landing page for each. The documentation can be divided into four broad categories: Site usage; Analysis methods; Dataset descriptors and Technical infrastructure and software documentation.

This report is available from all VEuPathDB sites, e.g. https://veupathdb.org/, from the Help menu.

About VEuPathDB.org

The Eukaryotic Pathogen, Vector and Host Informatics Resource (VEuPathDB) is one of two Bioinformatics Resource Centers (BRCs) funded by the US National Institute of Allergy and Infectious Diseases (NIAID), with additional support from the Wellcome Trust (UK).

These resources stem from support initially provided for the Plasmodium Genome Database by the Burroughs Wellcome Fund (2000-2) and a research grant from NIAID (2002-6). The BRC program was initiated in 2004 to provide public access to computational platforms and analysis

tools enabling collection, management, integration and mining of genomic information and other large-scale datasets relevant to infectious disease pathogens including their interaction with mammalian hosts and invertebrate vectors of disease. Two BRCs are currently funded:

- VEuPathDB focuses on eukaryotic pathogens and invertebrate vectors of infectious diseases, encompassing data from prior BRCs devoted to parasitic species (EuPathDB), fungi (FungiDB) and vector species (VectorBase).
- BV-BRC - PATRIC & VIPR focus on bacterial and viral pathogens.

VEuPathDB has also received funding from the Wellcome Trust (UK), Bill & Melinda Gates Foundation, and US Department of Agriculture to support informatics efforts focusing on kinetoplastida and fungal organisms with special emphasis on improving functional annotation for select genome sequences and families of genes.

VEuPathDB provides access to diverse genomic and other large scale datasets related to eukaryotic pathogens and invertebrate vectors of disease (see Genome Info and Stats). Organisms supported by this resource include (but are not limited to) the NIAID list of emerging and re-emerging infectious diseases.

Component web sites are constructed using a common infrastructure and standard data analysis and loading procedures, allowing the use of VEuPathDB as a single point of entry for each (or all) of these community resources, and the opportunity to leverage orthology for searches across taxa.

- Organisms - Genome Info and Stats provides a list of all organisms available in this website.
- Data Sets provides a list of all information in this website integrated into VEuPathDB, with relevant references.

Current Funding
The Eukaryotic Pathogen, Vector and Host Informatics Resources (VEuPathDB) is funded by the National Institute of Allergy and Infectious Diseases (NIH/DHHS) under Contract No. NIH HHS 75N93019C00077.

VEuPathDB also receives funding from the Wellcome Trust (UK) to support informatics efforts focusing on kinetoplastida and fungal organisms with special emphasis on improving functional annotation of genomes. Grant numbers: 212929/Z/18/Z and 218288/Z/19/Z.

About OrthoMCL

Current Release 6.3
In this release, 1 Peripheral species was added. Thus, OrthoMCL now predicts orthology for a total of 564 organisms (414 Peripheral and 150 Core organisms). Proteins from the 1 new species were mapped into Core ortholog groups. Then, a new set of residual ortholog groups (e.g. OG6r3_101799) were formed from the collection of all unmapped Peripheral proteins. See below for the methods.

To see the current set of organisms as well as their proteome sources and orthology statistics, go to Proteome Sources and Statistics.

Downloads: Go to the download site to obtain the protein sequences and ortholog groups used in this release.

Method for Forming and Expanding Ortholog Groups

Proteins are placed into Ortholog Groups by the following steps:

1. The OrthoMCL algorithm (see below) is employed on proteins from a set of 150 Core species to form Core ortholog groups. These species were carefully chosen based on proteome quality and widespread placement across the tree of life. Each Core protein is placed by the algorithm into a Core ortholog group consisting of one or more proteins. Core group names have the format OG6_xxxxxx (e.g., OG6_101327). OG6 refers to OrthoMCL release 6; for each sub-release (e.g., 6.1, 6.2, etc), the Core species and the Core ortholog group names will remain constant.
2. The proteins from hundreds of additional organisms, termed Peripheral organisms, are mapped into the Core groups. To do this, NCBI BLASTP is used to compare each Peripheral protein to each Core protein in the Core groups. (Note that Peripheral proteins that were previously added to the Core group are NOT used in the BLASTP.) Then, each Peripheral protein is assigned to the Core group containing the Core protein with the best BLAST score, but only if the E-Value is <1e-5 and the percent match length is >=50%.
3. All Peripheral proteins that fail to map to a Core group are collected and subjected to independent OrthoMCL analysis, forming Residual groups consisting of one or more proteins. Residual group names have the format OG6r1_xxxxxx (e.g., OG6r1_101327), where OG6 refers to release 6 and r1 refers to sub-release 1.
4. For each subsequent sub-release (which will occur every ~3 months along with other VEuPathDB sites), proteomes from additional Peripheral organisms will be processed as in steps 2 and 3 above. However, step 3 will differ slightly because the previous set of Residual groups will be disassembled, leaving the previous unmapped Peripheral proteins to be combined with the new unmapped Peripheral proteins. All of these proteins will be used to form new Residual groups (e.g., OG6r2_xxxxxx).
5. On occasion, the set of Core species will be re-defined, as more appropriate proteomes become available. In this case, new Core groups (e.g., OG7_xxxxxx) and Residual groups (e.g., OG7r1_xxxxxx) will be formed.

This design allows for the addition of proteomes at every sub-release (e.g., 6.1, 6.2, etc). Note that Core groups (e.g., OG6_101327) will remain between sub-releases, though these groups will expand as Peripheral proteins are mapped in. In contrast, Residual groups will exist only for that sub-release; thus, Residual groups are useful in allowing the user to find proteins related to their protein(s) of interest, but are not stable groups.

The OrthoMCL Algorithm

See the OrthoMCL Algorithm Document for a detailed description of the OrthoMCL algorithm.

In overview:

- All-v-all BLASTP of the proteins.
- Compute *percent match length*
  - Select whichever is shorter, the query or subject sequence. Call that sequence S.
  - Count all amino acids in S that participate in any HSP.
  - Divide that count by the length of S and multiply by 100
- Apply thresholds to blast result. Keep matches with E-Value < 1e-5 percent match length >= 50%
- Find potential inparalog, ortholog and co-ortholog *pairs* using the Orthomcl Pairs program. (These are the pairs that are counted to form the *Average % Connectivity* statistic per group.)
- Use the MCL program to cluster the pairs into groups

Background on Orthology and Prediction

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Detection of orthologs is becoming much more important with the rapid progress in genome sequencing (Glover et al. 2019).

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

OrthoMCL is similar to the INPARANOID algorithm (Remm et al. 2001) but is extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO (Lee et al. 2002), and an analysis using EC number suggests a high degree of reliability (Li et al. 2003).

We evaluated the performance of seven widely-used orthology detection algorithms that use three general prediction strategies: phylogeny-based, evolutionary distance-based and BLAST-based (Chen, et al. 2007). Specifically, we used Latent Class Analysis (LCA), a statistical technique appropriate for testing large data sets when no gold standard is available. Our results show an overall trade-off between sensitivity and specificity among these algorithms, with INPARANOID and OrthoMCL performing best with False Positive (FP) and False Negative (FN) error rates lower than 20%.

Software

OrthoMCL was originally implemented by Li Li. The software was not available for download.

Version 1.4 was developed as publicly available software by Feng Chen (This version is now not supported).

Version 2.0 was re-engineered to handle large-scale datasets (hundreds of genomes) by Steve Fischer, Mark Heiges, John Iodice, and Ryan Thibodeau.

Data Access Policy

All data on these websites are provided freely for public use, through the contributions of many researchers involved in generating genome sequences, functional genomics datasets, and additional information. These data often derive from ongoing research and are not guaranteed to be accurate. When using data obtained from VEuPathDB, it is important to cite the original publications and contributors. Please see Citing VEuPathDB.

Citing VEuPathDB in Publications and Presentations

If you use a VEuPathDB resource, we invite you to please cite the most relevant publication. This PubMed filter provides a list of the most recent VEuPathDB publications.

Please note that much of the data in VEuPathDB is provided by independent researchers, please cite them if you use their data. See Data Sets for a list of all information integrated into VEuPathDB, and related publications.

For acknowledgements in presentations, you may wish to use one or more of the following logos (right/control click to copy):

Additional resources leveraging the same infrastructure: MicrobiomeDB and ClinEpiDB.

Community Interactions and Data Submission Policies

VEuPathDB serves a global scientific community that demands direct active support and community involvement. VEuPathDB outreach activities include:

- Organizing and running hands on training workshops and webinars (Google Map).
- Developing educational material in the form of exercises and online tutorials.
- Responding to support emails for users who contact us directly by clicking the "Contact Us" links in the header or footer of any VEuPathDB webpage (average response time is 48 hours).
- Holding open community meetings/forums with our diverse user base. These meetings are held in person at scientific conferences or using an online conferencing platform.
- Attending national and international meetings with active participation in the form of posters, presentations or help desks.
- Authoring peer reviewed manuscripts.
- Maintaining active social media presence in the form of a FaceBook page and Twitter feed.
- Providing a clear data handling and release policy to investigators to encourage submission of data prepublication.

How to Submit Data to VEuPathDB

The Eukaryotic Pathogen, Vector & Host Informatics Resources (https://VEuPathDB.org) is a Bioinformatics Resource Center (BRC) operated under contract from the US National Institute of Allergy and Infectious Diseases (NIAID) and the Wellcome Trust. VEuPathDB is charged with ensuring that genomic (and other large-scale) datasets pertaining to supported pathogens are conveniently accessible to the worldwide community of biomedical researchers. This document summarizes policies associated with releasing datasets on VEuPathDB. Our goal is to help the communities that we serve ensure that their data are FAIR, Findable, Accessible, Interoperable and Reusable.

VEuPathDB welcomes submissions of genomic-scale data concerning eukaryotic microbes, fungi, vectors of human disease, and host-pathogen interactions. The VEuPathDB contract from NIAID provides support for biosecurity pathogens, including *Babesia*, *Cryptosporidium*, *Entamoeba*, *Giardia*, *Microsporidia* (various genera), *Toxoplasma*, *Plasmodium*, and related taxa (*Acanthamoeba*, *Gregarina*, *Neospora*, *Theileria*) and also arthropod vectors (ticks, mites, mosquitoes, kissing bugs, tsetse flies, sand flies, lice, etc.) of human disease, as well as a snail that serves as an intermediate host, and comparator species. Support for kinetoplastid parasites

(*Crithidia*, *Endotrypanum*, *Leishmania*, *Trypanosoma*) is provided by The Wellcome Trust. The FungiDB project encompasses a large (and growing) number of species supported by both NIAID and the Wellcome Trust. Please contact us if you have data from other species that should be incorporated into VEuPathDB! **Please review our Data Submission Policy**.

Our most common data types include transcriptomics, proteomics, metabolomics, epigenomics, population-level and isolate information. In one form or another, VEuPathDB currently represents datasets in the following categories:

- Sequence (genomic [nuclear and organellar])
- ESTs and RNA-seq, generated on various platforms)
- Host-response data
- Comparative genomic information
- DNA polymorphism and population genetics data
- Sequences and metadata pertaining to field and clinical isolates and collections (with geo-spatiotemporal and other metadata)
- Chromatin modification data (ChIP-chip and ChIP-seq)
- Manually curated and automatically generated gene models and other annotation (GO terms, InterPro domains, etc.)
- Transcript and proteomic profiling
- Host response data sets (multiple platforms)
- Interactome data
- Protein structural information
- Metabolic pathways and metabolomics data
- Phenotype information, reagents (clones, antibodies, etc.)
- Publication references
- Image data, etc.
- Restriction Fragment Length Polymorphism

We also accept other genomic-scale data and are open to suggestions. Use the Contact Us link to make suggestions. We look forward to working with you!

How to submit data for integration in VEuPathDB

To submit your data for integration, fill out the appropriate VEuPathDB Dataset nomination form listed below. If your data cannot be submitted via our forms, use the Contact Us link to send a brief description (two or three sentences) of your data.

Genomes & high throughput sequencing data (e.g. RNA-Seq, ChiP-Seq, isolates typed by Whole Genome Sequencing or by sequencing limited genetic loci) must be available in The International Nucleotide Sequence Database Collaboration (INSDC) such as NCBI GenBank, EMBL-EBI ENA or DDBJ.

Once the dataset is prioritized for loading, we will export the data directly from INSDC. Note: while genome sequences must be available in INSDC, functional annotation (e.g. gene names, GO terms, etc.) can be submitted directly to VEuPathDB.

Tell us about your data as early as possible, to allow ample time for scheduling into VEuPathDB release cycles. Depending on the dataset type, we can provide instructions on how to transfer

your data to us (e.g. formats of proteomics datasets may differ depending on the nature and scale of the data to be transferred), or we may be able to facilitate data submission to a repository (e.g. GenBank, GEO/ArrayExpress, etc.).

VectorBase resource: Gene manual annotations (change of exon-intron boundaries, creation of new genes) and metadata (gene names/symbols and functional description) can be submitted via Apollo (Coming soon). If you submit a gene annotation before you submit a manuscript for publication, we can generate GeneIDs that can be linked out to within the publication. Gene deletions are not handled via Apollo, please Contact Us with supporting evidence. Metadata can also be submitted by sending a spreadsheet file for batch submissions, follow this link for information about the 12 columns heading that are required. To add publications to a gene send us the corresponding PubMed link.

Genome Sequence and/or Annotation

Genomes must be available in The International Nucleotide Sequence Database Collaboration (INSDC) such as NCBI GenBank, EMBL-EBI ENA or DDBJ.

- If your genome **IS** uploaded to a repository, complete the Genome Sequence and/or Annotation Description Form making sure to include the accession numbers of your data when prompted. We will download your data from the repository.
- If the annotation file **is not** uploaded to a repository, use the Contact Us form to send us the genome annotation file only (e.g. gff file format).

High Throughput or Next Generation Sequencing – RNA, DNA or ChIP Sequencing

We prefer to download the raw read data in FASTQ format from a sequence read archive. We integrate the data into the database using the raw reads and use the raw reads during future database releases to remap or update our analyses when necessary.

How to transfer a copy of your data to VEuPathDB:

- Upload your data to a sequence read archive such as the European Nucleotide Archive or NCBI's Sequence Read Archive and provide us the accession numbers. We will export the data directly from INSDC
- Complete the appropriate data description form making sure to enter your data archive accession numbers when prompted

RNA-Seq Data Description Form

DNA-Seq Data Description Form

ChIP-Sequencing Data Description Form

Microarray

Transfer a copy of your data to VEuPathDB using one of these options:

- Upload your data to a repository such as GeneExpression Omnibus. Complete the data description form linked below making sure to enter your data archive accession numbers when prompted. We will export the data directly from a repository.
- Send your data as an attachment to an email. Use the Contact Us form to send us an email. Files (e.g. CEL, CSV) must include expression levels and probe set information. Pay special attention to clearly indicate the identity of columns in the data files you transferred to VEuPathDB.

Microarray Data Description Form

Proteomics

Excel or tab delimited text files are preferred. We can accommodate xml file format. Required columns include gene IDs, peptide sequences, peptide counts and scores.

How to transfer a copy of your data to VEuPathDB:

- Send your data as an attachment to an email. Use the Contact Us form to send us an email.
- Complete the Proteomics Data Description Form making sure to clearly indicate the content of each column in your file.

Quantitative Proteomics

Excel or tab delimited files are preferred. We can accommodate xml file format. Required columns include gene IDs and scores.

How to transfer a copy of your data to VEuPathDB:

- Send your data as an attachment to an email. Use the Contact Us form to send us an email.
- Complete the Quantitative Proteomics Data Description Form making sure to include a description of data columns, for example, time course units and arrangement if not apparent from column headers.

ChIP-chip

Your data files should include expression levels and probe set information.

Transfer a copy of your data to VEuPathDB using one of these options:

- Upload your data to a repository such as Gene Expression Omnibus and complete the ChIP-chip Data Description Form making sure to enter the archive accession numbers (if any) for your data when prompted.
- Send your data as an attachment to an email. Use the Contact Us form to send us an email.

Isolates typed by sequencing limited genetic loci

If your data **IS** uploaded to GenBank, use the Contact Us form to tell us about your data. Note: GenBank Isolate records and the associated metadata are automatically updated with each VEuPathDB release. There is no need to complete our Isolate Submission Form.

If your data **IS NOT** uploaded to GenBank, we can facilitate this upload. Complete the Isolate Submission Form linked below and we will use the information to generate a GenBank submission for your isolates. The new isolate records will be downloaded to VEuPathDB with the release. Use the Contact Us form to send us instructions for retrieving your data.

Isolate Submission Form

Help for submitting Isolate Data

Isolates or Strains typed by High Throughput Sequencing

We prefer to receive the raw read data in FASTQ or FASTA file format. We integrate your data into the database using the raw reads. We also use the raw reads during future database releases to remap your data when the reference genome is reloaded and to update our analyses when needed.

How to transfer a copy of your data to VEuPathDB:

- Upload your data to a sequence read archive such as DNA Data Bank of Japan, the European Nucleotide Archive or NCBI's Sequence Read Archive. We will retrieve your data using the read archive's accession numbers for your data set.
- Complete our DNA Seq Data Description Form making sure to enter the read archive accession numbers for your data when prompted. We also ask that you complete an abbreviated Abbreviated Isolate Submission Form to describe meta data associated with your isolates.

General Data Submission – Use the Contact Us form to tell us about data that does not fit any of the above categories

Data submission and release on VEuPathDB databases

issued February 2010, most recent revision 02 April 2020

The Eukaryotic Pathogen, Vector & Host Informatics Resources (http://VEuPathDB.org) is a Bioinformatics Resource Center (BRC) operated under contract from the US National Institute of Allergy and Infectious Diseases (NIAID) and the Wellcome Trust. VEuPathDB is charged with ensuring that genomic (and other large-scale) datasets pertaining to supported pathogens are conveniently accessible to the worldwide community of biomedical researchers. This document summarizes policies associated with releasing datasets on VEuPathDB and affiliated knowledgebases. Our goal is to help the communities that we serve ensure that their data are FAIR, Findable, Accessible, Interoperable and Reusable.

General principles:

- ***Data providers define the schedule for data release (in consultation with funders, publishers, etc).*** While there is no point in providing VEuPathDB with data that will never become public, deposition does not in itself authorize immediate release. Data become accessible to the public only when the data providers and VEuPathDB staff agree that it is accurately represented and ready to go live. Note that knowledgebase staff are not active research scientists; they are distinct from researchers in the groups responsible for VEuPathDB, who see new data only when it becomes accessible to the general public.
- ***Data providers know their data best.*** We expect to work with those who generated the underlying data to determine how best to analyze and represent new data types. This typically means taking in relatively raw data – often earlier, and in a more unprocessed form than the published dataset – and building an in-house analysis pipeline to ensure that all comparable datasets are handled similarly.
- ***The earlier we learn about new datasets, the easier it is to schedule timely release.*** The nature of our knowledgebase production, and competing demands from the many communities we support, means that several months' notice are often required to prepare for release. Note that it is often possible to use a preliminary dataset for planning, which can be swapped for the final version before public release.
- ***Experience has shown that data not deposited prior to publication often fails to emerge at all!*** After publication, it may be difficult to focus on tracking down the raw data, associated metadata, analysis methods, etc. It is never too early to discuss planned datasets with the VEuPathDB team!
- ***While not required, pre-publication data release often results in favorable attention from scientific colleagues (including journal editors and grant reviewers).*** Note that all major scientific journals now agree that early release of genomic-scale datasets does not compromise publication.

Why submit my data to VEuPathDB?

- Inclusion in VEuPathDB facilitates your own analysis of the data, in the context of other genomic-scale experiments already available from researchers around the world.
- Electronic access permits others to analyze your data in greater depth than possible in print (even in advance of publication, if you wish to allow this).
- Availability within VEuPathDB keeps your data alive on a highly visible genomics knowledgebase resource: VEuPathDB is accessed by ~13,000 unique users each month.

How do I submit data to VEuPathDB?

- Fill out the appropriate form to indicate the data availability
- Contact the VEuPathDB by clicking the ‘Contact Us’ link on any VEuPathDB page or emailing us at help@VEuPathDB.org.
- Tell us about your data as early as possible, to allow ample time for scheduling into VEuPathDB release cycles.
- Once you tell us about your data, we will provide instructions on how to transfer your data to us (formats may differ depending on the nature and scale of the data to be transferred).

- In order to avoid any confusion and ensure accuracy, we adhere to strict Standard Operating Procedures (SOPs), as outlined below.

What data types are supported by VEuPathDB?

In one form or another, VEuPathDB currently represents sequence (genomic [nuclear and organellar], ESTs and RNA-seq, generated on various platforms), host-response data, comparative genomic information, DNA polymorphism and population genomics data, sequences and metadata pertaining to field and clinical isolates and collections (with geo-spatiotemporal and other metadata), chromatin modification data (ChIP-chip and ChIP-seq), manually curated and automatically generated gene models and other annotation (GO terms, InterPro domains, etc.), transcript and proteomic profiling, host response data sets (multiple platforms), interactome data, protein structural information, metabolic pathways and metabolomics data, phenotype information, reagents (clones, antibodies, etc.), publication references, image data, etc. We can support additional data types as needed.

Please let us know if you have data to provide that is not currently supported! What species are supported by VEuPathDB?

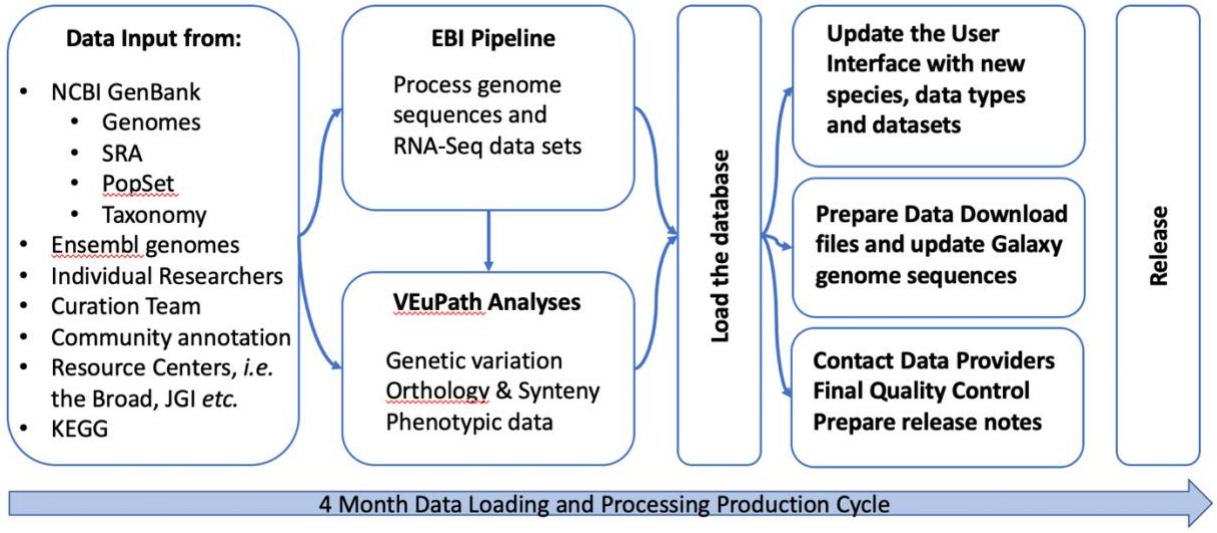
The VEuPathDB contract from NIAID provides support for biosecurity pathogens, including *Babesia*, *Cryptosporidium*, *Entamoeba*, *Giardia*, *Microsporidia* (various genera), *Toxoplasma*, *Plasmodium*, and related taxa (*Acanthamoeba*, *Gregarina*, *Neospora*, *Theileria*) and also arthropod vectors (ticks, mites, mosquitoes, kissing bugs, tsetse flies, sand flies, lice, etc.) of human disease, as well as a sail that serves as an intermediate host, and comparator species. Support for kinetoplastid parasites (*Crithidia*, *Endotrypanum*, *Leishmania*, *Trypanosoma*) is provided by The Wellcome Trust. The FungiDB project encompasses a large (and growing) number of species supported by both NIAID and the Wellcome Trust. ***Please contact us if you have data from other species that should be incorporated into VEuPathDB!***

The Data Production Cycle

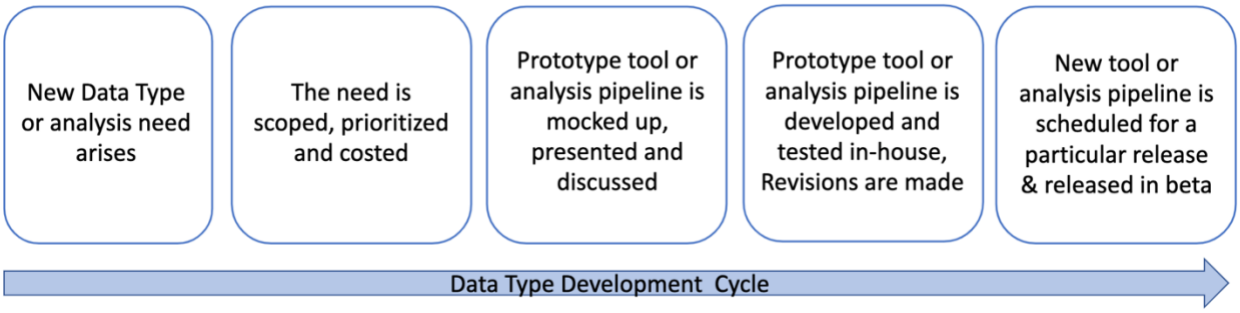
Guiding Principle:

- We balance the need to process as many datasets as possible from known data types in our production pipeline with the need to scale in capacity and containerize analyses for these same data types and the need to continually build new tools and infrastructure in our data type development to accept new data types and facilitate emerging community analysis needs as they emerge.

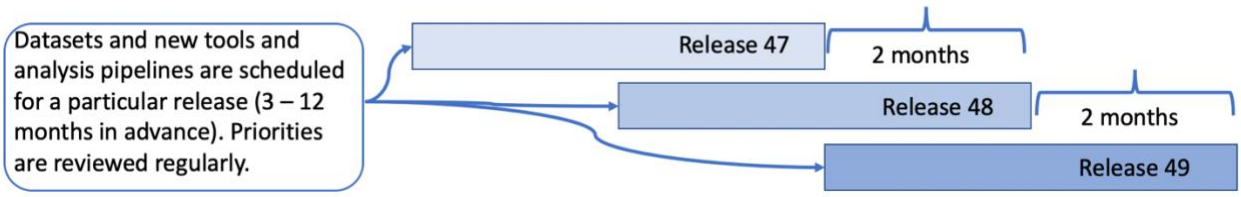
(A)



(B)



(C)



High level views of the VEuPathDB Production Cycle (A), Datatype Development Cycle (B), and Release Timeline (C; overlapping 4 month cycles of production and data type development with bimonthly releases).

Data Management SOPs (Standard Operating Procedures) for VEuPathDB databases

VEuPathDB routinely handles datasets provided prior to publication, in addition to those already in the public domain. In order to ensure timely and accurate data integration we strictly adhere to the following Standard Operating Procedures (SOPs):

1. Datasets come to our attention in several ways, including

- Direct contact from researchers generating the data (during the earliest stages of project design, as
    - data is being produced, in the course of data analysis, or in the context of manuscript preparation).
    - Information provided by members of the VEuPathDB or larger pathogen community.
    - Information obtained by VEuPathDB staff at meetings and conferences.
    - Publicly available information from the scientific literature, genomic dataset repositories, *etc*.

** Note that VEuPathDB can often facilitate data deposition in the appropriate archival repositories (GenBank, SRA, GEO/ArrayExpress, etc.)*

2. ***Decisions to include a dataset in VEuPathDB are based on value to the research community.*** When prioritizing data for integration, we rely heavily on discussions with active researchers, including the scientific advisory committees established for each of the taxonomic groups supported by VEuPathDB and the objectives and priorities of our funders. ***Please contact us if you are interested in participating in these discussions.***
3. ***Regardless of how we first learn about a given dataset, communication is established with the original data producer*** through email, teleconference, and/or face-to-face meetings to discuss the desirability and feasibility of integration into VEuPathDB. In the course of these discussions, we consider what data are likely to be available, data formats and transfer protocols, questions the community may wish to ask of this data, and ways to represent or display such information. We also collect appropriate metadata (regarding samples, experimental protocols, *etc.*), and information on data sources, data providers, appropriate citation, *etc.*
4. ***Data provided to VEuPathDB is housed on secure servers and never shared outside of VEuPathDB staff without prior consent of the data provider.*** Note that database staff are not active researchers; they are distinct from students and postdocs in the groups responsible for VEuPathDB, who see new datasets only when they become accessible to the general public.
5. Datasets are assigned a provisional release date, in consultation with the data provider. ***Scheduling a dataset does not mean that it will be released without final examination and approval by the data provider!*** We operate on the assumption that those who generate the data are best placed to evaluate its proper integration and representation in the knowledgebase. Note that this 'golden rule' applies to both published and unpublished data.
6. Three to four months before the scheduled release date, the data provider is contacted by the ***Outreach*** team, to ensure that we have the most up-to-date version of the data, along with appropriate metadata and information on data sources and citations. The ***Data Loading*** team then processes and integrates this data into our internal knowledgebases.
7. After data loading is complete, the ***Data Development*** team begins to analyze and develop searches against the data. At this point we will likely communicate with the data provider, if questions arise.
8. ***Once data development is underway, the data provider is given access to a***

password-protected version of the VEuPathDB website containing their data. This development site is similar to the current production knowledgebase, except that it also includes new data from the provider. We also provide instruction on how to search and view these new data, including sample searches integrating new data with relevant information already available in the knowldegebase. Important questions to consider include:

- Does the knowledgebase accurately represent your data?
- Are the values and/or graphical displays provided appropriate?
- Are the questions that one can ask of your data appropriate?
- Are there additional questions that you would like to see implemented?
- Are the data appropriately described, including relevant metadata and reference / citation details?

9. ***A series of exchanges typically ensues,*** in which we work iteratively with data providers to address any concerns, with changes reviewed on the password protected site so that providers can view and interrogate their data in the context of the rest of the database.
10. ***Public release is only considered after everyone is satisfied with how the data is represented.*** If the provider is not yet ready to authorize data public release, data is rescheduled for a future release, and removed from the development site before it goes live.
11. Once data are approved for public release, a description is included in the 'News' accompanying the next release, ***highlighting new datasets and functionality, and acknowledging all data providers.***
12. ***Post-release quality assurance*** provides the opportunity to modify displays and develop new queries if/as appropriate.

How to Use Our Sites

We provide our users with a variety of mechanisms to learn about how the VEuPathDB site works and can facilitate their research. In addition to the instructions provided here, users should know that they can learn about VEuPathDB in person, via recorded instructional material or via help located throughout the website, detailed in the list below.

Online instructional material

a. Users can request that a member of the VEuPathDB outreach staff attend a lab meeting or institutional or regional workshop to ask questions and have training in a specific topic
b. User can participate in or view previously recorded webinars (online recordings of training sessions conducted by VEuPathDB staff) that are provided as YouTube recordings.
c. Users can browse and download "quick start" materials and tutorials to follow at their own pace, each with color pictures and well explained prompts to help users

explore the VEuPathDB site or one of its projects, e.g. VectorBase, FungiDB or PlasmoDB. All project use the same underlying architecture and software, only the local environment is customed so user can learn from a tutorial on any organism.

d. VEuPathDB host several workshops annual and all training materials (step by step tutorials with exercises and answers) are provided for all to use and can be easily downloaded.
e. Context sensitive help is available via the many help icons located on the web site, particularly on search pages.
f. Textual descriptions are associated with search pages to provide information about the data type and how to use the search.

To quickly access these online resources, please visit:
https://beta.veupathdb.org/veupathdb.beta/app/static-content/landing.html

Print-based instructional material

OrthoMCL Tutorial

Basic OrthoMCL Functionality

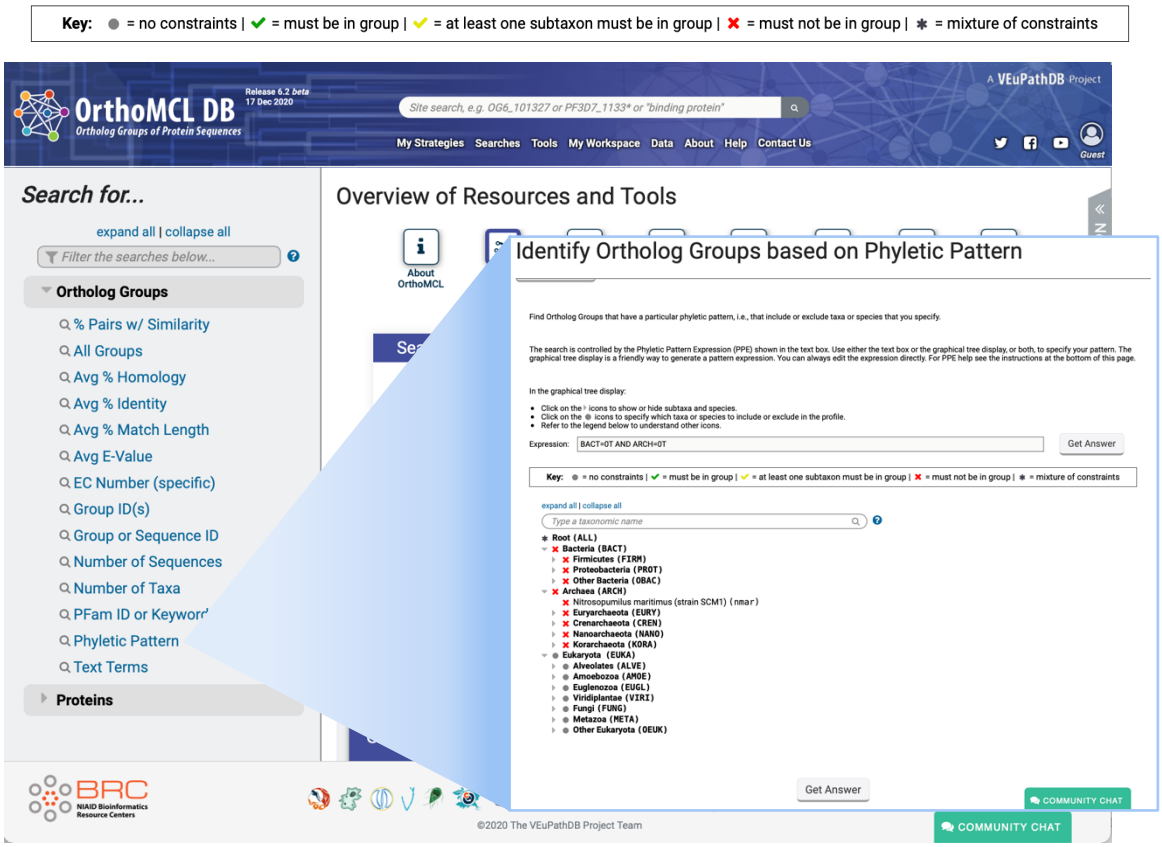
http://orthomcl.org/

Learning objectives:

- *Run searches in OrthoMCL*
- *Run phyletic pattern searches using check boxes or an expression*
- *Combine searches using the strategy system*
- *Explore individual ortholog group pages*
- *Explore the group cluster graphs*

1. Using the Phyletic Pattern search in OrthoMCL

The “Phyletic Pattern” search is an ortholog group search – look under the ortholog groups category and explore the available searches. Can you find the one called “Phyletic Pattern”? There are two ways to specify a phyletic pattern:



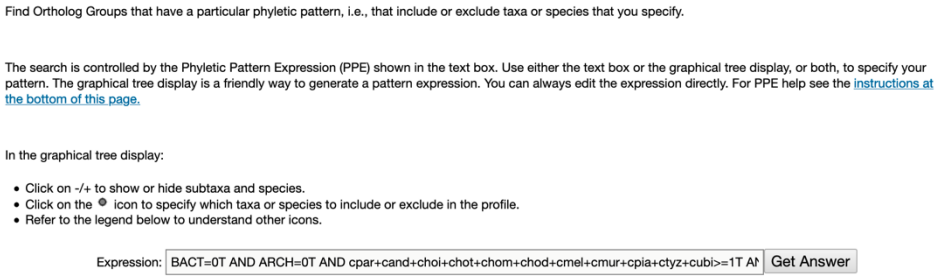
1. Using the expression box. Type the expression using hints available at the bottom of the search page.

2. Using the selectable tree menu. Click on the circle next to the taxon you want to include or exclude.

a. How many protein groups <u>do not</u> contain orthologs from bacteria and archaea?

b. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea. If you are getting frustrated trying to figure this one out, you have a right to be! You cannot answer this question by using the check boxes. However, OrthoMCL has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Can you figure out what expression to use to answer this question? (hint: scroll down to the bottom of the search page to find additional information about expression parameters.

Before looking at the answer below, try this on your own to see if you can figure it out.



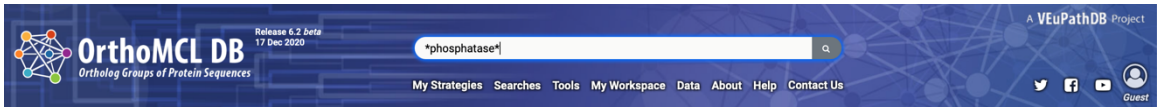
BACT=0T AND ARCH=0T AND cpar+cand+choi+chot+chom+chod+cmel+cmur+cpia+ctyz+cubi>=1T AND gass+gadh+gasb+gabb+gase+gmur>=1T

All VEuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under *Genes -> Orthology and Synteny -> Orthology Phylogenetic Profile*. This search is very useful to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus of interest but not present in the host as these genes may make good drug targets or vaccine candidates.

2. Combining searches in OrthoMCL

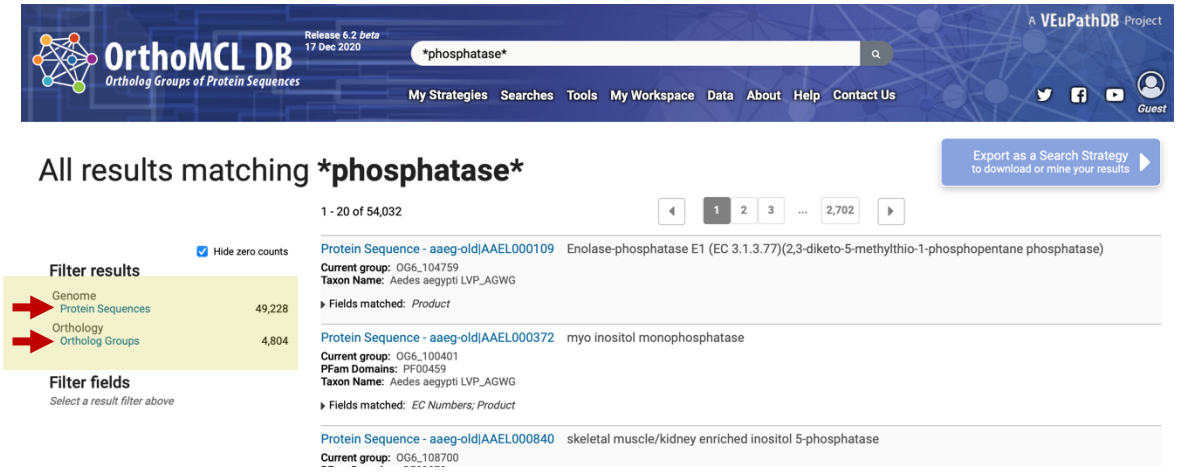
Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

a. Use the site search box at the top of the page in the header **to find OrthoMCL groups** that contain the word "*phosphatase*" (note that the search should be run without the quotation marks but with the asterisks).



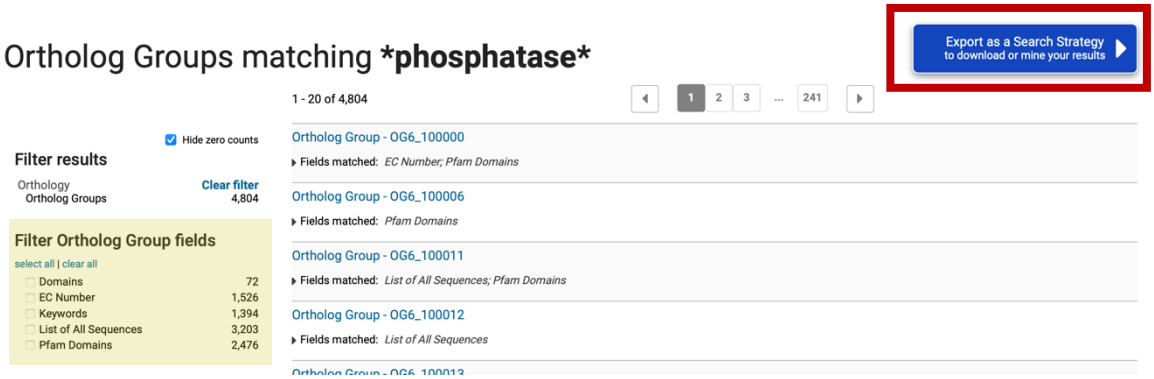
- How many proteins did you identify?
- How many groups did you identify?

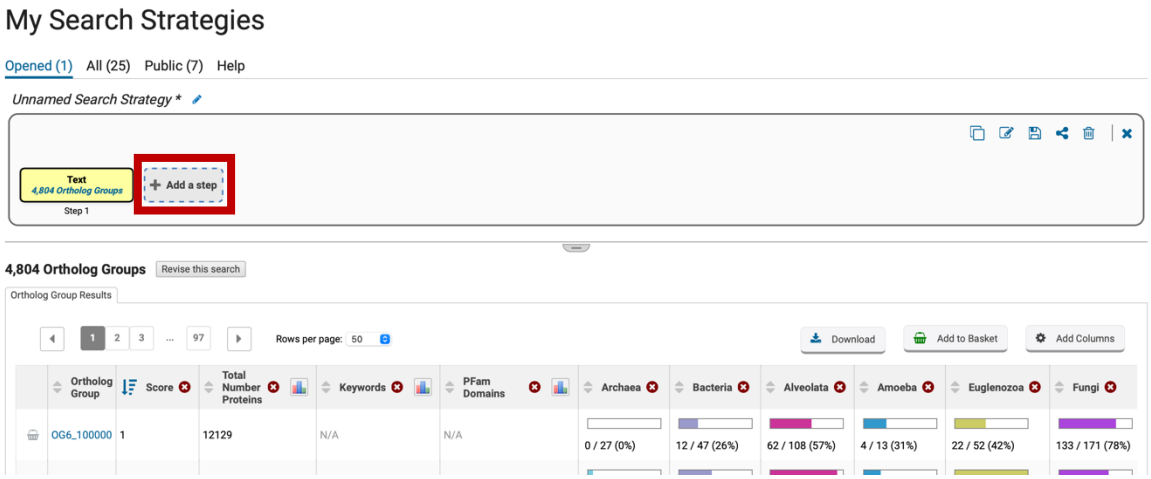
Note: that numbers in screen shots will likely be different on the site due to frequent updates.



b. Click on ortholog groups to filter the results to only show the ortholog groups containing the word phosphatase. Notice that you can filter the group results even further by the group fields. Also, notice that the export as a Search Strategy button is now active. (Note: this is because all the results are now only one type of record: groups.)

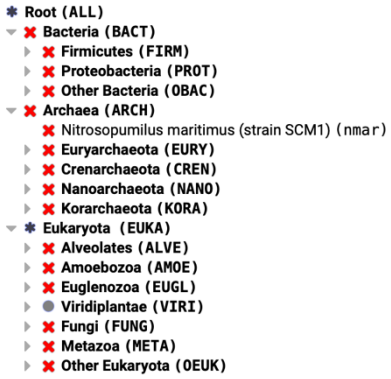
c. Export the ortholog group results as a search strategy by clicking on the blue “Export as a Search Strategy” button at the top right of the page.

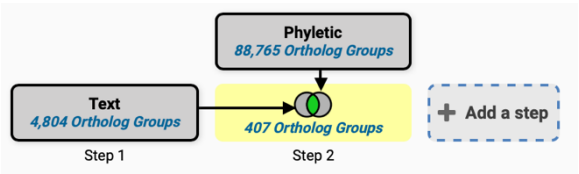




d. Add a step and run a phyletic pattern search for groups that contain any plant protein but do not contain any other organism outside plants. (hint: make sure everything has a red x on it except for plants (Viridiplantae (VIRI)), which should be a grey circle).

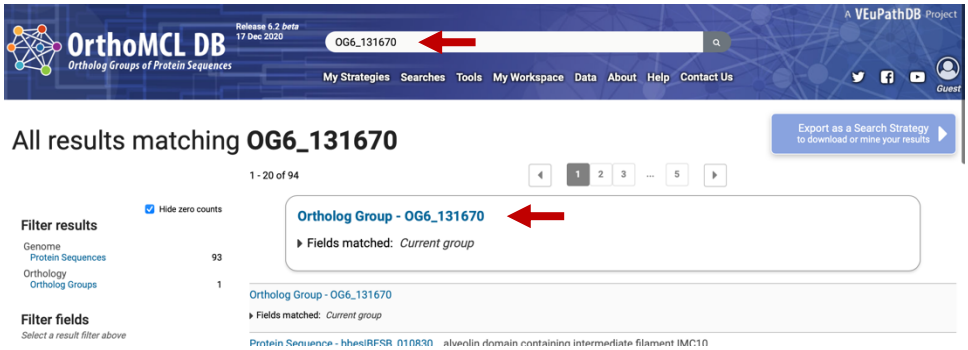
e. How many groups did you return?





3. **Exploring a specific OrthoMCL group - examining the cluster graph.**

a. Visit the OrthoMCL group OG6_131670. You can quickly get to this group by first running a site search with the group ID. Click on the group ID to get to its page.

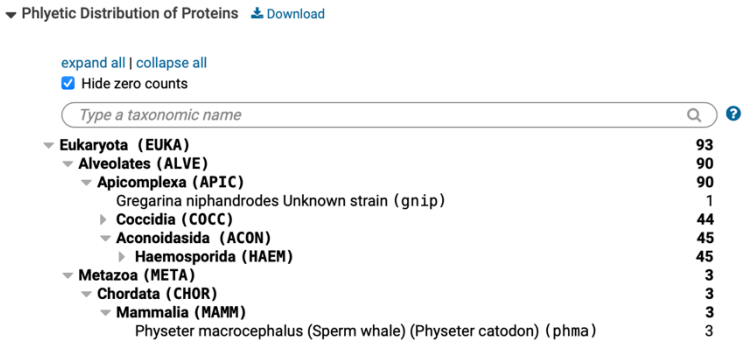


b. The group page is divided into 5 sections:

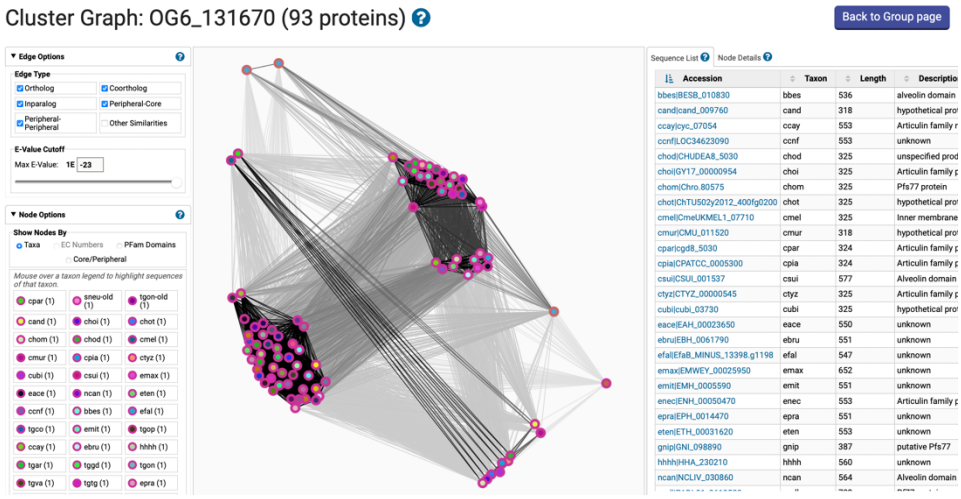
i. Phyletic distribution
ii. Group summary
iii. List of proteins
iv. PFam domains
v. Cluster graph

Examine each of the above sections - is it clear what each section contains?

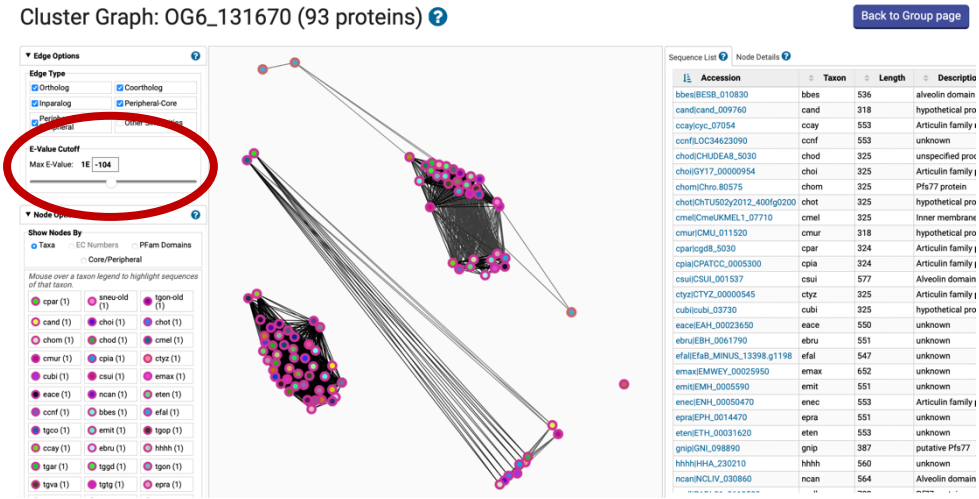
c. Examine the phyletic distribution tree. What taxa does this group contain?



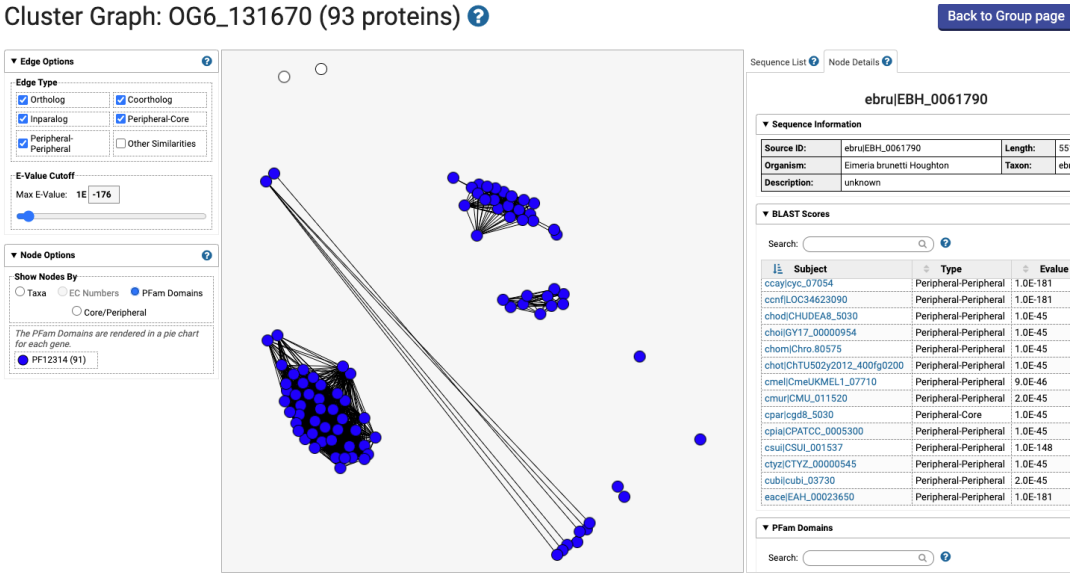
d. Examine the cluster graph for this group (*hint: go to the cluster graph section of the page and then click on the "Click to open the Cluster graph in a new tab"*)



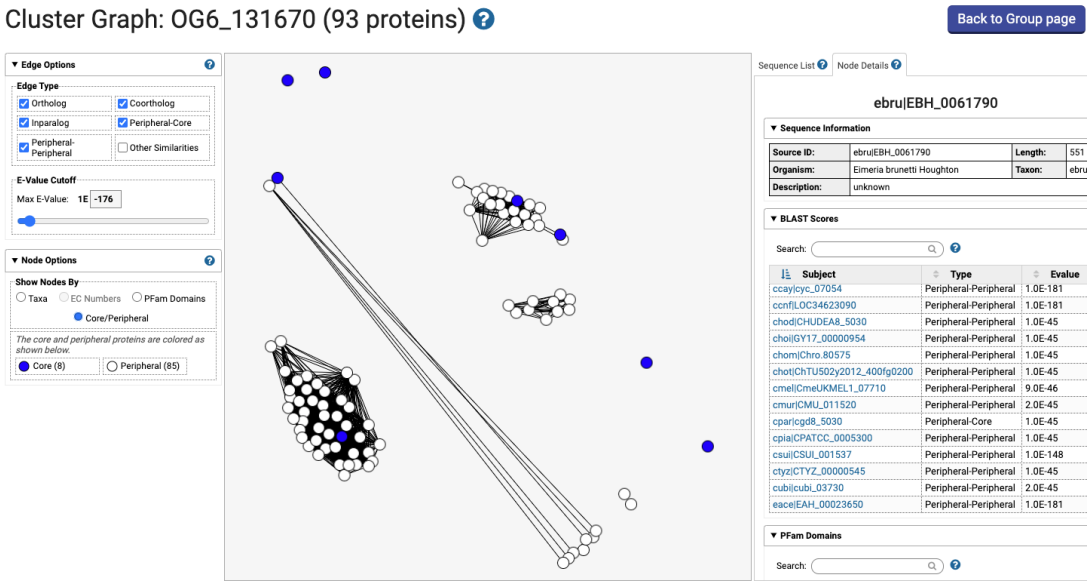
You can interact with the cluster graph. For example, move the slide to increase the E-value cutoff stringency (e.g., to a more negative number). Can you identify subclusters? Click on the nodes in the graph – notice how the organism is updated on the right.



On the left of the page in the Node Options panel, click on PFam Domains to see which proteins have the various PFam domains.



And in the same panel, click on Core/Peripheral to observe which proteins were derived from Core species and which proteins were derived from Peripheral species. Proteins from Core species were used in the initial OrthoMCL algorithm to form Core ortholog groups. Proteins from Peripheral species were mapped into these Core groups by sequence similarity (determined by BLAST score).



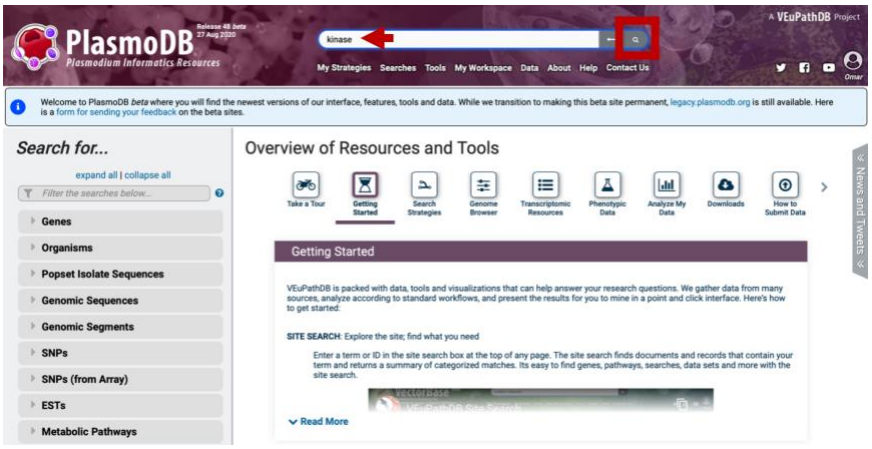
Site Search

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

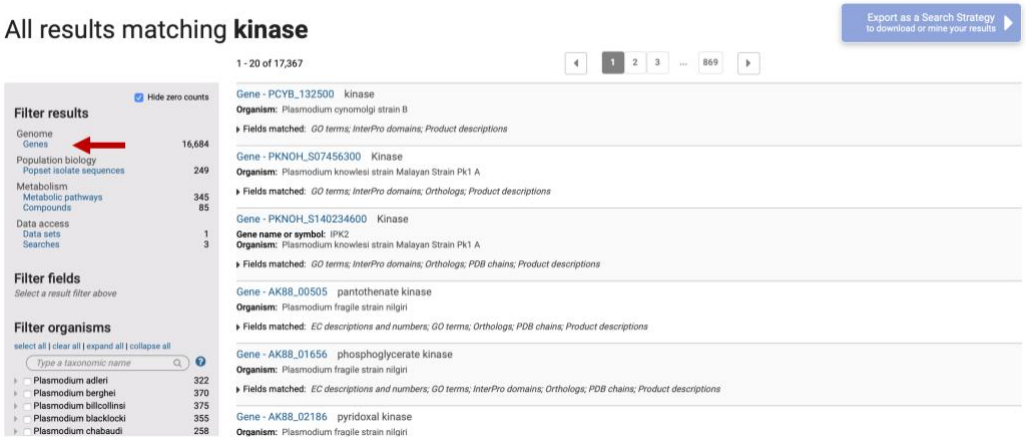
Learning objectives:

- 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

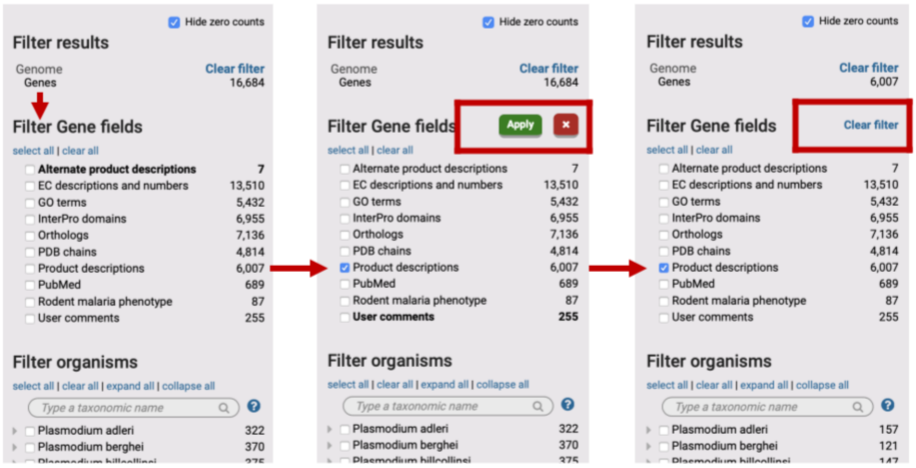
1. Enter the word *kinase* in the site search window (top center of the page, arrow in the image below). Then click enter on your keyboard or click on the search icon (square in the image below).



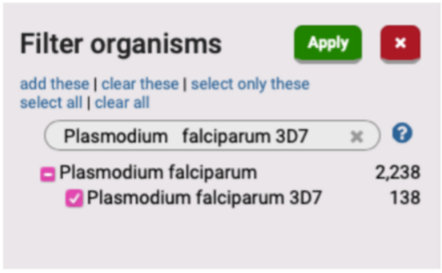
2. How many results with the word kinase did you get? Are all the results genes? Explore the filter panel on the left side of the webpage. Filter the results so that you only view gene results (hint: click on the word *genes* in the *Filter results* section; arrow in image below).

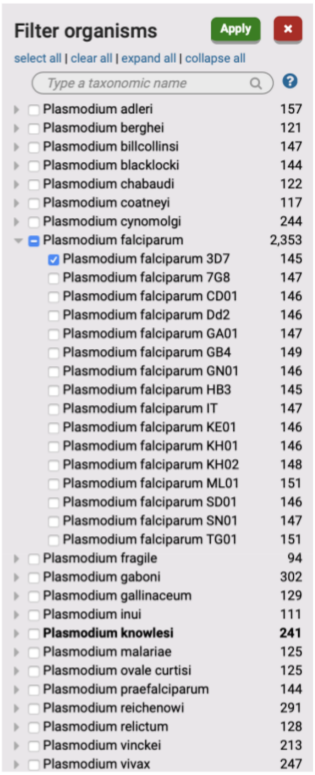


3. How many of the genes included the word kinase in their product descriptions? Notice that once you filter the result by genes (click on the *Genes* filter), the fields section expands to reveal additional filtering options. Once you select the *Product descriptions* field you are provided the option to *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).

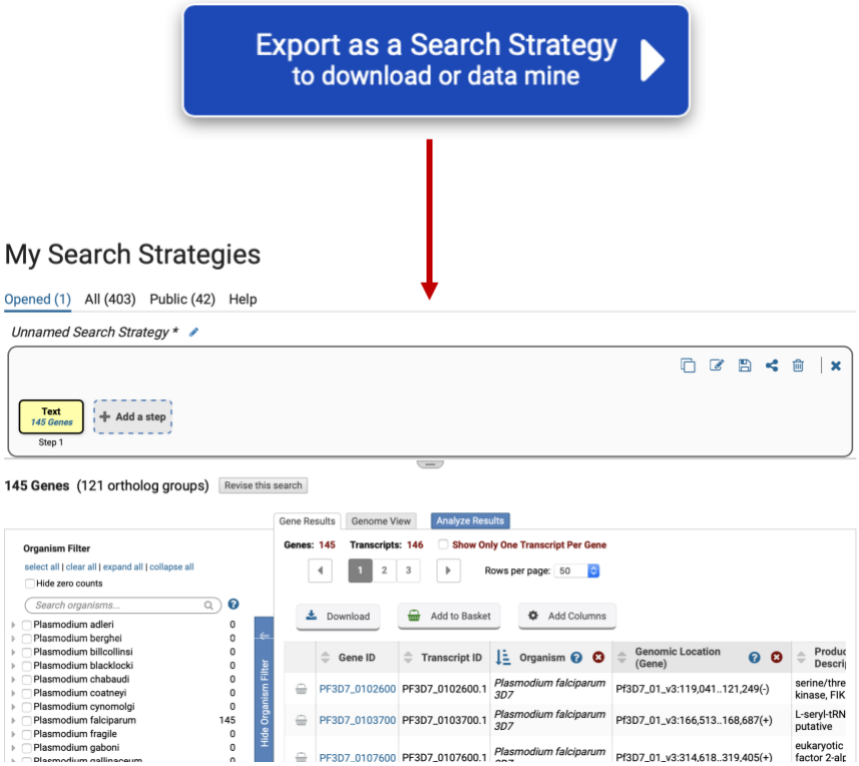


4. How many of the above genes are found in *Plasmodium falciparum* 3D7? How did you find this number? (hint: explore the *Filter organisms* section of the results filter). Select the correct organism and apply the filter.



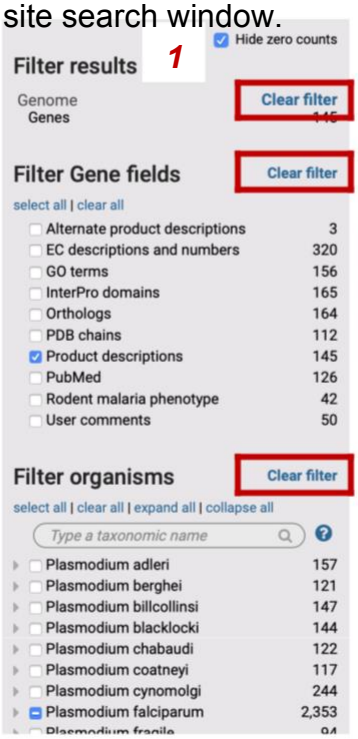


5. Export the results to a search strategy. (hint: to achieve this click on the blue *Export as a search strategy* button at the top right-hand side of the results).

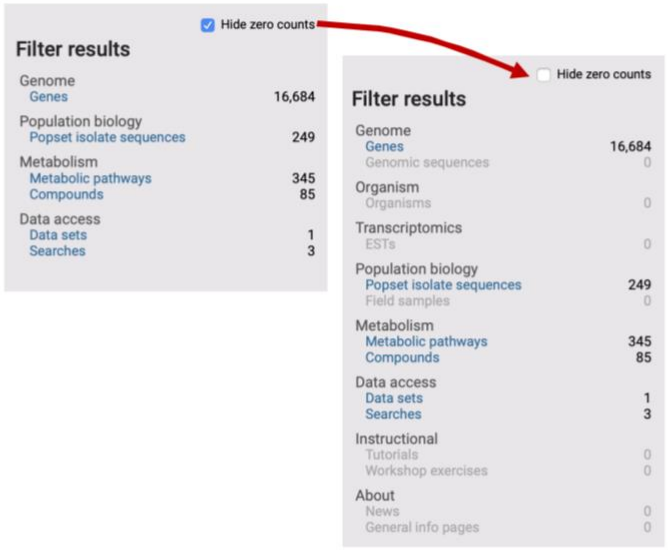


6. Return to the site search results page. How did you do this? (hint: you can achieve this in two ways: 1. Click on your browser's back arrow. 2. Click on the back to results arrow in the site search window. Notice that your previous results and filter settings were preserved.

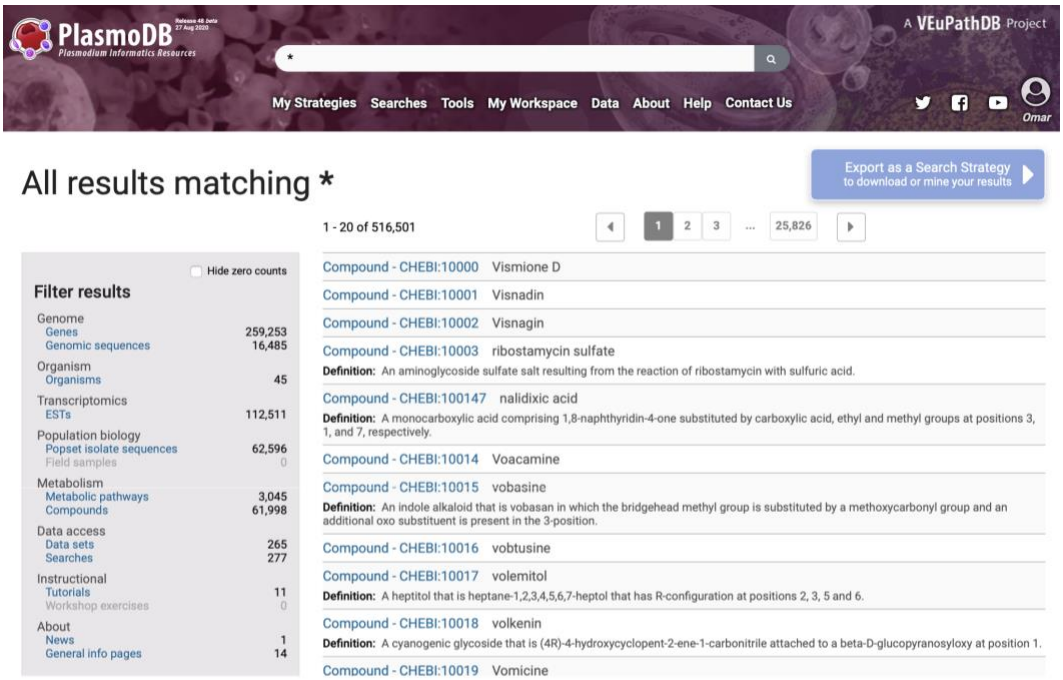
7. Clear all filters. How did you do this? (hint: you can achieve this in two ways: 1. You can click on each of the clear filter options in the filter results panel on the left (boxes below). 2. You can click on the single *clear filters option* in the site search window.

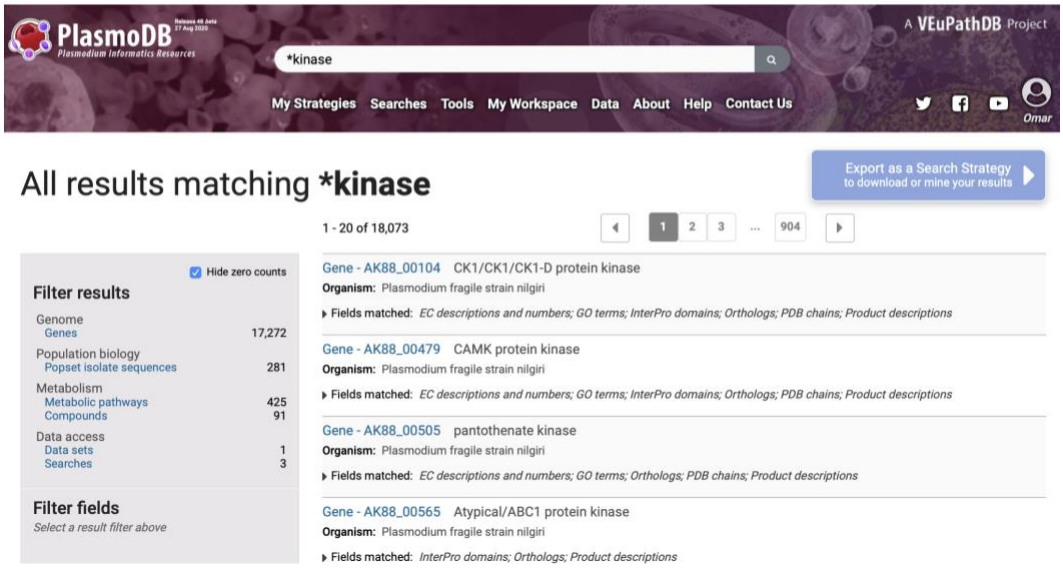


8. Try the *Hide zero counts* check box in the *Filter results* panel. What does this do?



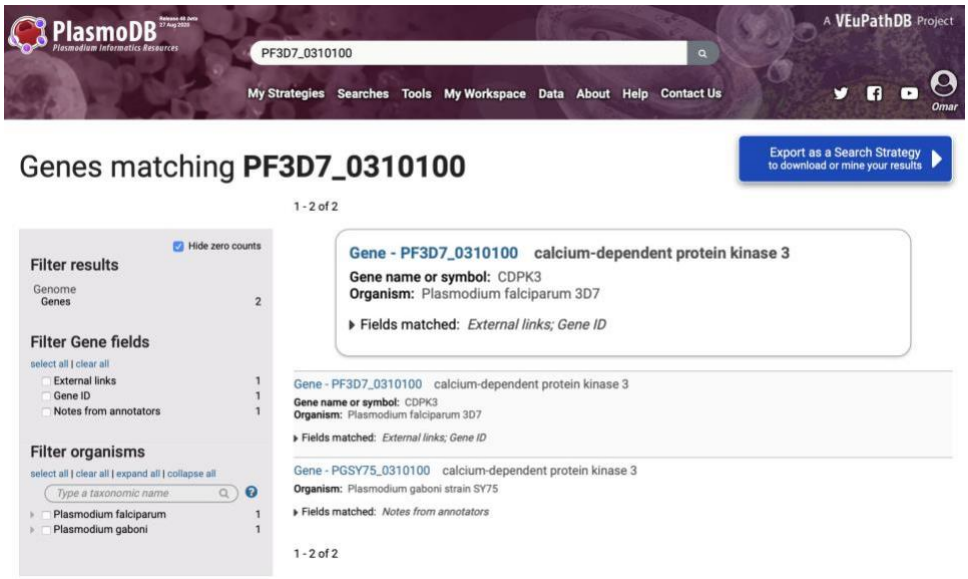
9. Try running a search with a wild card. The wild card is denoted by an asterisk *. The wild card can be used alone to retrieve all results available to the site search or combined with a word such as **kinase* to retrieve compound words ending with the word kinase like phosphofructokinase. As usual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





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

PF3D7_0310100



Notice that the gene of interest appears at the top for easy access. You can click on the Gene ID to go the gene page.

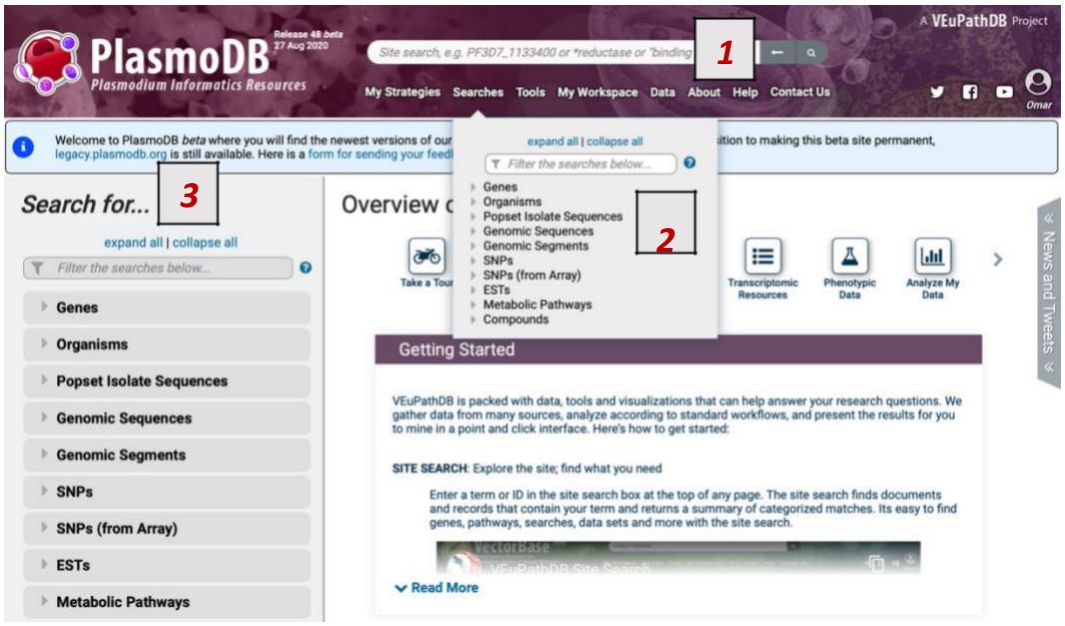
Search Strategies

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

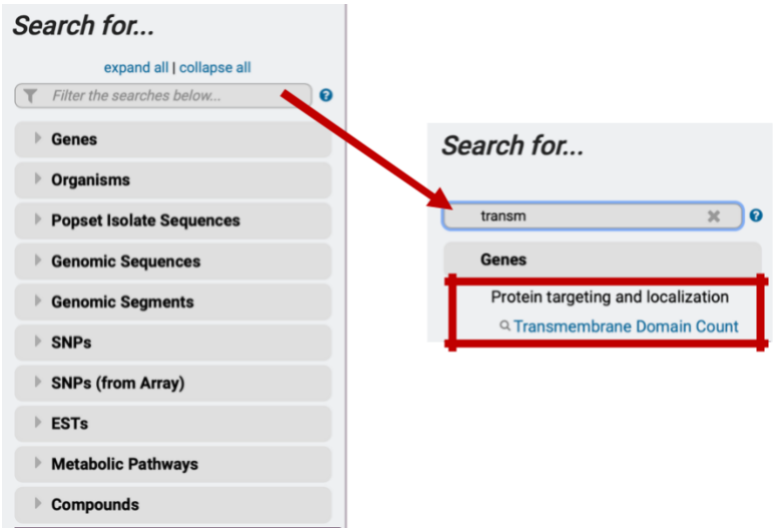
Learning objectives:

- Running a search to start a search strategy
- Adding steps in a search strategy
- Adding and sorting results
- Revising steps

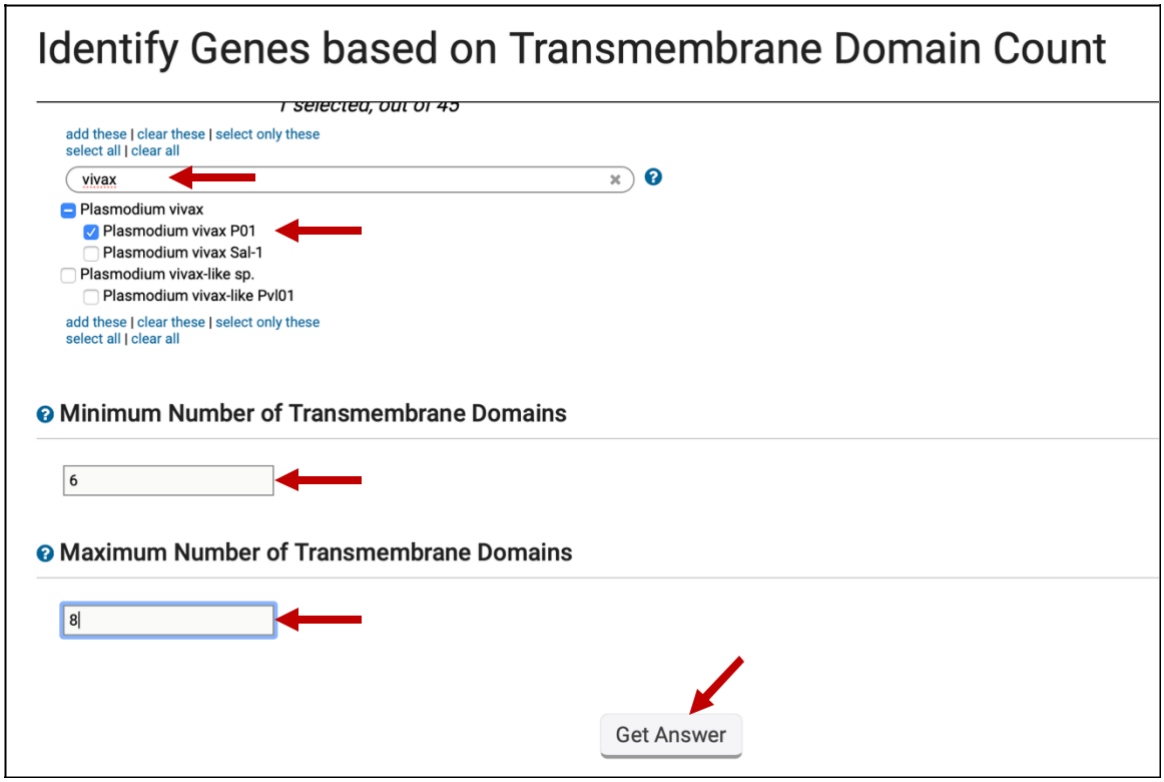
There are three options to start a Search Strategy. 1) From the "Site Search" box ---> Export as Search Strategy, 2) In the site header from the "Searches" menu and 3) In the home page (left hand side) from the "Search for …" section.



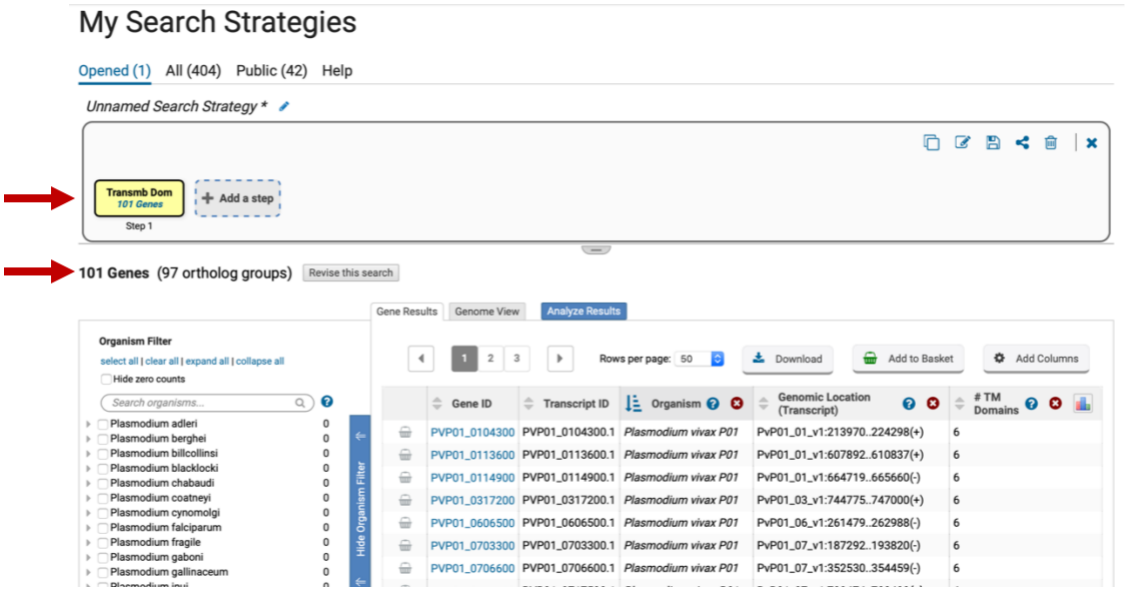
1. Go to the home page and in the "Search for ..." section on the left, filter the searches by typing the word *transmembrane* to find the ***Transmembrane Domain Count*** **search** in the filtered results.



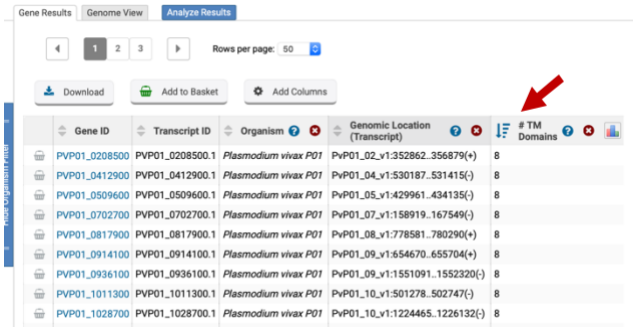
2. Click on the transmembrane (TM) domain count search to get to the search page. Configure this search to find all genes from *Plasmodium vivax* P01 that have at least 6 TM domains and at most 8 TM domains. See image below if you need help with the configuration.

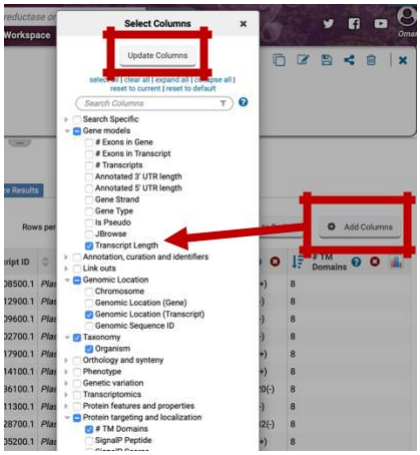


3. How many genes did you obtain? (hint: look at the number results in the strategy step in yellow, or the number right above the results and below the search strategy).

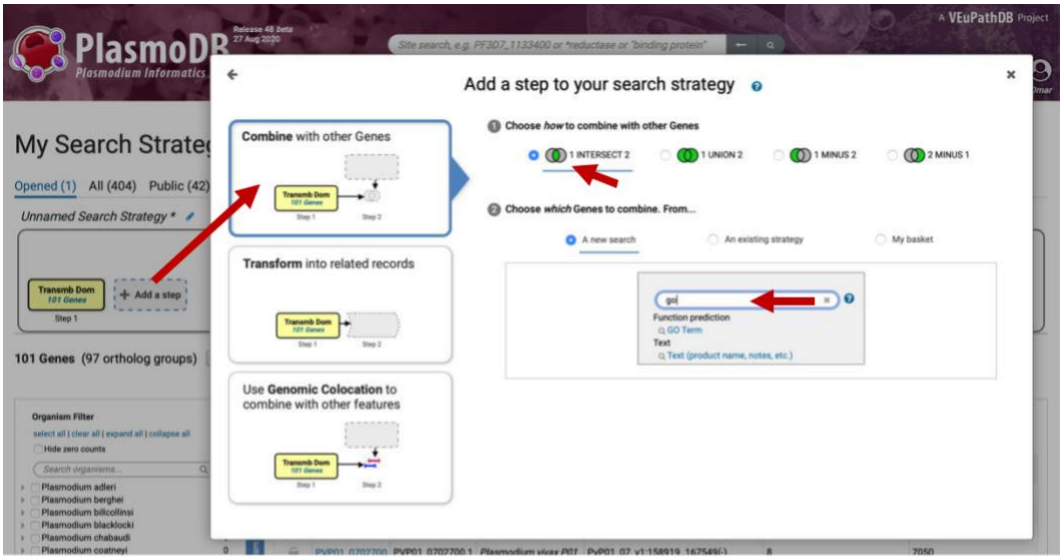


4. Explore the results table. Try the following things:
   a. Sort the #TM domain column to show genes with 8 TM domains first.
   b. Add a column for transcript length (Click on add columns and find the transcript length column, then click on update columns).





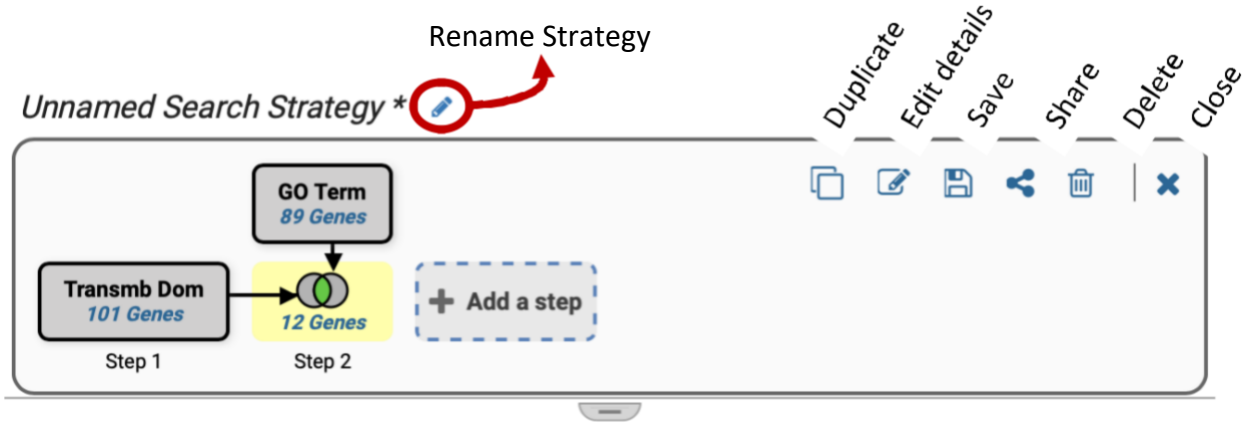
5. Add a step to your strategy. Click on the add step button then find the search for genes with GO Terms. When you find it select it and configure the search to find all genes with the GO term “Transporter activity”. See screen shots below if you need help.



Notice that you have different options on how to combine results from searches in your strategy. What do you each of the operations do?

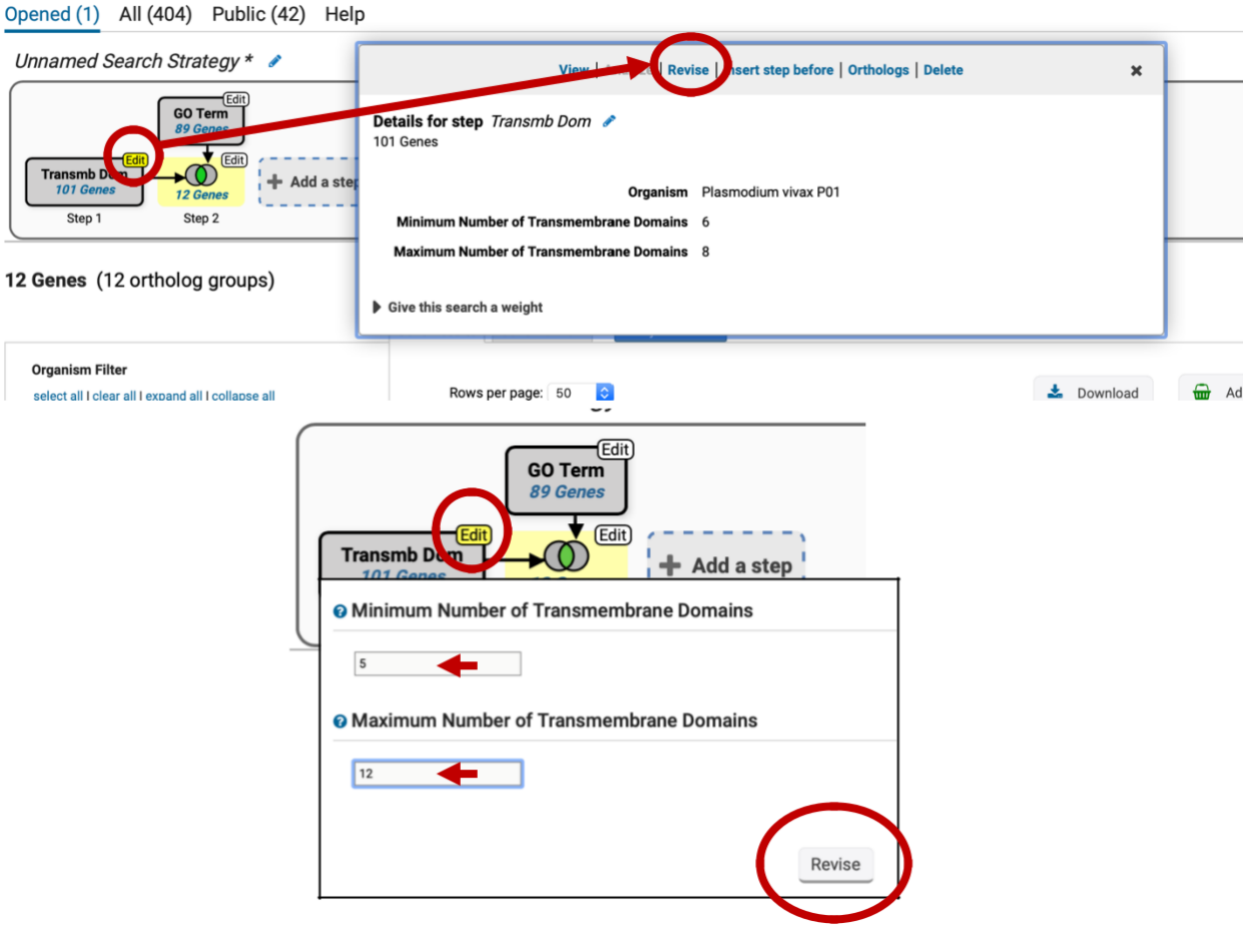
Operator	Combined Result will contain
	IDs in common between the two lists
	IDs from list 2 and list 3
	IDs unique to 2
	IDs unique to 3
	IDs whose features are near each other (collocated) in the genome

6. You can rename, duplicate, delete, save and share strategies (saving and sharing strategies requires creating and account and logging in). You can also rename each



individual step in a strategy.

7. Revising a step in a strategy. You can revise any step in a strategy by moving your mouse over the step you want to revise until you see the edit button appear on the step.
8. Revise the first step in your strategy and change the TM domain parameter to include genes with a minimum of 5 TMs and a maximum of 12 TMs. How does this change your final results?



Advanced Search Strategies

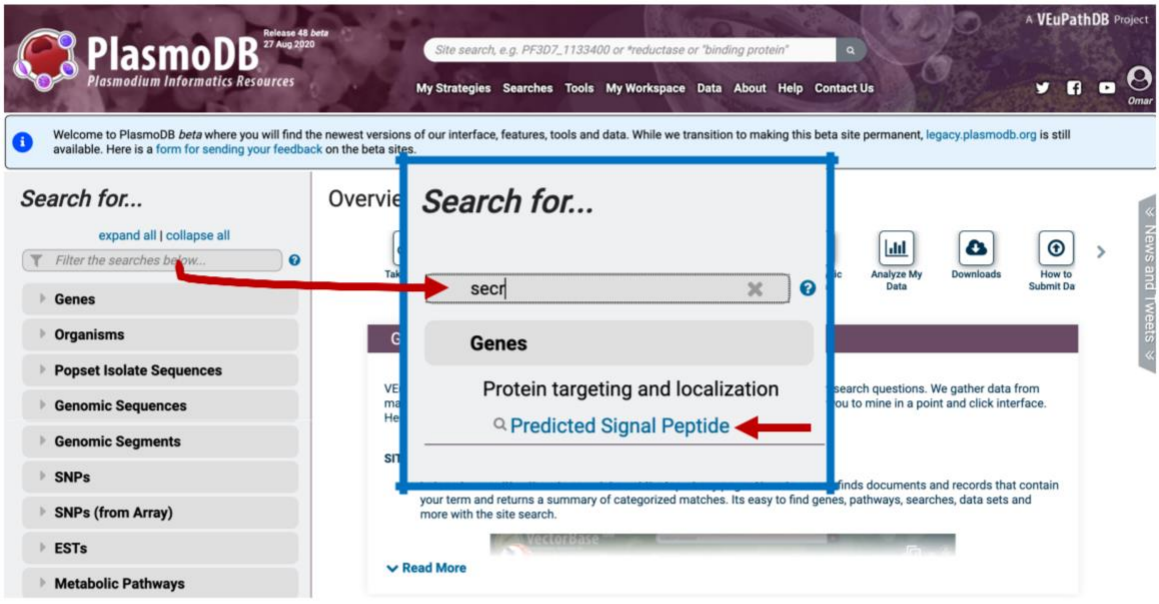
Note: *this exercise uses PlasmoDB.org as an example database, but the same functionality is available on all VEuPathDB resources.*

Learning objectives:

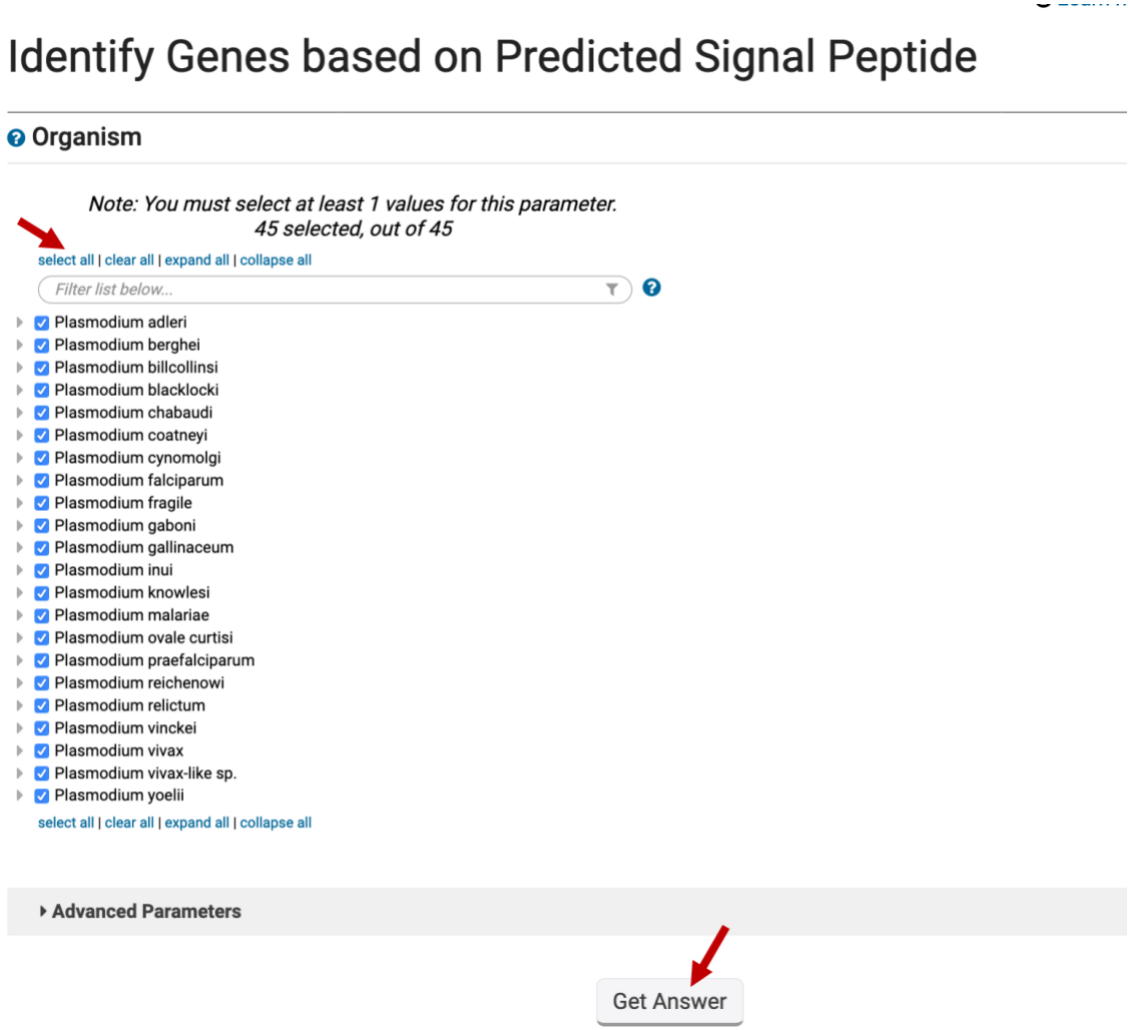
- Integrate diverse datatypes in a search strategy
- Leverage orthology and phylogenetic profile searches

This exercise walks you through the process of building a multi-step strategy, integrating different datatypes. The final search strategy identifies plasmodium genes that are likely secreted, or membrane bound, highly polymorphic, “essential” for parasite survival, not conserved in mammals and expressed in liver stages of the Plasmodium life cycle. There are many ways to build these strategies and order the steps to reach a similar answer.

1. Identify all genes in PlasmoDB that are predicted to have a secretory signal peptide as defined by SignalP. An easy way to identify a search type is to filter the searches on the left of the home page. Start typing a word to identify the search type. For example, start typing the word "secreted”, you should see the searches being filtered even before you finish typing the complete word.

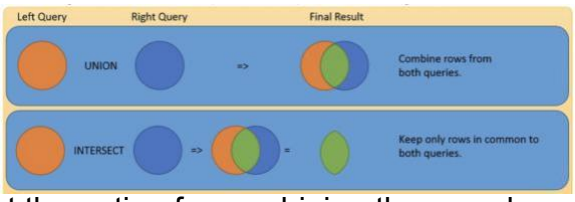


2. Click on the search for genes by predicted signal peptide. On the next page select all organisms and click on the get answer button at the bottom of the page.

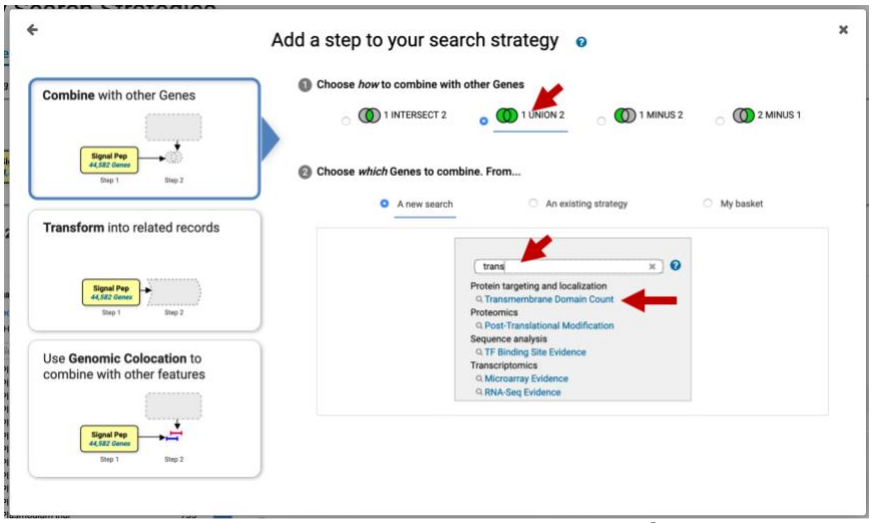


3. The next step is to combine the signal peptide results with results of genes that are predicted to have at least one transmembrane domain (TM). Click on the add step button in the search strategy panel.

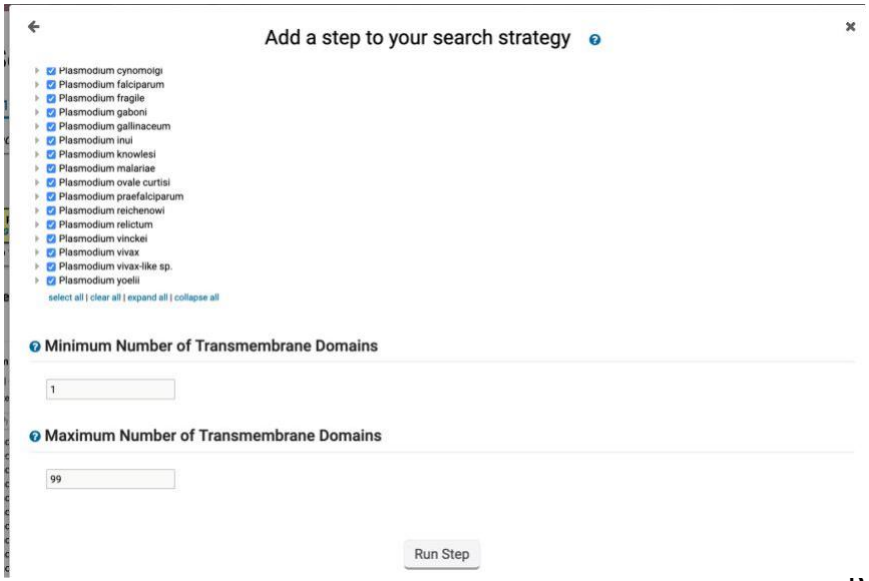
The popup window offers you option to add additional steps and ways to combine the searches (intersect, union, minus). For this exercise we are interested in finding genes that a signal peptide or a TM domain or both. What operation will you use to combine the searches – Union or Intersect?



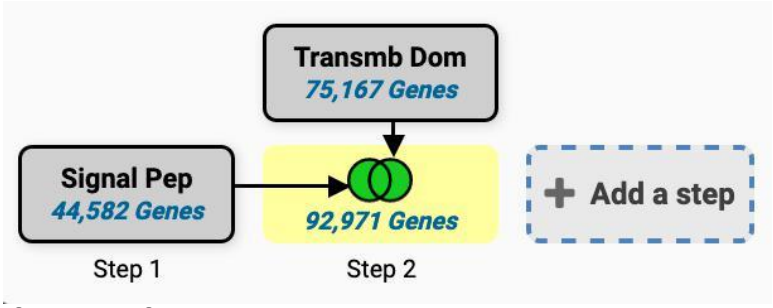
Once you select the option for combining the searches, find the search for transmembrane domain count. Notice that you can use the same query filtering mechanism as before. Start typing transmembrane to find this search. Once you find it click on to open the search parameters.



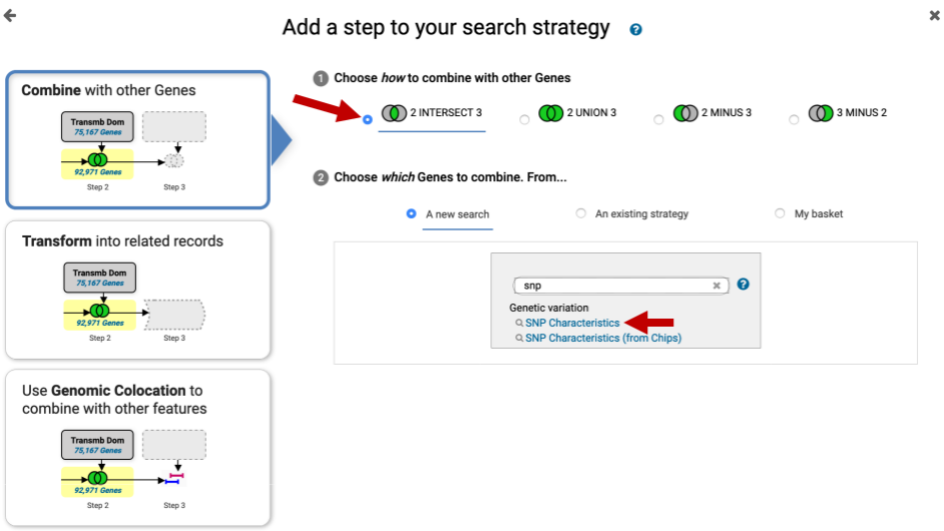
4. For the TM search, again select all organisms, use the default parameters and click on the get answer button.



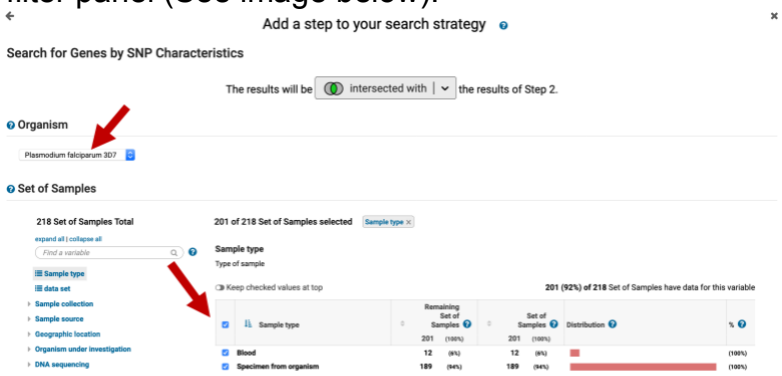
5. How many genes did you get? Since you used a union the number of results should be more that each of the individual steps that were combined.



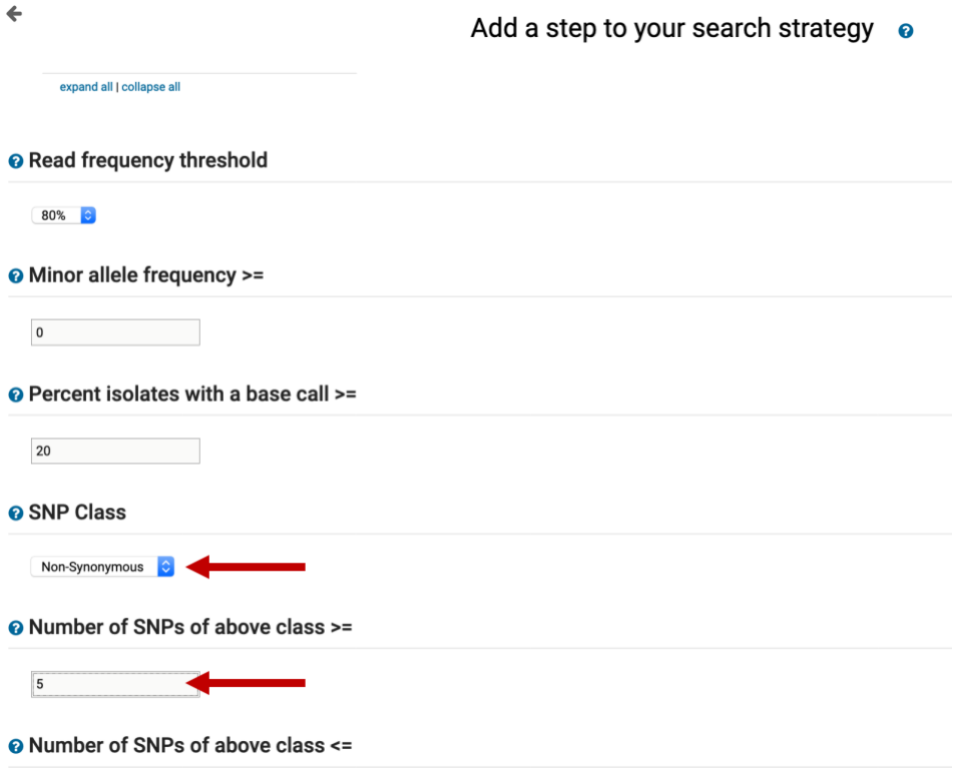
6. Next, identify genes from step 2 that contain at least 5 non-synonymous SNPs (non-synonymous SNPs are single nucleotide polymorphisms that result in an amino acid change). Were you able to find the SNP search by clicking on add step and filtering the searches with a keyword? Which operation will you select to combine the searches?



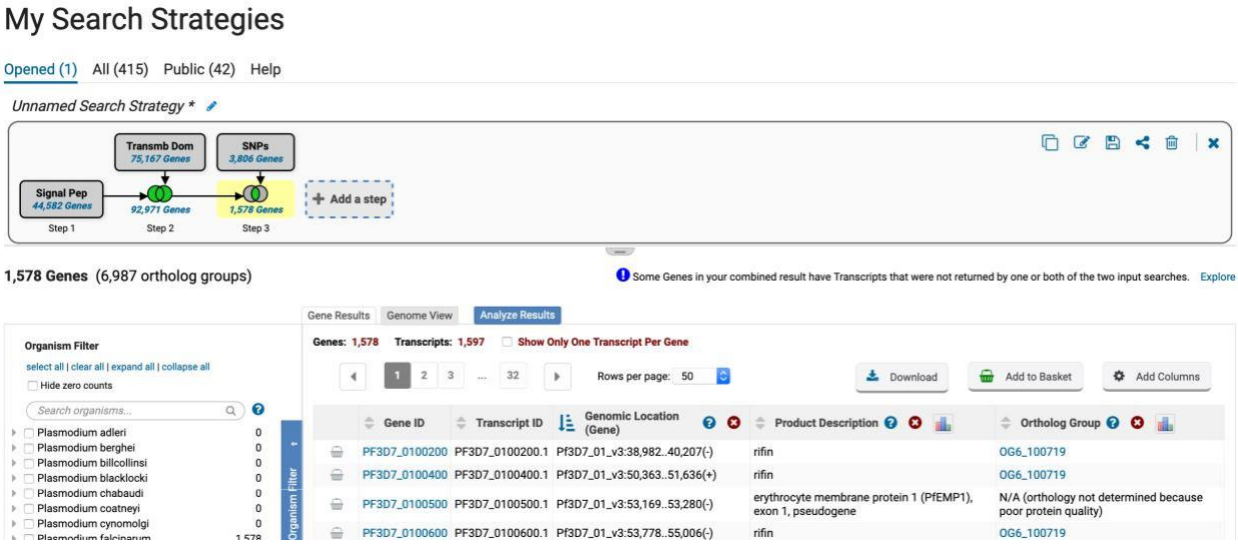
7. On the Genes by SNP characteristics search popup, select Plasmodium falciparum from the drop down and select all available isolates by selecting the checkbox at the top of the filter panel (See image below).



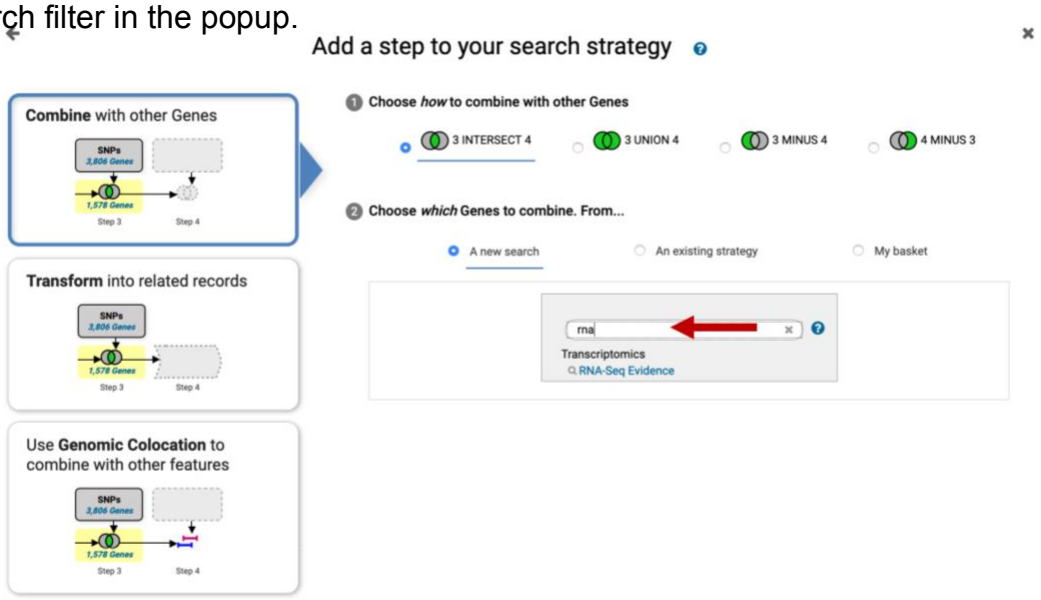
8. Next scroll down and select the following parameters. SNP class = Non-synonymous. Number of SNPs of above class >/= 5. After you select these parameters, scroll down to the bottom and click on Run Step.



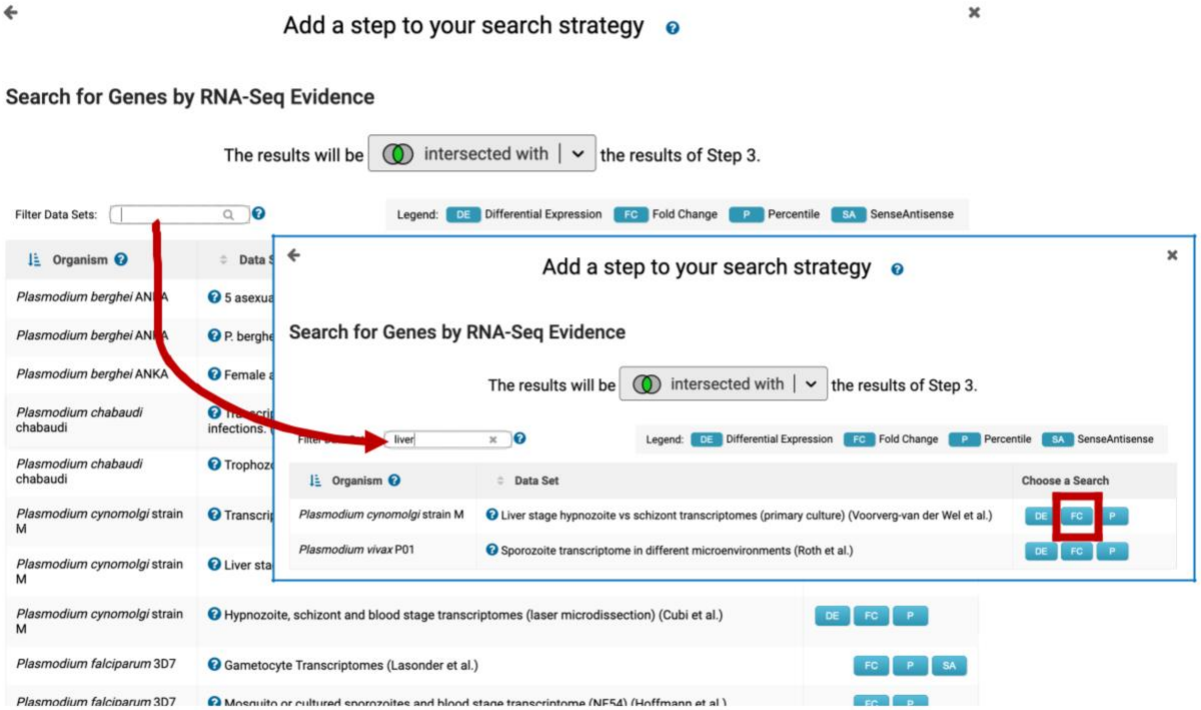
What do the results look like? What species are represented in the results? Is this surprising? Remember that your last search only queried *P. falciparum* data.



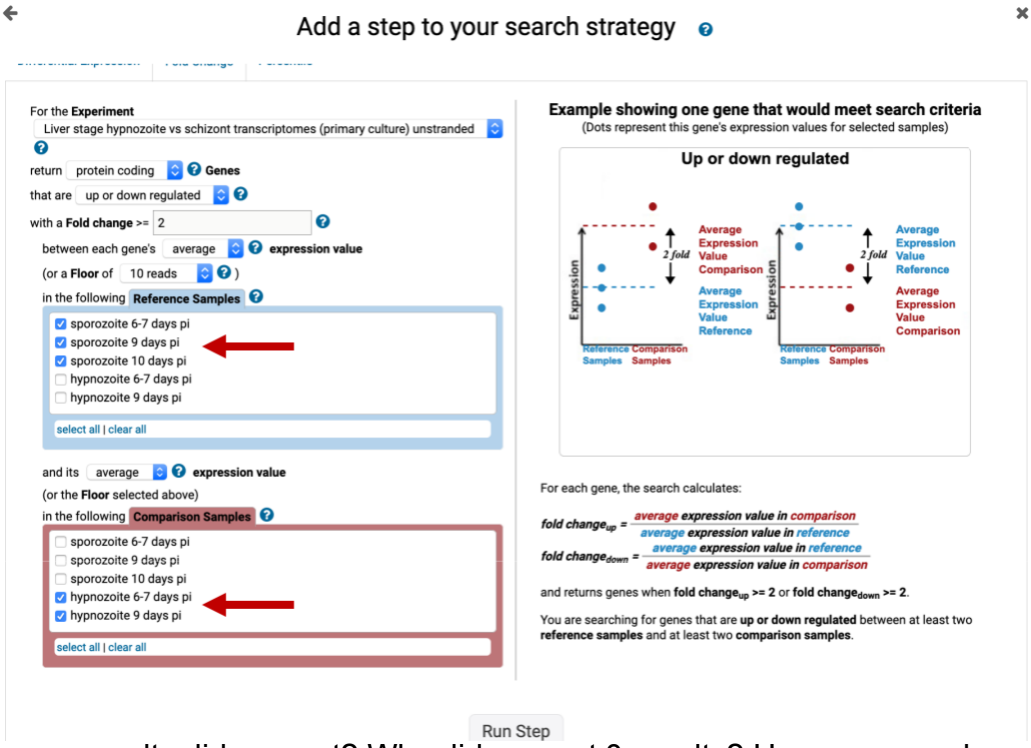
9. Determine how many of these genes are also differentially expressed in liver stages. Click on add step then search for the RNA-seq search. Type RNA in the search filter in the popup.



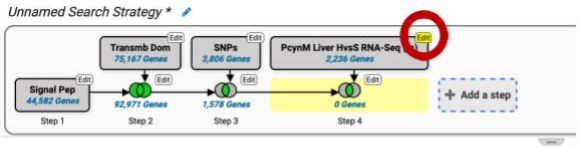
10. On the next page find data that queries liver stages. You can filter the data by typing the word liver in the filter box at the top of the page. This should yield two datasets from *P. cynomolgi* and *P. viviax*. For this exercise, select the fold change query for the *P. cynomolgi* dataset: Liver stage hypnozoite vs schizont transcriptomes (primary culture) (Voorverg-van der Wel et al.).



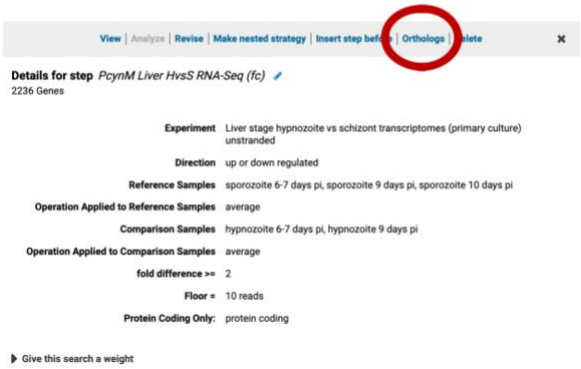
11. Configure the RNA-Seq search to identify genes that are differentially regulated by at least 2-fold between all the hypozoite stages and the sporozoite statges. For example, select the hypnozoite stages in the reference selection box and the sporozoite samples in the comparator selection box, then click on run step.



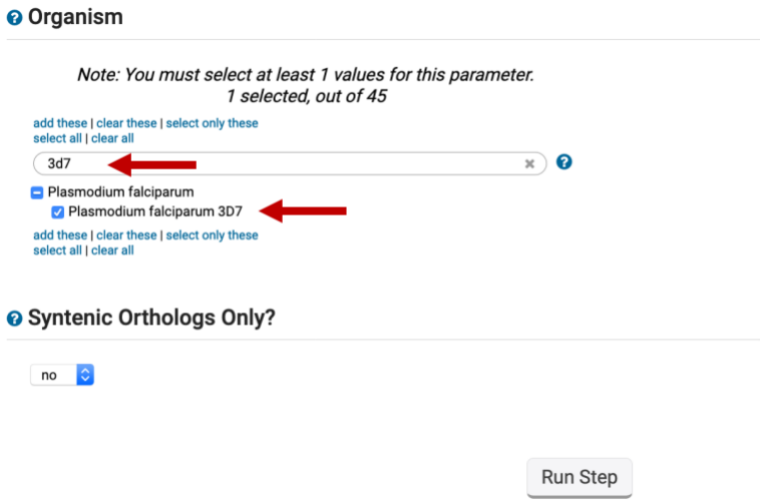
12. How many results did you get? Why did you get 0 results? How can you change this? *Remember that the previous search was a list of P. falciparum genes and this RNA-Seq was from P. cynomolgy. What you would like to do is convert the P. cynomolgy genes into P. falciparum genes.* To do this folloow these steps:
    a. hover your mouse of the RNA-seq step then click on the edit option on that step.



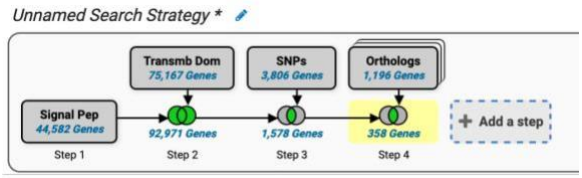
    b. In the popup window, click on the **orthologs** link.



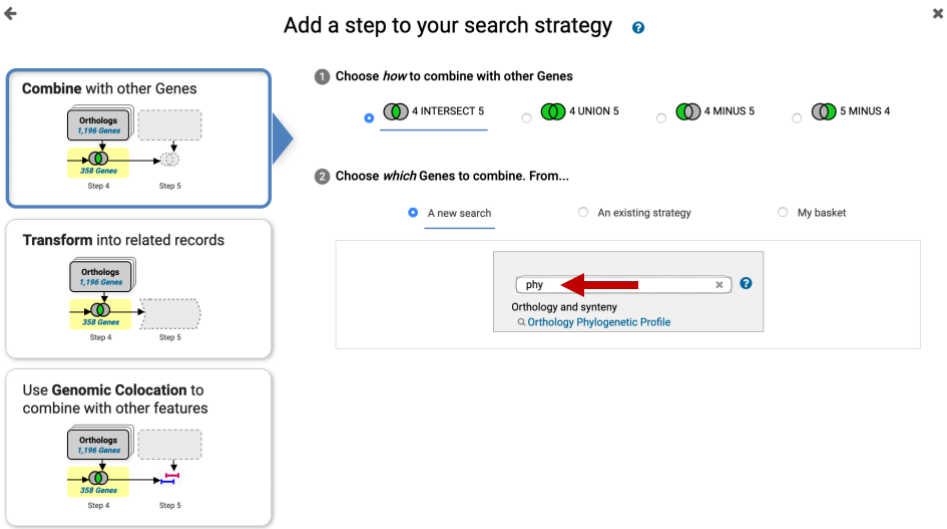
c. In the next window select which organism(s) you would like to transform to. For this exercise select *P. falciparum* 3D7 and click on run step.



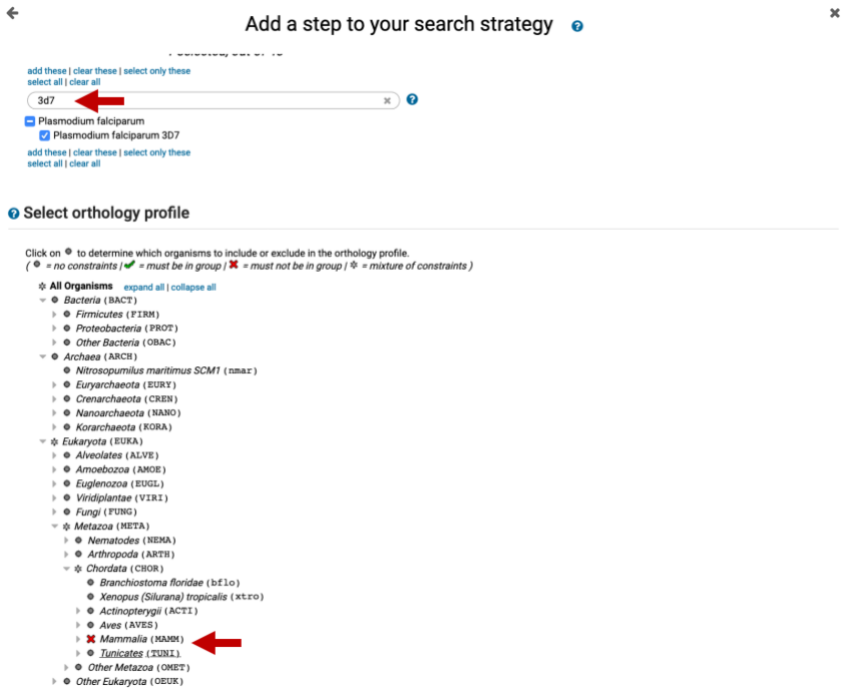
d. Did you get results now?



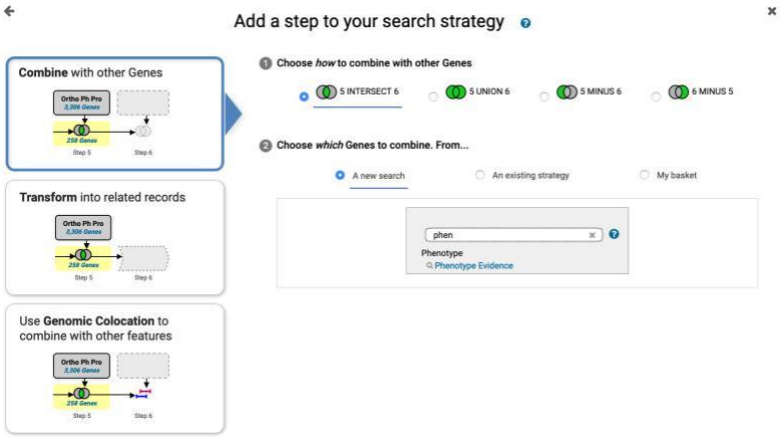
13. Next identify how many of these genes do not have orthologs in mammals. To do this add a step for genes based on orthology phylogenetic profile. Again you can filter the searches by typing the word “phylogenetic”.



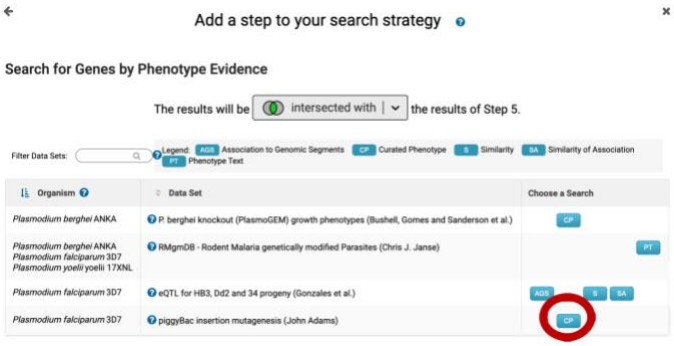
On the next page select *P. falciparum* 3D7 the configure the phylogenetic profile by finding Mammalia under Chordata which are under Metazoa. Click twice on the circle next to Mammalia – it should become a red x (See image below).



14. Determine if a mutation in any of these genes affects fitness. Click on add step and find the search for phenotype evidence.

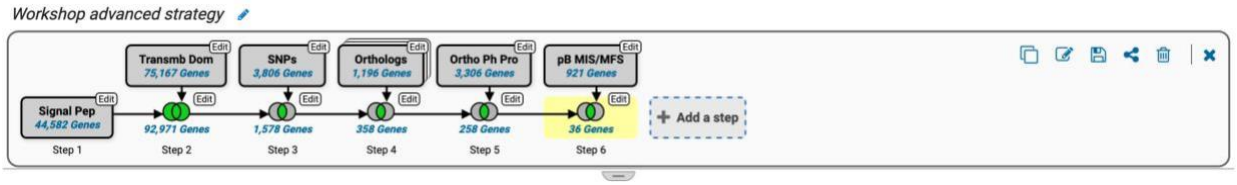


15. Select the *P. falciparum* piggyBac insertion mutagenesis (John Adams) experiment.



16. On the next page select the Mutan Fitness Score (MFS) option and choose any score range – generally the more negative the bigger the effect is on fitness. For this example a score range of -4.078 to -3.07 was chosen

Explore your final results. Do they make sense/plausible? Note that you can revise any of the steps in the strategy to explore the data further. You can also save your strategy and share it with others or make it public. Here is a link to this search stragey:



https://plasmodb.org/plasmo/app/workspace/strategies/import/fd387e8d3acda856

Public Strategies

Users can share their strategies publicly so that others may use them. The public strategies link is located under the **About** menu followed by **Community** and **Public Strategy.** A table of available public strategies will appear and there is a filter box located at the top of the public strategies so that you can search for the author or subject of the strategy among other items. The public strategies for PlasmoDB, as an example, are located at: https://plasmodb.org/plasmo/app/workspace/strategies/public

Exploring the Gene Page

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

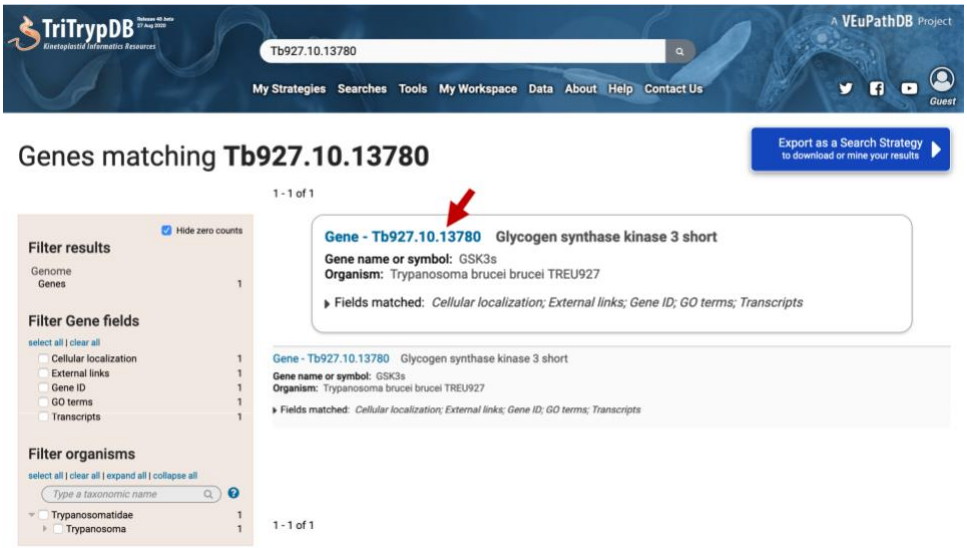
Learning objectives

Gene pages:

Become familiar with the information in gene pages
Navigate to and from the gene pages

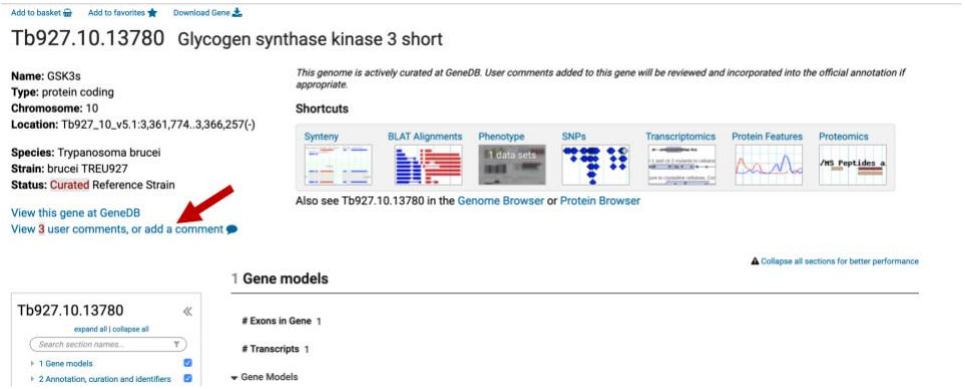
1. **Navigation to the Gene pages**

   For this exercise visit the gene page for Tb927.10.13780 (Glycogen synthase kinase 3). 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.



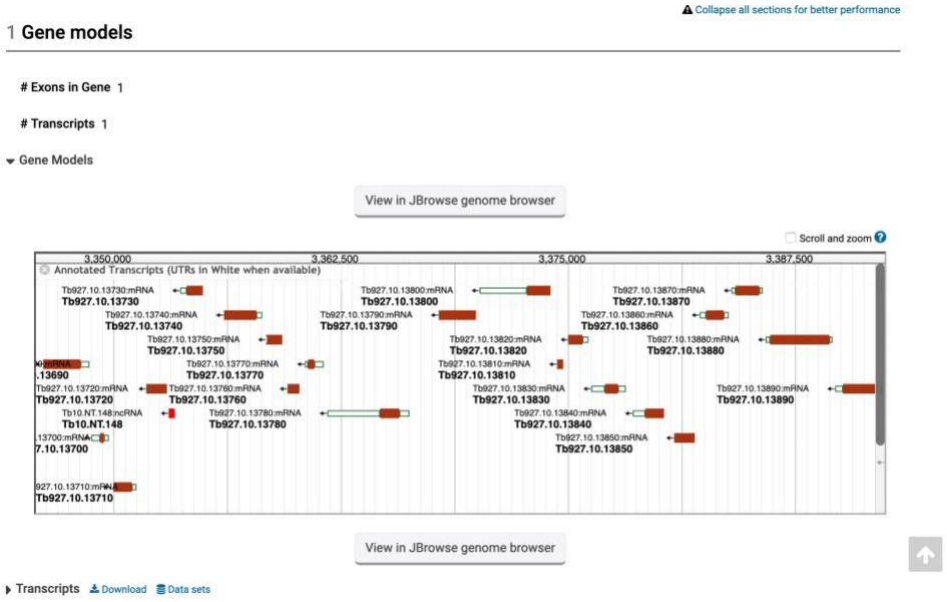
2. **Explore the top section of the gene page**

What information is in this section? Can you easily find which chromosome this gene is located on? Does this gene have any user comments?

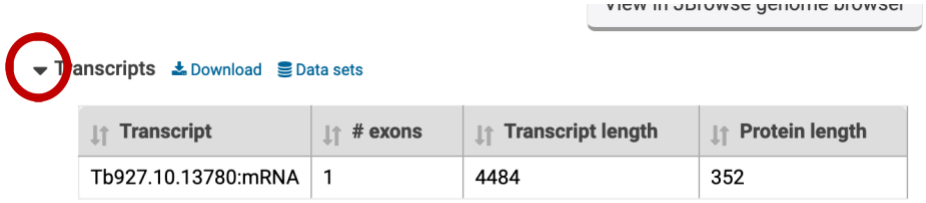


3. Explore the gene model section.

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?



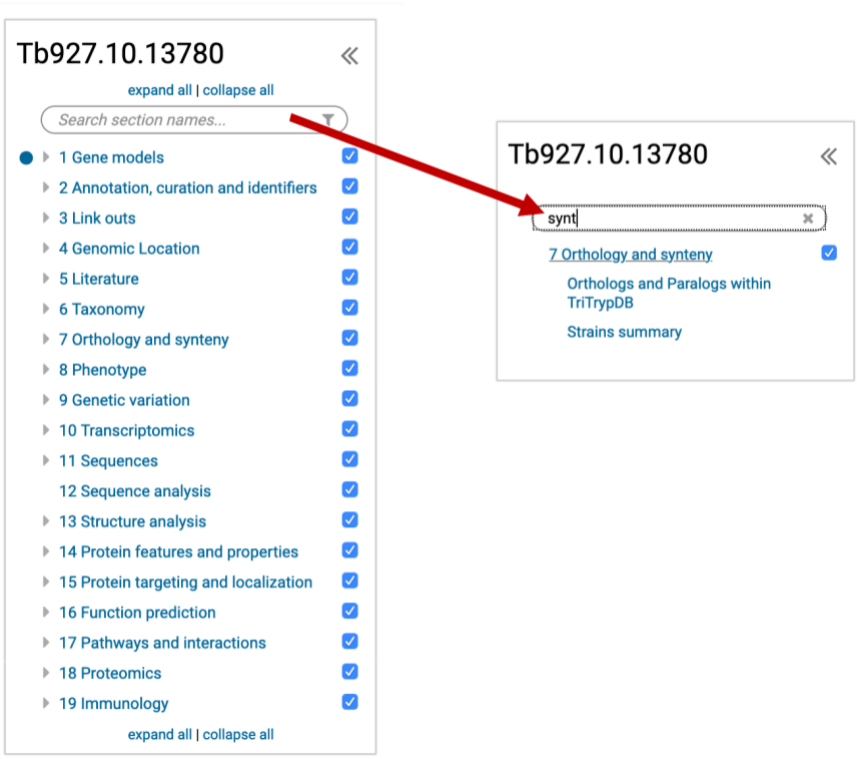
How long is the transcript? You can determine transcript length by expanding the Transcripts section.



4. Content navigation.

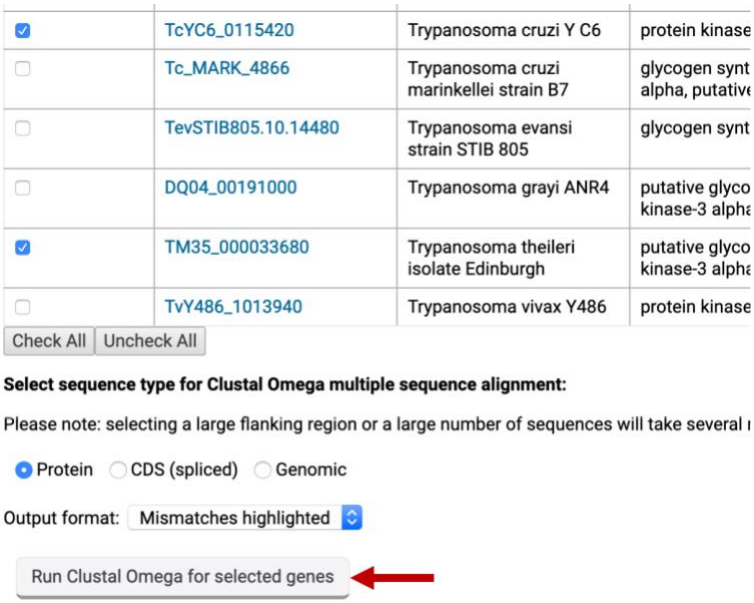
How do you find/navigate to the different sections of the page? Use the “Contents” menu on the left side, type a keyword and click on the menu, click on the work to

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.



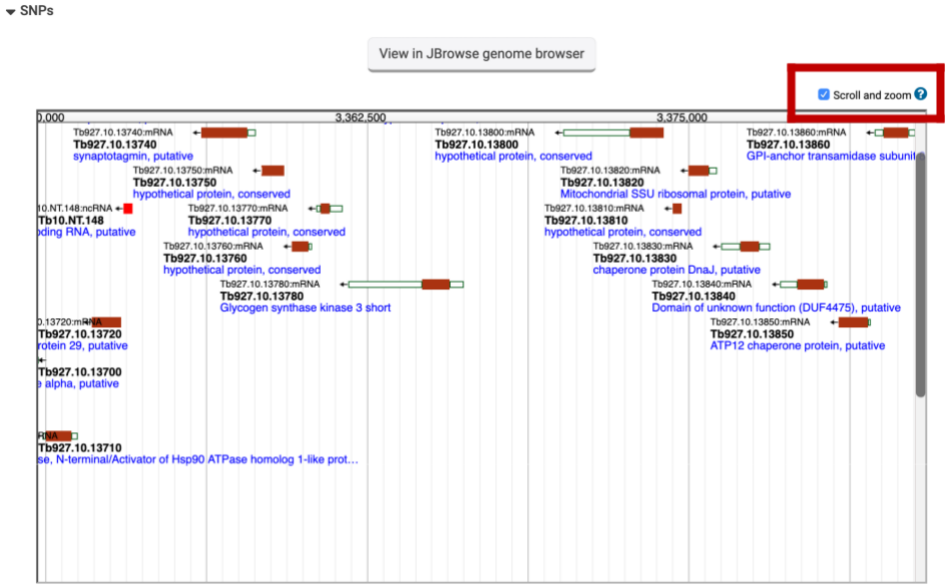
5. Running an alignment of selected sequences

a. Expand the “Orthologs and Paralogs in TriTrypDB” 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.

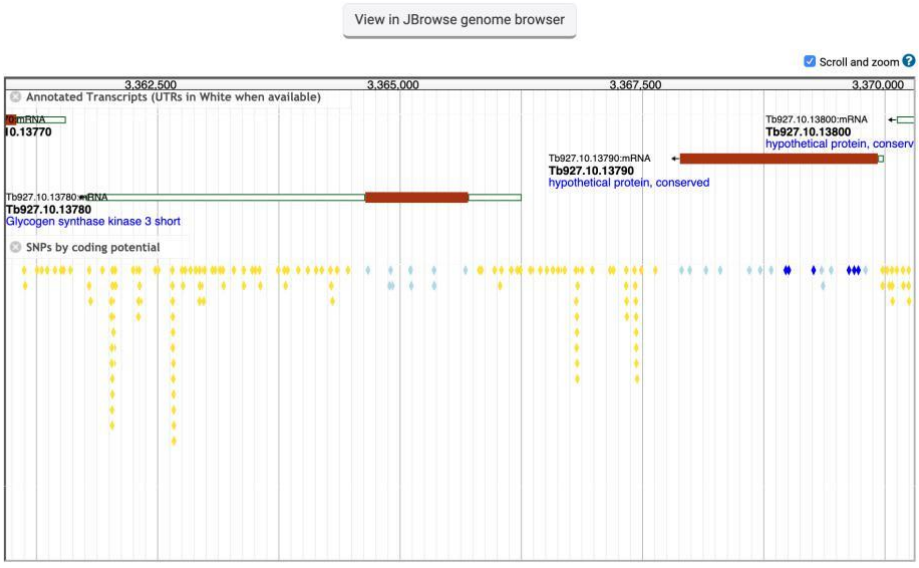


6. Exploring the genetic variation section

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 as an example database, but the same functionality is available on all VEuPathDB resources.*

Learning objectives:

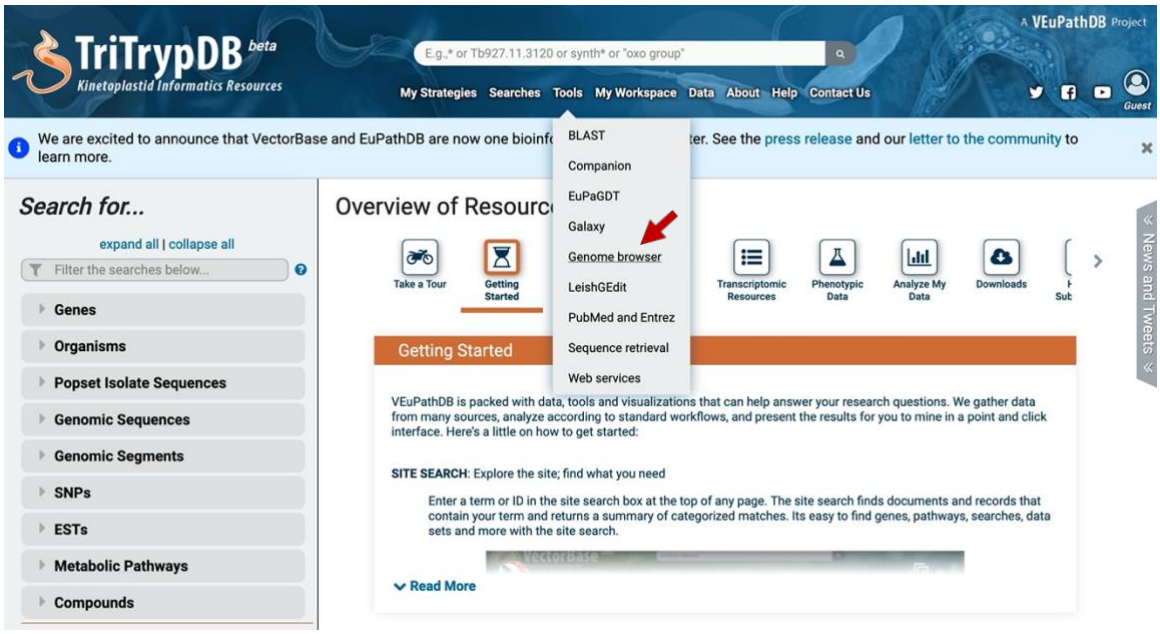
- Navigate to the genome browser
- Use the menu and navigation bars
- Run searches
- Add pre-loaded data tracks
- Upload your own data tracks
- Configure tracks
- Download track data

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/

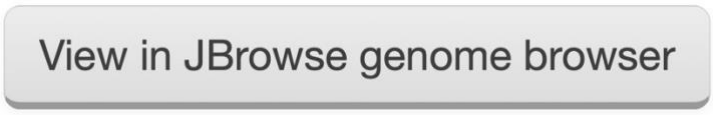
Links to the genome browser are available from multiple locations:

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



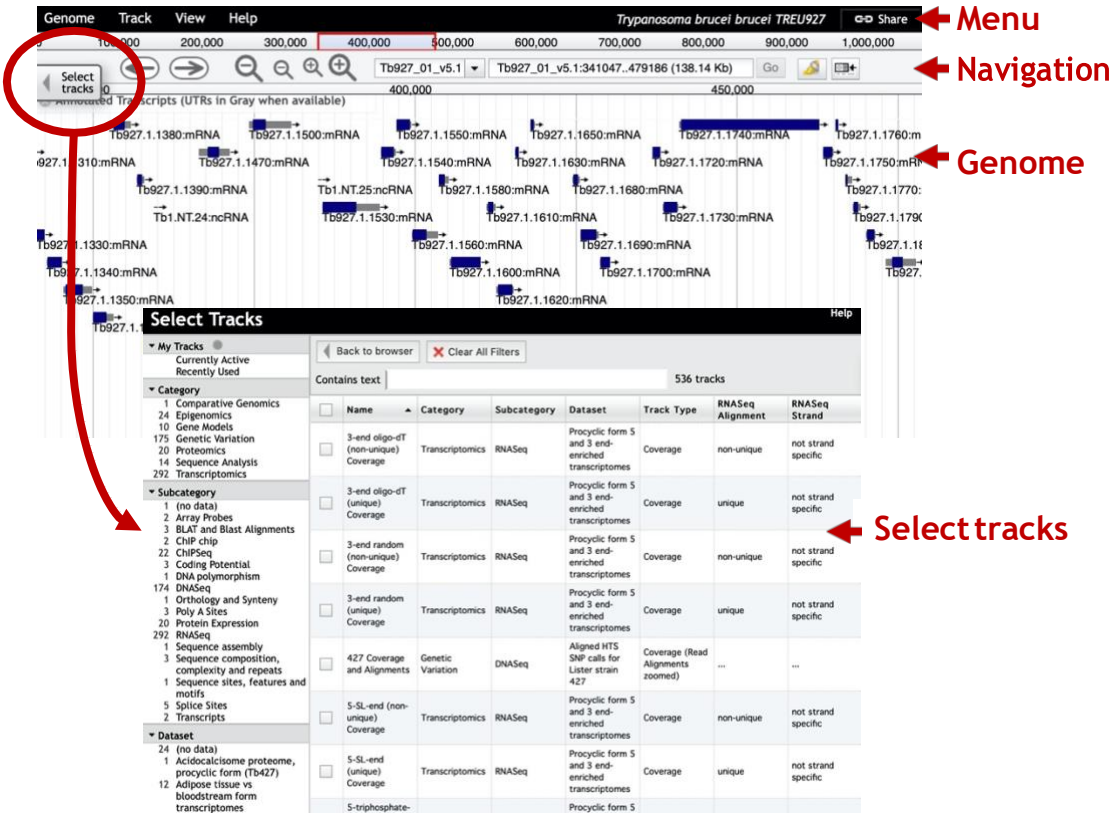
b. From record pages such as gene, SNP or genomic sequence pages – these links are usually to a specific JBrowse configuration that includes data relevant to the section on that record page. For example, a JBrowse link from an RNAseq dataset on the gene page would display the gene of interest

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.

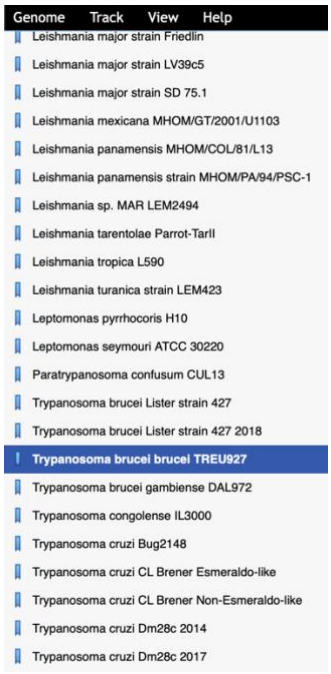
a. Use any of the above described JBrowse linking strategies to get to the genome browser.
b. Once in JBrowse examine the following features:
   i. The **menu bar**: located at the top of the JBrowse frame. This includes the Genome menu, Track menu, View menu, Help menu and the Sharing link. What do 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

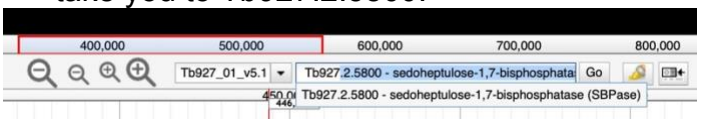
check box next to the track name.

3. Navigating to a specific gene in JBrowse.

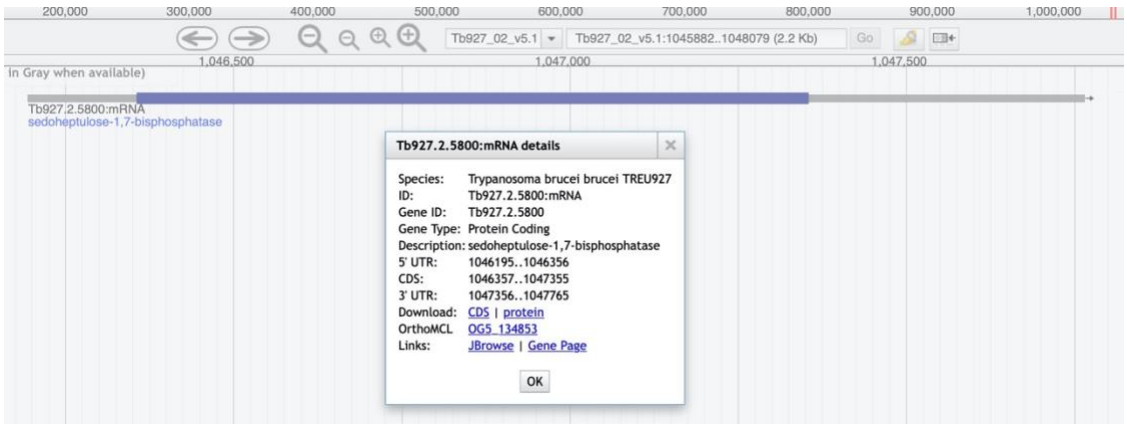


The goal of this step is to navigate to the sedoheptulose-1,7-bisphosphatase (SBPase) gene of *T. brucei* 927.

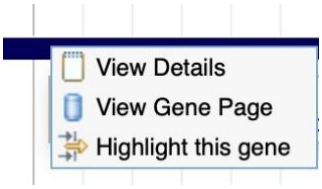
a. Make sure the *Trypanosoma brucei* brucei TREU927 genome is selected from the genome menu.
b. Start typing the word sedoheptulose in the search box. After a few seconds you should see the result of the search (do not hit enter). Select the gene from the search dropdown. This will take you to Tb927.2.5800.



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?

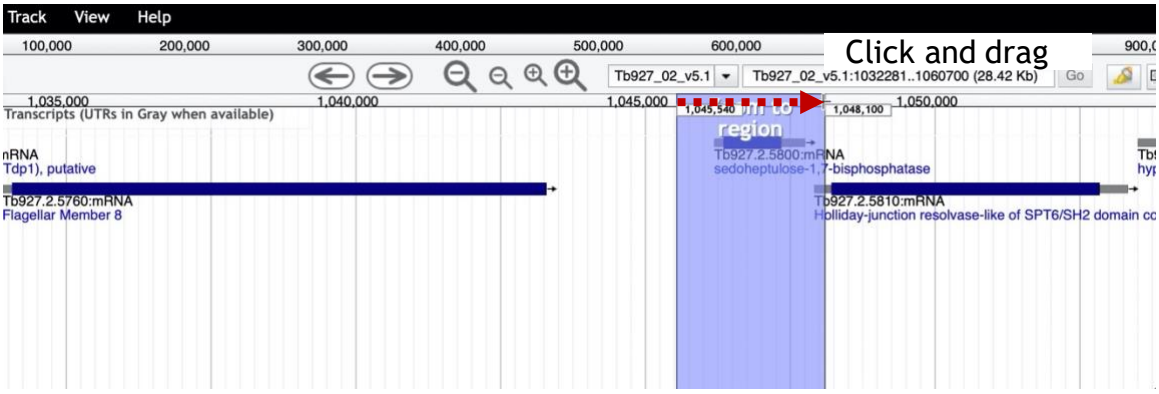


d. You can also right click (or control click) on a feature to display the context menu which provides quick links to highlight a feature, go to the feature page (like the gene page) or get the info popup (the same one you get when you click on the feature).



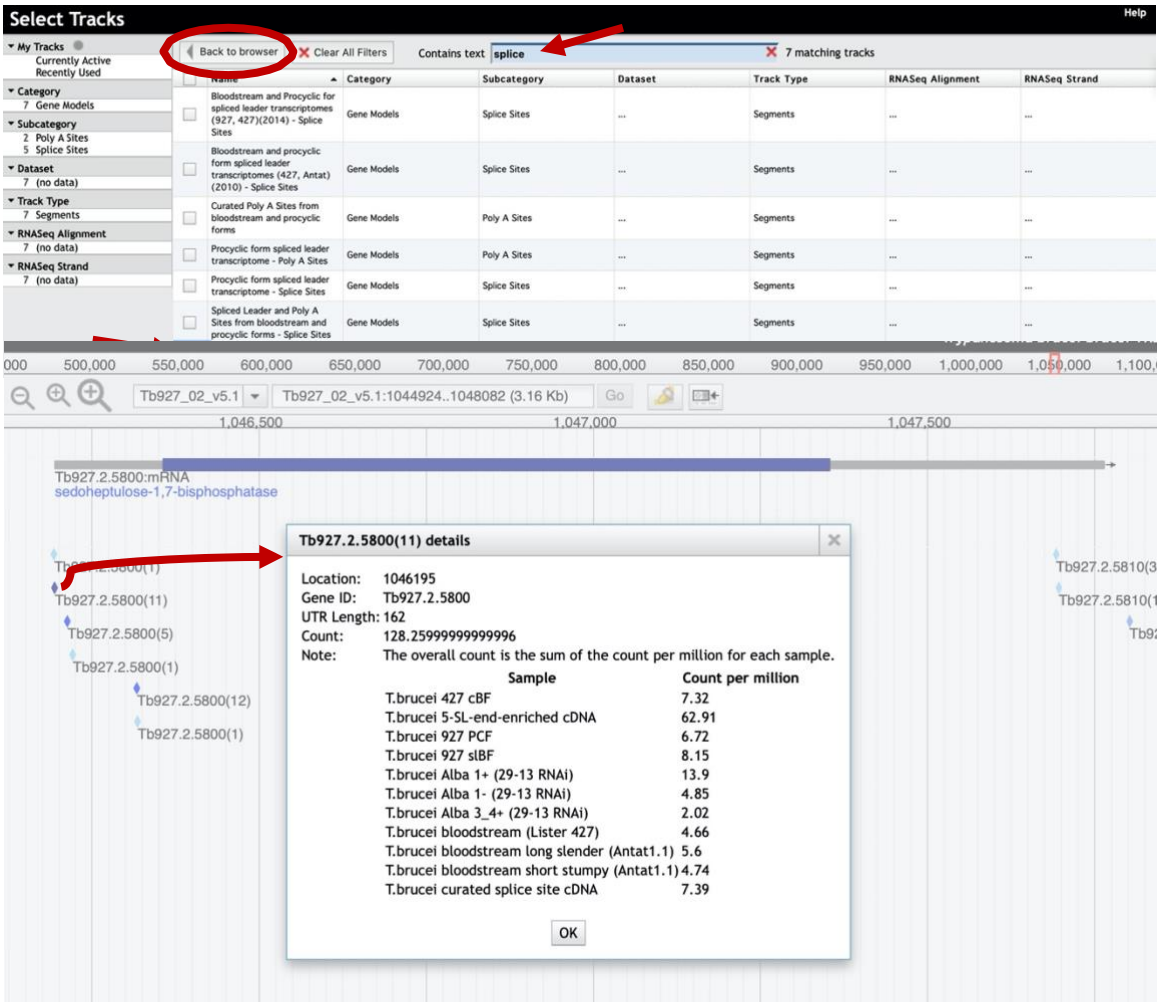
e. What genes are immediately upstream and downstream of SBP? (Hint: use the zoom out button in the navigation bar). What is the difference between the small and large zoom buttons? (*Tip1:* another way to zoom in and out is by clicking on shift and the up or down arrows. What happens if you click shift

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).

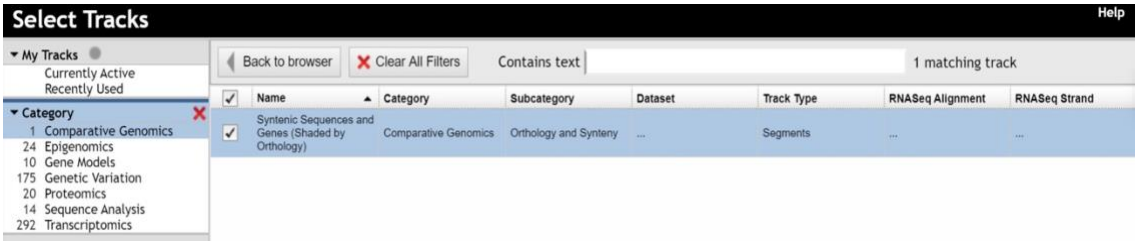


What additional data track would be useful for you to assess this? (hint: Click on the “Select Tracks” button to reveal all available tracks. Now type the word “splice” in the “contains text” box. This will filter all tracks that contain the word splice. Find the one called “Unified Splice Leader Addition 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.

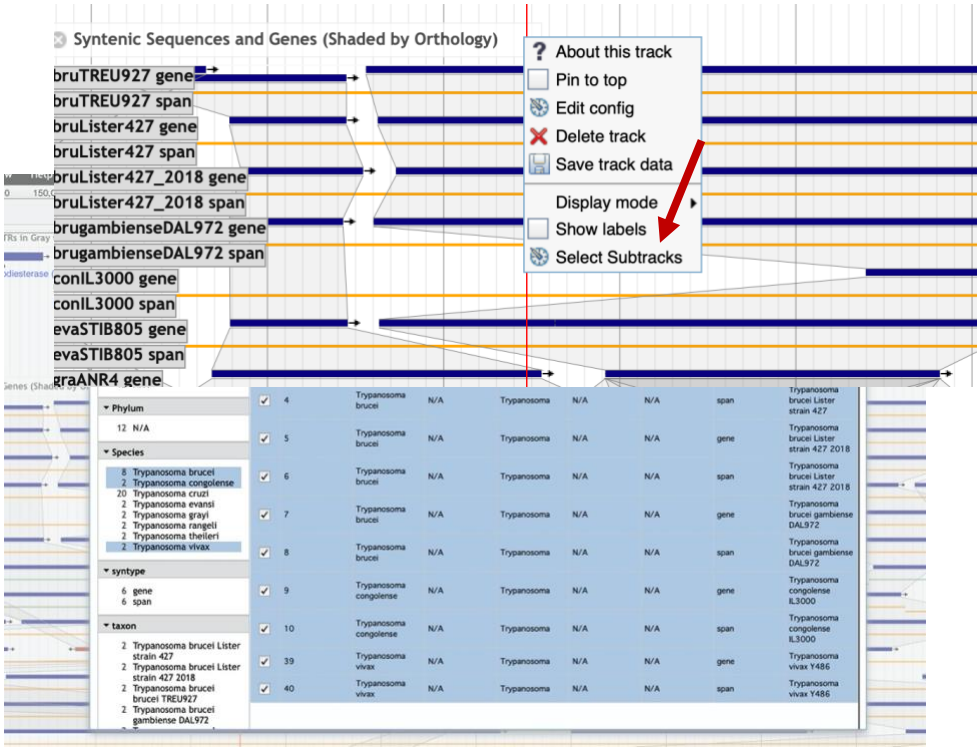
Synteny helps define conservation of homologous genes and gene order between genomes.

- 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



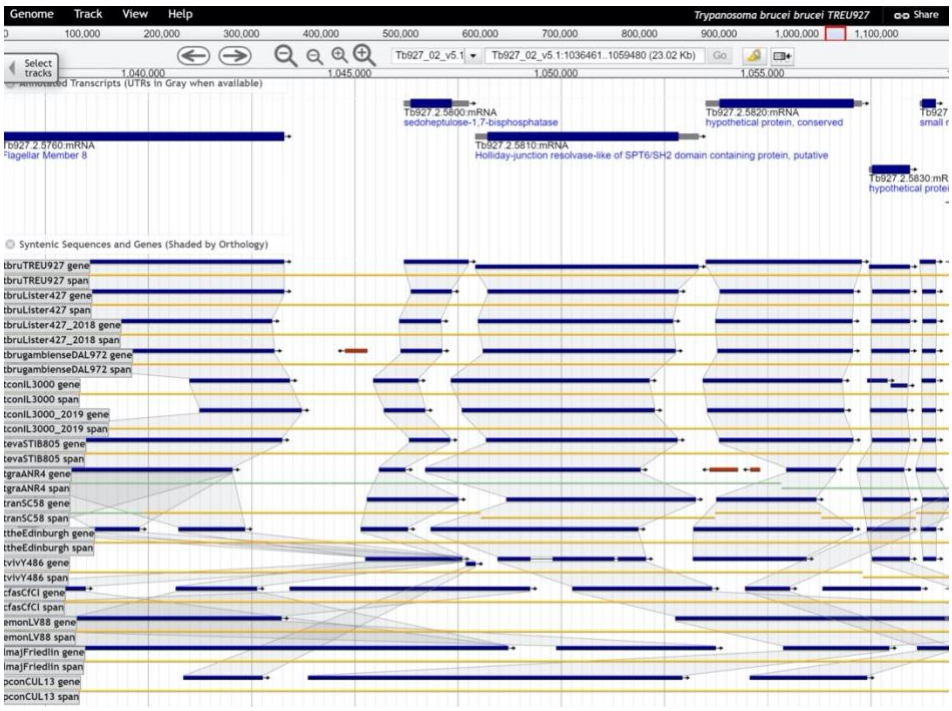
tracks.
- Return to the browser by clicking “Back to Browser” and zoom out so you can see a couple of genes on either side of SBP (does not have to be exact)
- Configure the synteny track to include the following species subtracks: *Trypanosoma brucei 927, T. brucei 427, T. brucei gambiense, T. congolense,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

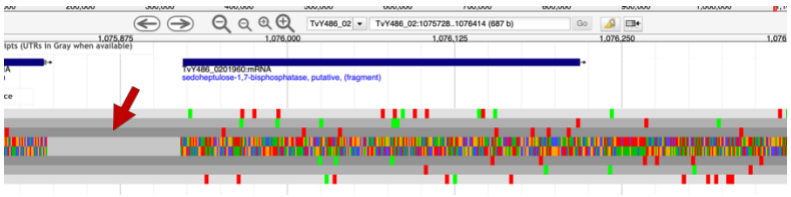


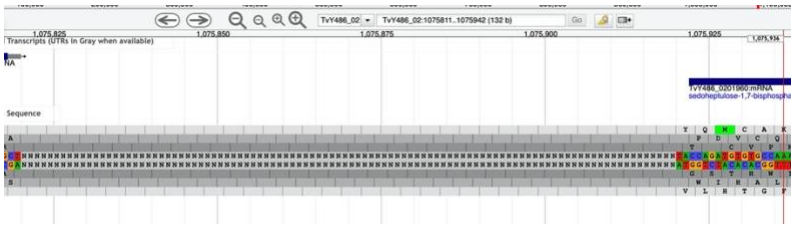
- In the next popup first uncheck all organisms, second use the filters on the left to select Trypanosoma, third select the species of interest (note that you should select both the gene and span subtracks for each species), fourth click on the save button at the bottom of the popup.

- 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?

6. Exploring other data tracks in Jbrowse.

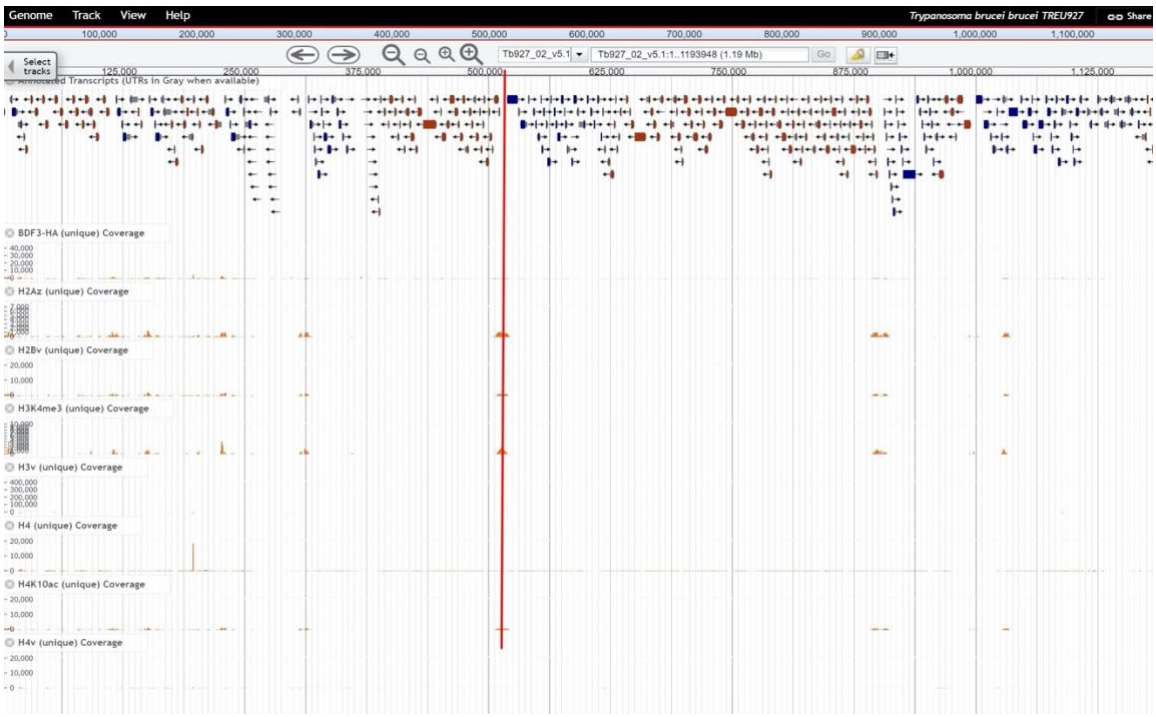
For this example, we will view *T. brucei* data, so the data tracks you turn on will display data only if the data is aligned to the *T. brucei* genome. Return to the SBPase gene in *T. brucei* by searching for the gene ID in the (Tb927.2.5800) in 'Landmark or Region' to redirect the browser. Then zoom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

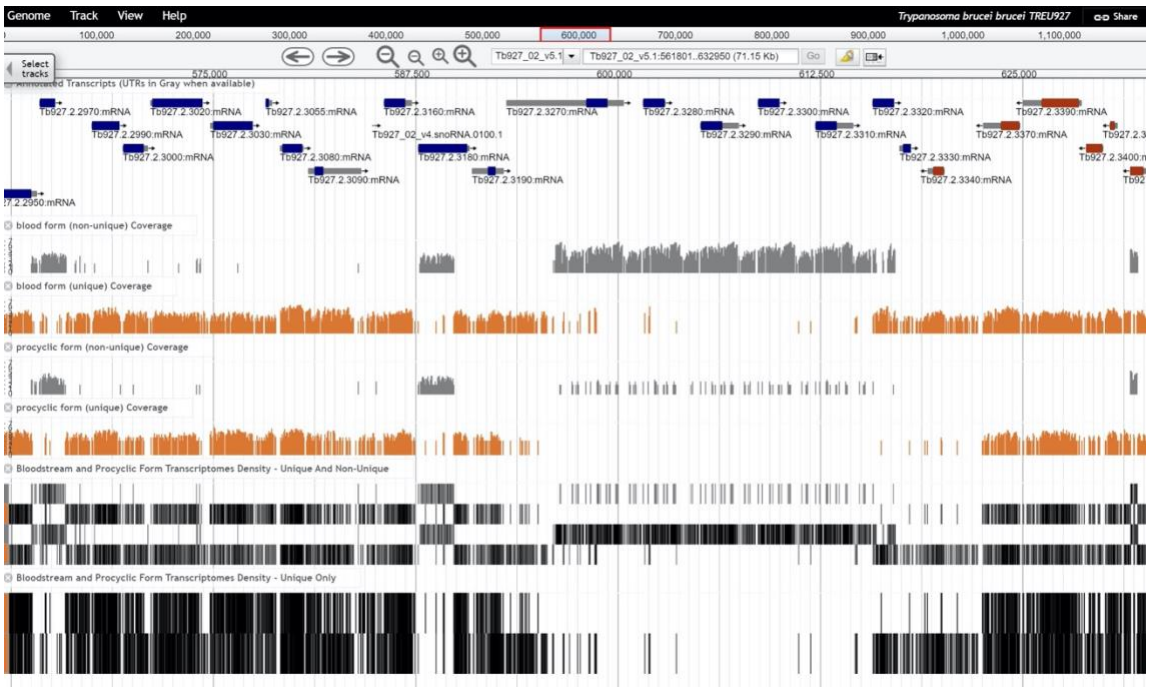
- You may need to adjust the y-axis scaling to bring the tracks into proper view (try

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.

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

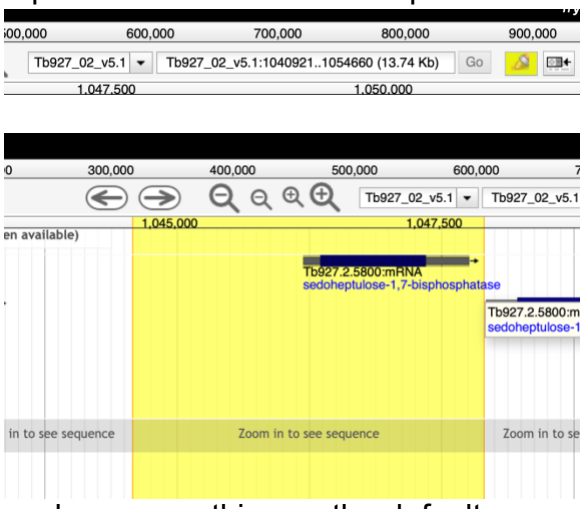
a. Downloading sequence in FASTA format from a region of interest:

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.

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

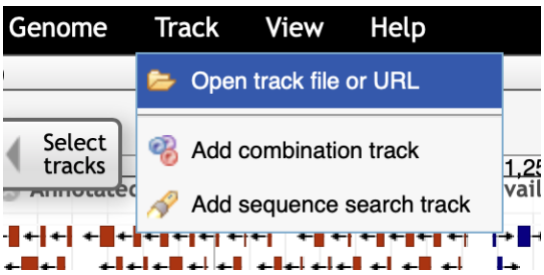
ii. Scroll down to the bottom of the page and download the bigwig file with the http link.

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

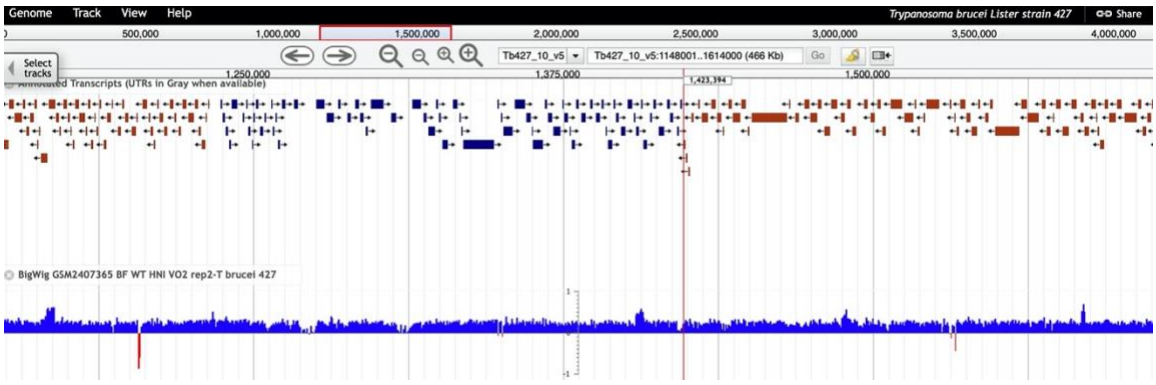
iii. Once the file is downloaded go to JBrowse and select *Trypanosoma brucei brucei* Lister 427 as the reference genome (hint: use the Genome link in the menu panel, top left).

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.



Interpreting RNA-seq data (Browser Exercise II)

In previous exercises, you spent some time learning about gene pages and examining genes in the context of the JBrowse genome browser. It is important to recognize that gene models (structural annotation) are often open to interpretation, however, especially with respect to:

- transcript initiation and termination sites (5' and 3' untranslated regions, or UTRs)
- alternative processing events … if you sequence deep enough, virtually *all* genes (in organisms that process transcripts) display alternative splicing, even for single exon genes
- the potential significance of non-coding RNAs

Even heavily curated genomes (*Plasmodium falciparum, Trypanosoma brucei, Saccharomyces cerevisiae*) do not fully reflect all available knowledge about stage-specific splicing, as new information is emerging all the time! In addition, many gene models were computationally derived using methods that may have not replied on experimental evidence supporting intron/exon boundaries (e.g. RNAseq data).

In this exercise, we will explore genome browser track configuration options in greater detail, focusing on the interpretation of RNA-seq datasets, and using this information to examine the differentially-spliced HXGPRT gene of T. gondii. You will then apply your newfound skills to examine other genes that may be alternatively spliced … and report your findings back to the group as a whole.

The screen shot below (Fig. 1) shows a sample of data tracks that can be turned on and configured in JBrowse. There are a number of tracks that are worth examining which help in determining the accuracy of annotated gene models and that help in defining possible alternative splice variants of a gene. The link below will display the JBrowse view from figure 1, except for any special configurations with are not stored in the URL. For example, tracks 1c and 1d are collapsed in figure one but will appear expanded in the JBrowse view after clicking on the link:

https://tinyurl.com/y5lhcv3f

- What evidence do each of the tracks provide?
- Are the ChiP-ChIP and Chip-seq tracks similar in what they show?
- How many alternative splice variants of HXGPRT would you be willing to annotate based on the evidence?

Are there other data tracks that might be useful to examine?

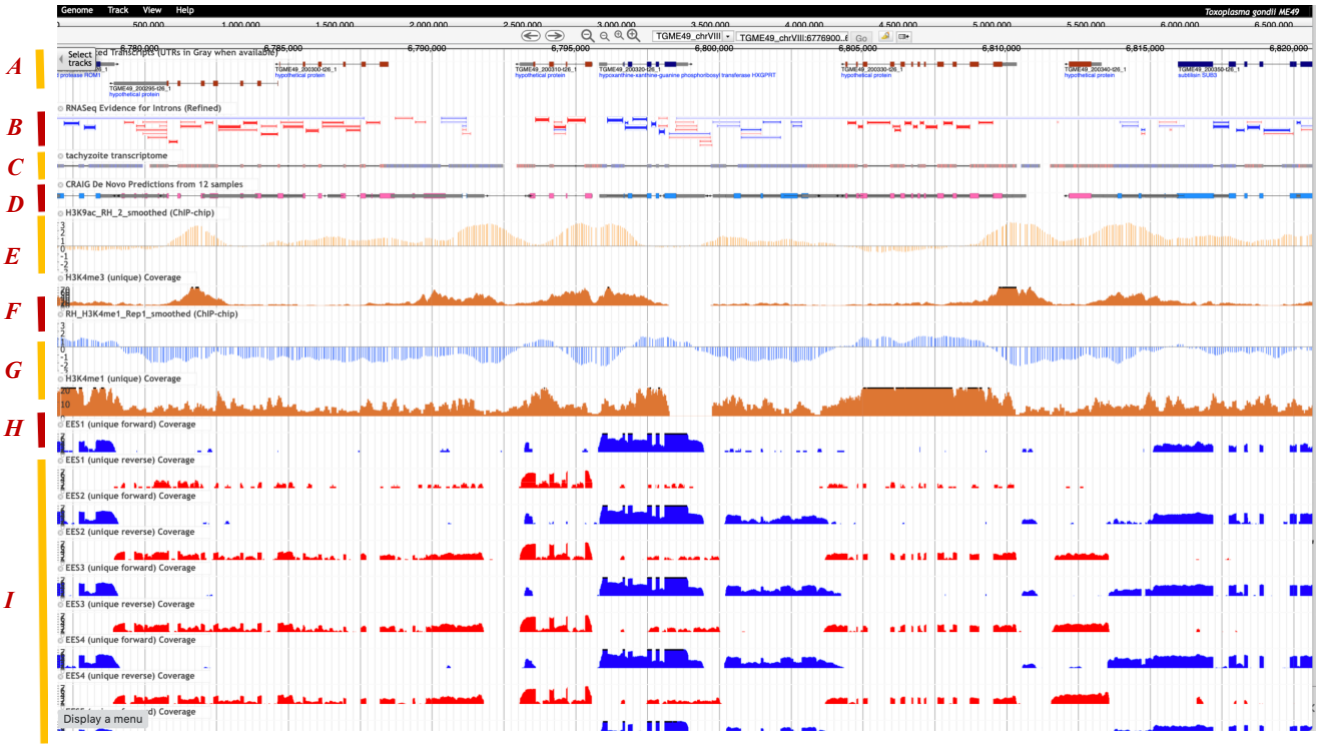


Figure 1: Screen shot from ToxoDB JBrowse. A. Official gene models. B. Splice junction evidence based on available RNAseq data. C. Nanopore long-read transcriptomic data (collapse view). D. Alternative gene models using RNAseq evidence from 12 experiments (collapsed view). E. Chip-ChIP H3K9ac. F. Chip-Seq H3K4me3. G. Chip-ChIP H3K4me1. H. Chip-Seq H3K4me1.
I. RNAseq coverage from Toxoplasma gondii strain CZ clone H3 in feline enteroepithelial stage (strand specific).

Working in groups of four, please select at least two genes from this list to evaluate, based on RNA-seq data and any other available evidence. See if you can discover which exon(s) were represented … and determine whether these genes are actually alternatively spliced (constitutively or stage-specifically). We will then reconvene to hear a brief report from each group.

TgME49_200320 (HXGPRT)	TGME49_211420	TGME49_281440
TGME49_246490	TGME49_214440	TGME49_279390
TGME49_256650	TGME49_250115	TGME49_202770
TGME49_283540	TGME49_261720	TGME49_217490
TGME49_226410	TGME49_268610	TGME49_292150
TGME49_225730	TGME49_270520	TGME49_276170
TGME49_213610	TGME49_280380	TGME49_266640
TGME49_213660	TGME49_293720	TGME49_266920
TGME49_297160	TGME49_248445	TGME49_299010
TGME49_211250	TGME49_230180	

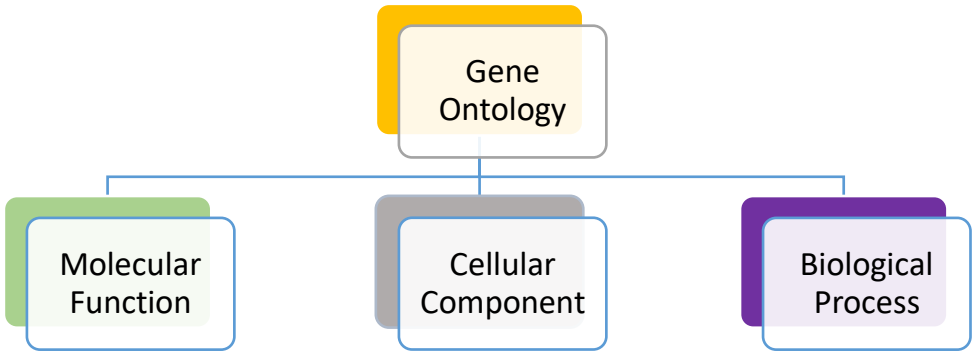
Gene Ontology (GO) Enrichment

Learning objectives:

- Run a GO enrichment analysis
- Explore GO enrichment results

Background:

The gene ontology describes the knowledge of biological sciences and divides this knowledge into three broad categories: Molecular function, cellular component and biological process.



Activities at the molecular level performed by gene products, e.g. Toxin activity, catalytic activity of transporter activity

Where a gene product performs its function, e.g. Cilium Mitochondrion, plastid, Golgi etc…

Processes accomplished by multiple activities, e.g. pyrimidine biosynthesis

To learn more about Gene Ontology, please visit:

http://geneontology.org/docs/ontology-documentation/

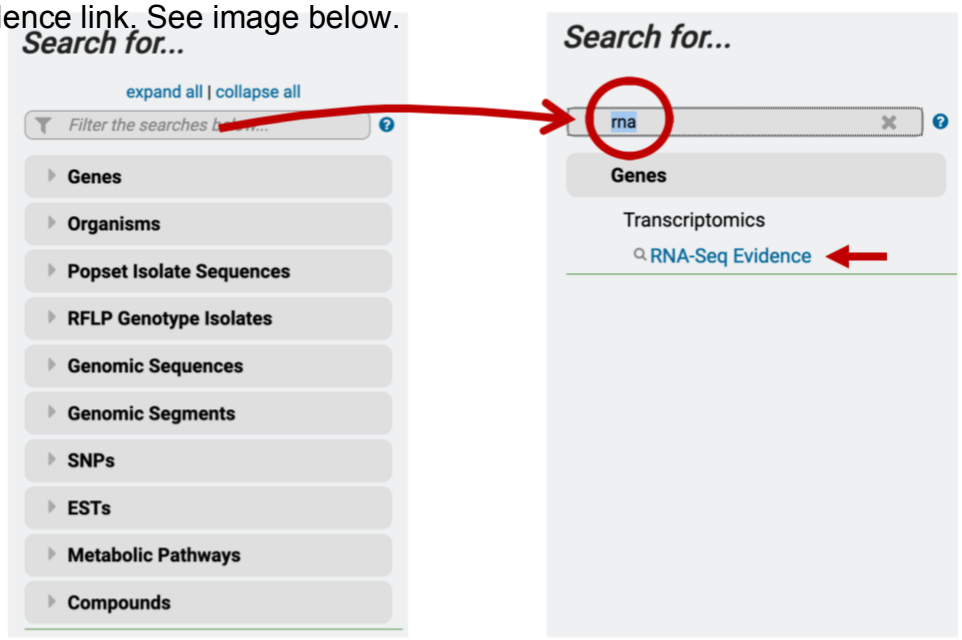
A gene can be assigned a GO term either manually (by an annotator evaluating experimental evidence) or computationally (based on the GO terms of genes that share sequence or functional domains). These GO terms can be used to test whether your set of genes are enriched for a molecular function, cellular component, or biological process.

For example: Does my list of genes have an over-representation of specific GO terms compared to the rest of the genome?

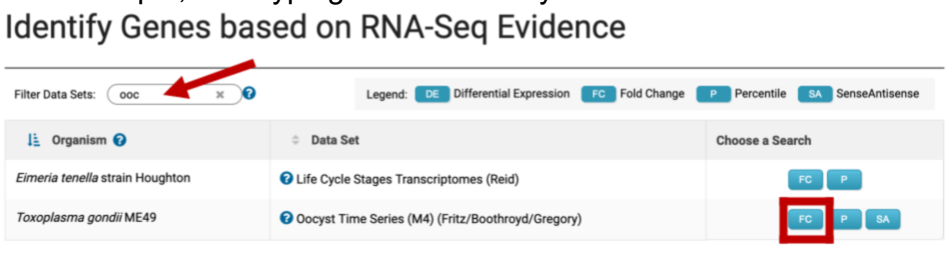
A standard enrichment method employs Fisher's exact test, a statistical test that evaluates a 2x2 contingency table (in this case, number of genes in my set *versus* number of genes from genome not in my set, and number of genes with GO term Z *versus* number of genes without term Z). This test produces a p-value between 0 and 1, where p £ 0.05 is considered significant (that is, less than 5% probability that the enrichment is due to chance). However, the test is performed for each of the 100s of GO terms, increasing the chances that a GO term will be incorrectly considered enriched (a false positive, or type I, error). Thus, the original p-value must be adjusted for so-called multiple hypothesis testing, resulting in an adjusted p-value such as the Benjamini-Hochberg false discovery rate (FDR) or Bonferroni adjusted p-value.

1. In order to run a GO enrichment analysis, you need a list of genes to test. This can be a list of gene IDs from your experimental results that you can upload using the ID search or a gene list resulting from a search you conducted on a VEuPathDB website. For this example, in ToxoDB, we will identify genes that are differentially regulated over time.

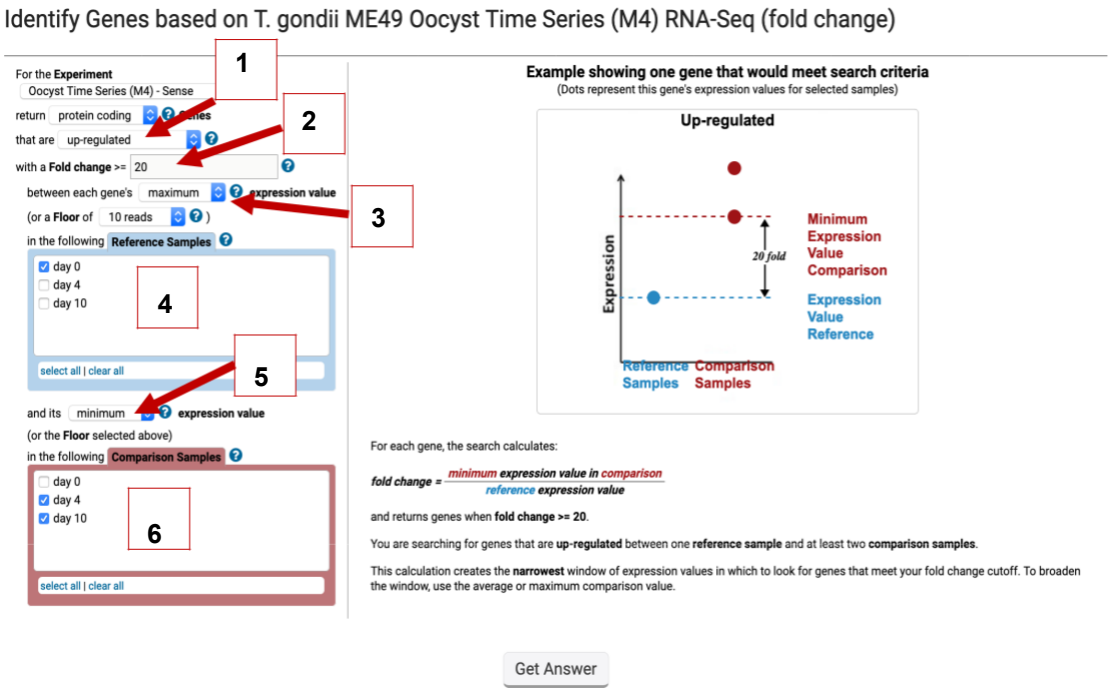
   a. Navigate to the RNA-Seq searches and find the data set called "**Oocyst Time Series (M4)"** from Fritz *et al.* A quick way of getting to the RNA-Seq searches is to type 'rna' in the filter box on the left of the home page and click on the RNA-Seq Evidence link. See image below.



   b. The RNA-Seq evidence page includes a list of all data sets that are loaded in the website. To quickly find a dataset, you can start typing key words in the "Filter Data Sets" box. For example, start typing the word "oocyst".



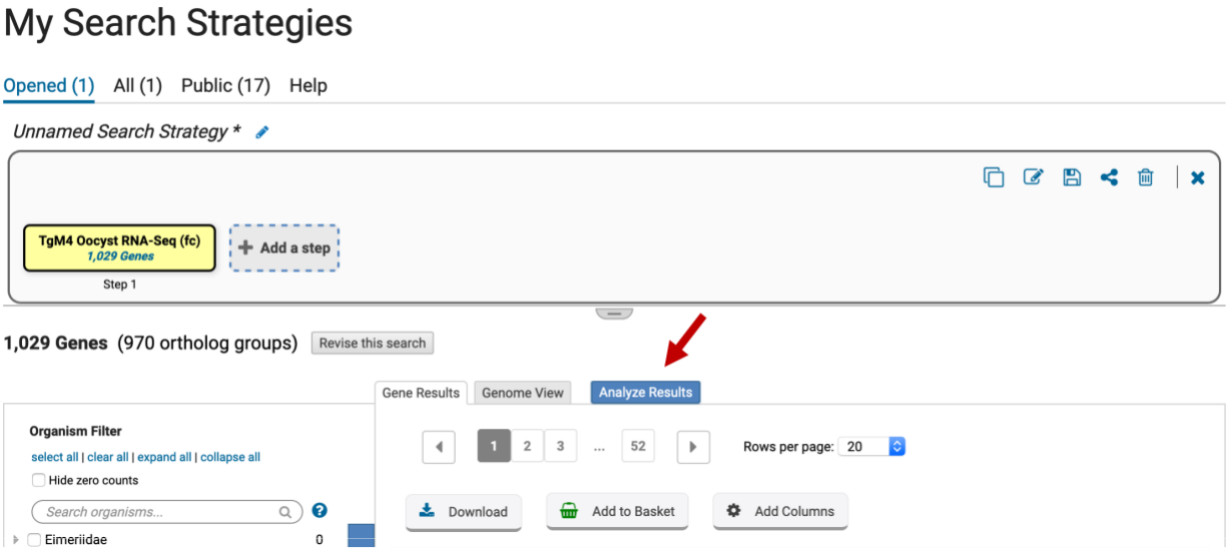
c. Once you find the data set of interest, click on the fold-change (FC) option. This will open a search page that contains the parameters that you can manipulate to search this data set. For this exercise, identify genes that are upregulated by 20-fold in the day 4 and day 10 time points compared to the day 0 time point. Parameters to set:
    2. Up-regulated
    3. 20-fold
    4. Maximum
    5. Day 0
    6. Minimum
    7. Day 4 and 10



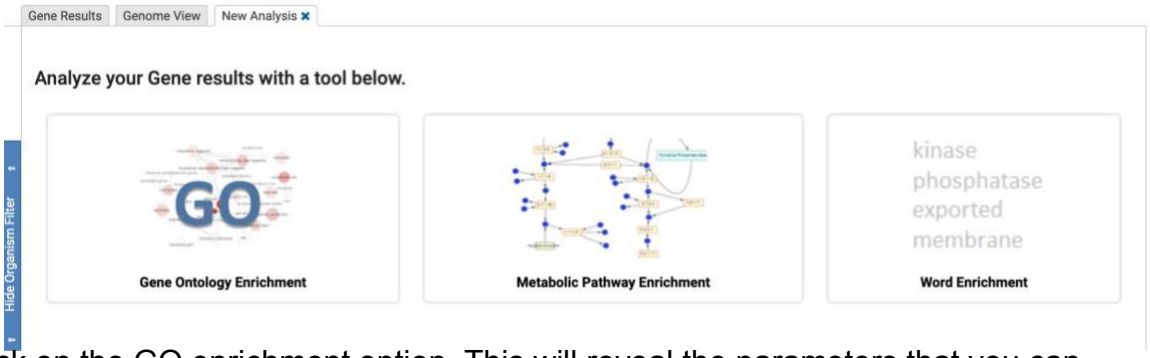
d. Once you have set the parameters, click the “Get Answer” button at the bottom of the search. This will return a one-step search strategy. How many genes did you get?

2. To run a GO enrichment analysis on these results, do the following:

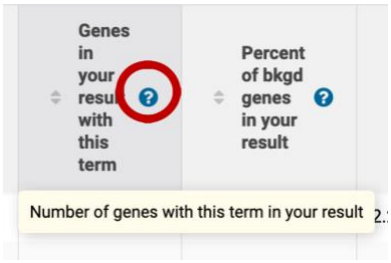
a. Click on the Analyze Results tab just above the list of genes (arrow in image below).



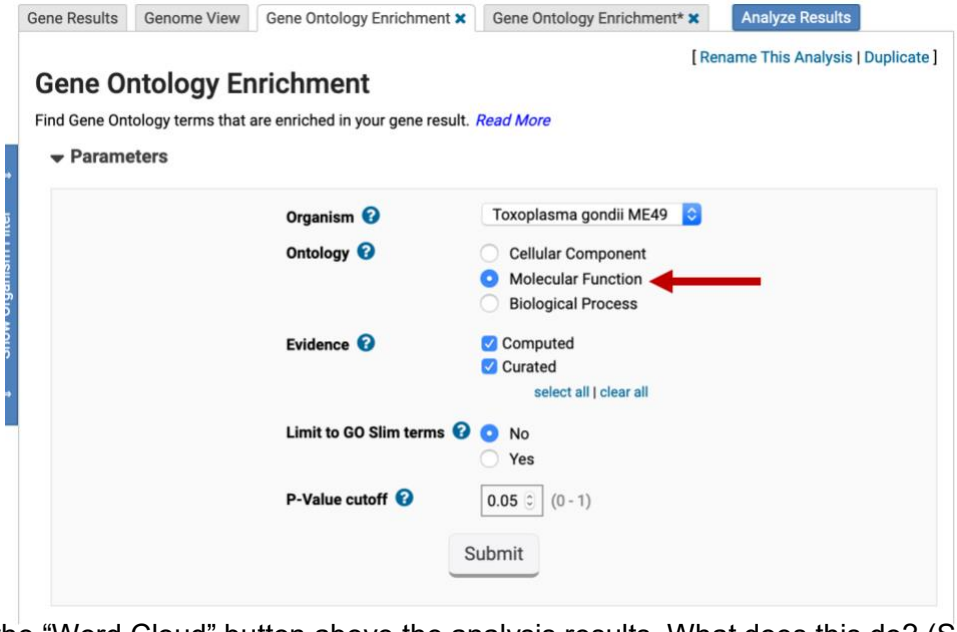
b. Click on the “Analyze Results” tab to reveal the different analyses that you can run on your results. Besides GO enrichment, what other analyses are available?



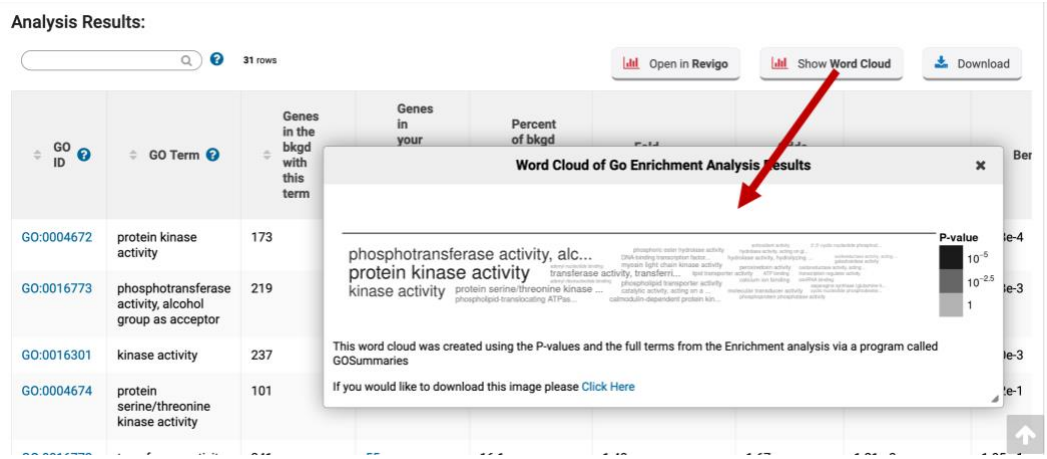
c. Click on the GO enrichment option. This will reveal the parameters that you can modify. For the purpose of this exercise, keep all the defaults and click on “Submit”.
d. What is the top enriched GO term from this analysis?
e. What do each of the columns in the analysis table represent? (Hint: move your mouse over the question mark next to each column header to get more information.)



f. Try rerunning the GO enrichment analysis, but this time select the Molecular Function ontology. What is the top enriched GO term?



g. Click on the “Word Cloud” button above the analysis results. What does this do? (See image below).



Additional resources:

Gene Ontology:
http://geneontology.org/docs/ontology-documentation/
Enzyme Commission numbers:
https://www.qmul.ac.uk/sbcs/iubmb/enzyme/
More info on Fischer's exact test:
http://www.biostathandbook.com/fishers.html
Fisher's Exact Test and the Hypergeometric Distribution (the M&M example):
https://youtu.be/udyAvvaMjfM
Some more info about Odds ratios:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/
False discovery rates and P value correction:
http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/
GO Slim:
http://www-legacy.geneontology.org/GO.slims.shtml
REVIGO:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800

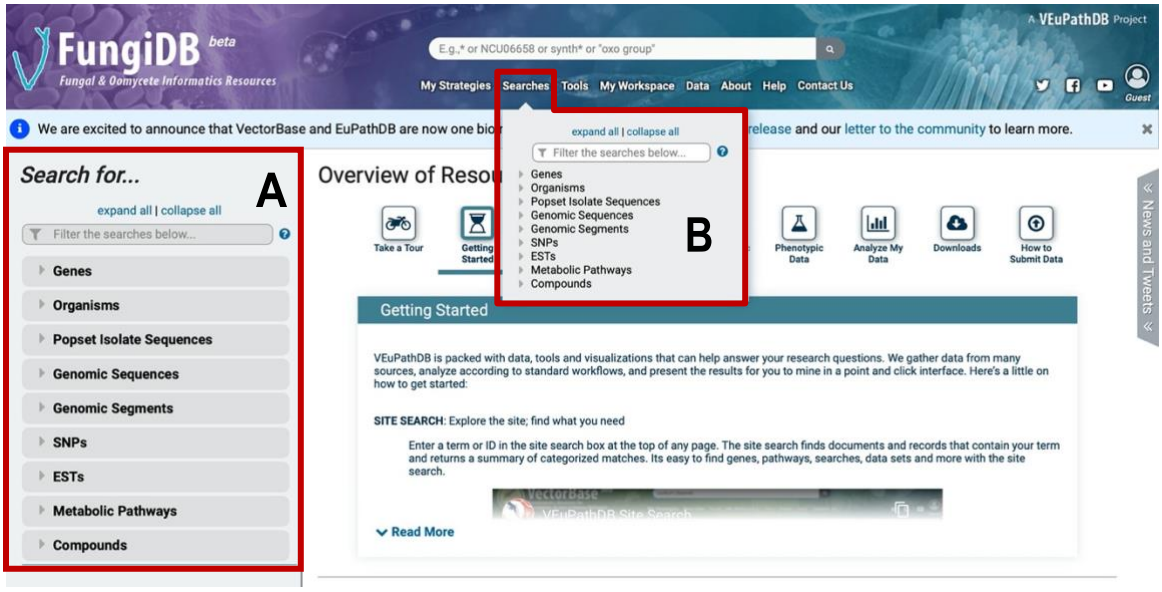
Regular Expressions & Genomic Colocation

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

Learning objectives:

- Run a regular expression search on amino acid sequences
- Run a regular expression search on nucleotide sequences
- Use the genomic colocation search

Protein or nucleotide sequences can be identified using the regular expression searches in VEuPathDB. This search is very useful to identify patterns of sequences.
Searches can be accessed from categorized menus in the left search for panel (A) or from the searches menu in the header (B).

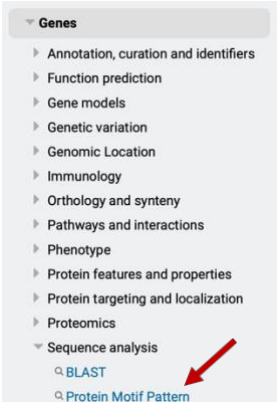


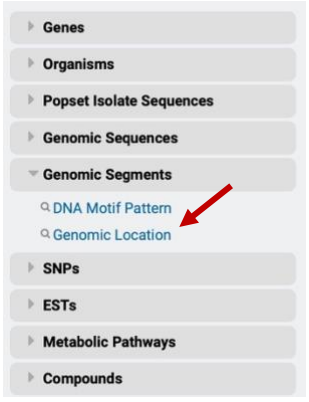
Accessing the protein motif pattern search:

- Click on the *Genes* category then click on the sequence analysis category

Accessing the DNA motif pattern search:

- Click on the *Genomic segments* category





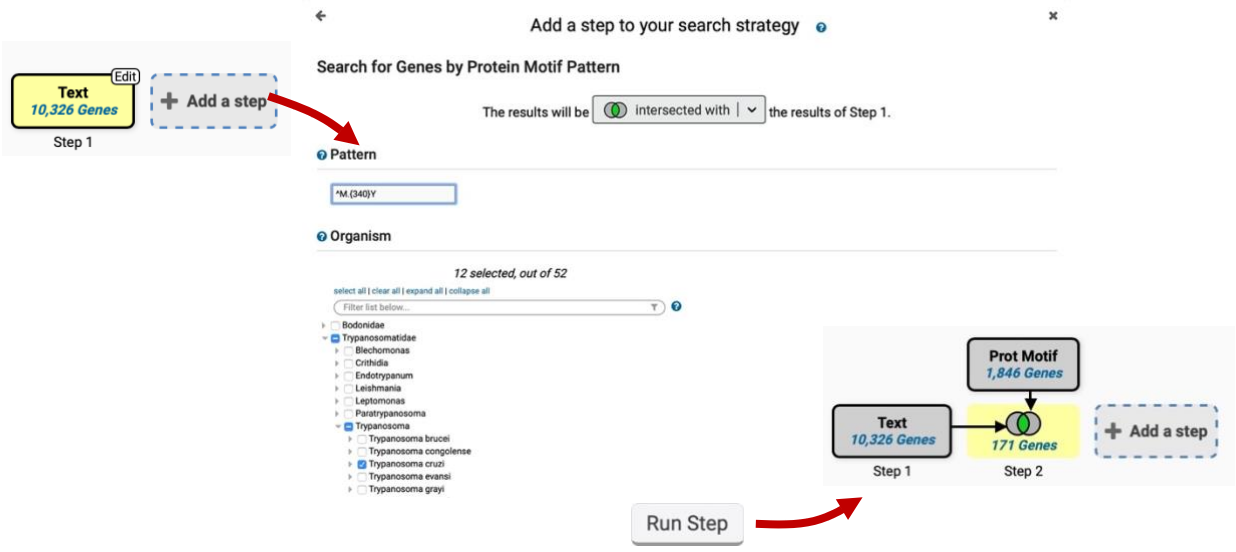
Note: the appendix at the end of this document includes additional regular expression help.

1. **Using regular expressions to find motifs in TriTrypDB: finding active trans-sialidases in *T. cruzi* (TriTrypDB).**

   a. *T. cruzi* has an expanded family of trans-sialidases. In fact, if you run a text search for any gene with the word “trans-sialidase” in its *product description*, you return over 10000 genes among the strains in the database!!! Try this and see what you get.

   b. Not all of the genes returned in (a) are predicted to be active. It is known that active trans-sialidases have a signature tyrosine (Y) at position 342 in their amino acid sequence. Add a motif search step to the text search in ‘a’ to identify only the active trans-sialidases.

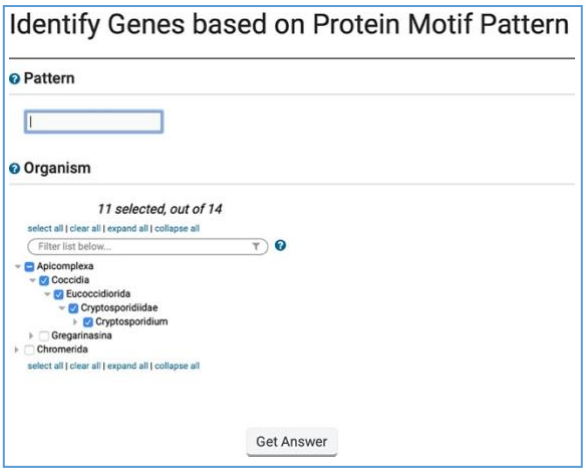
   c. Write a regular expression that defines a protein sequence that starts with a methionine, and is followed by 340 of any amino acids, followed by a tyrosine ‘Y’.



2. **Find Cryptosporidium genes with the YXXΦ receptor signal motif (CryptoDB)**

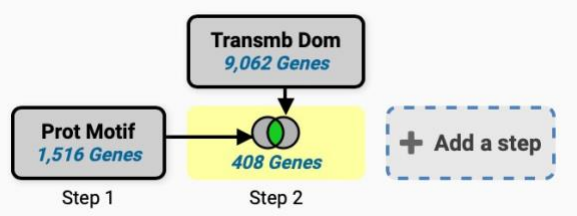
   The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine, threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal

end of the protein. ****Note**: *do not look for the Φ symbol on your keyboard – this will not work. Rather you should use the amino acid symbols.*



a. Use the "protein motif pattern" search to find all *Cryptosporidium* proteins that contain this motif anywhere in the terminal 10 amino acids of proteins. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed by any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine).

b. How many of these proteins also contain at least one transmembrane domain.



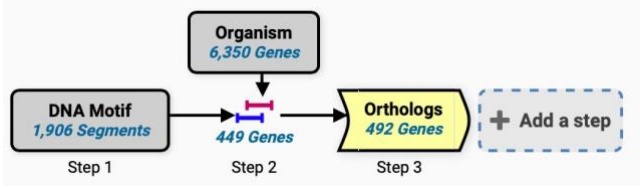
c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression).

Note: if you need help with the regular expression the answers are in appendix B.

3. Find fungal genes downstream of a regulatory DNA motif (FungiDB).

Transcriptional start sites are often located within a certain distance upstream of the genes or gene clusters that they regulate. In fungi, DNA motifs are also important for regulation of processes linked to host cell invasion or production of secondary metabolites. Readily available genomic data facilitate the discovery of regulatory motifs via examination of orthologous sequences.

The goal of this exercise is to identify all genes harboring upstream CACGTG motif, known for its role in transcriptional regulation. We will start our search in an extensively studied model

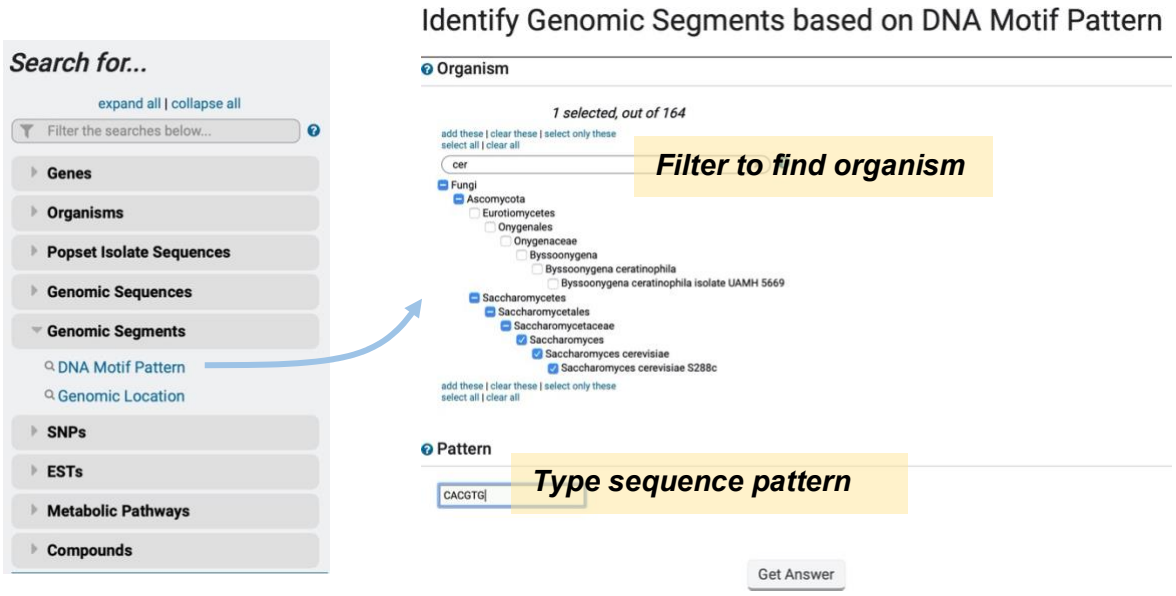


organism *Saccharomyces cerevisiae* and expand our search to *Fusarium graminearum*.

Here is a summary of the search strategy:

a. Find the CACGTG DNA motif in the *Saccharomyces cerevisiae* genome.

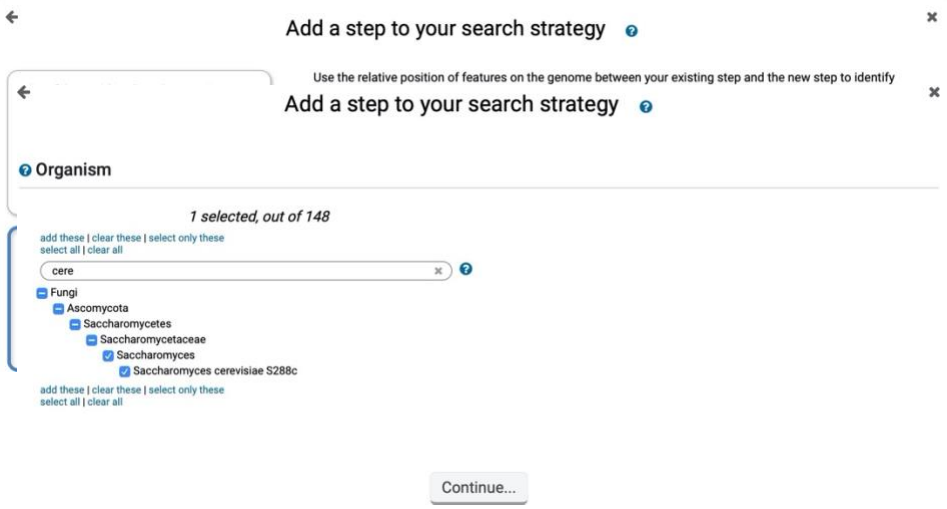
1. Select the "Search for genomic segments (DNA motif)" menu from the Search menu and look for CACGTG in *S. cerevisiae*.
2. Your search returns over 1900 DNA segments containing GACGTG motif. Next, let's look for putative regulatory targets of this motif by searching for genes that are located 600bp downstream of this sequence.



b. Identify genes with the CACGTG motif located 600bp upstream of an open reading frame.

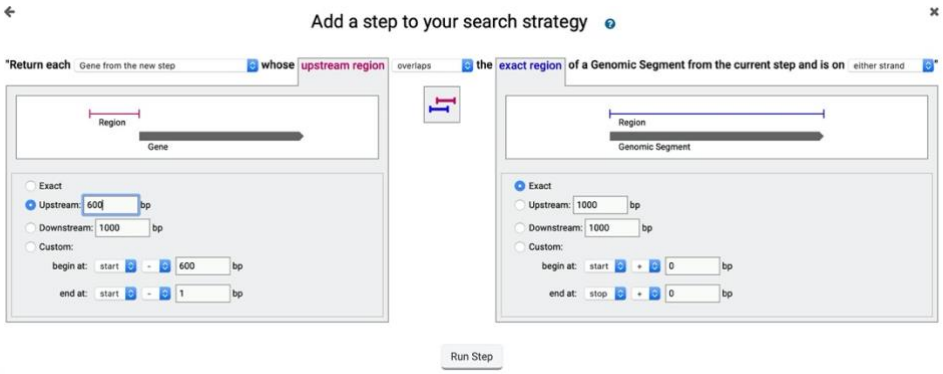
EuPathDB offers a colocation function to identify genomic features within a specified distance of each other. Run a search for all genes in *Saccharomyces cerevisiae* and use the colocation tool to identify genes that contain the CACGTG motif in their upstream regions. Follow these steps:

1. Click “Add Step”. Choose the option on the left called “Use **Genomic Colocation** to combine with other features” then select the *organism* gene search which can be found under the *Taxonomy* category.



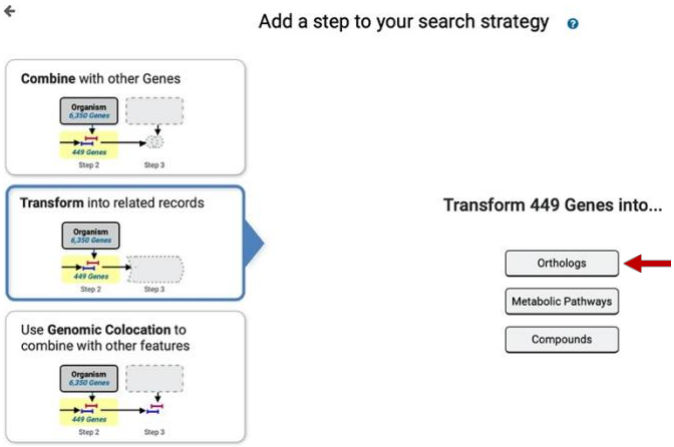
2. On the next page select *Saccharomyces cerevisiae* from the taxonomy browser and click on continue.

3. Configure the parameters on the next page to return each gene from step 2 whose upstream region (600bp) overlaps the exact region of a Genomic Segment in Step1 (CACGTG) and is on either strand.

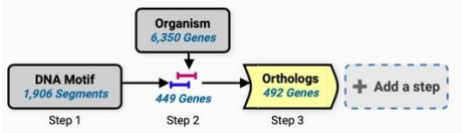


c. Identify orthologs S. cerevisiae genes in Fusarium graminearum.

All VEuPathDB sites offer tools to transform results between record types. The “Transform by Orthology” tool uses orthology clusters assigned by the OrthoMCL algorithm to enable transformation of a list of genes from one or more species to another (or more) species.



- click on add step then select the **Transform** into related records option from



the left side of the popup.
Next click on the *Orthologs* option.

- Select *F. graminearum* in the next popup window and click on *Run Step*.
-

Appendix A

Regular expression help

The following codes can be used to represent classes of amino acids.

AA property	Amino acids	Code
Acidic	**DE**	**0**
Alcohol	**ST**	**1**
Aliphatic	**ILV**	**2**
Aromatic	**FHWY**	**3**
Basic	**KRH**	**4**
Charged	**DEHKR**	**5**
Hydrophobic	**AVILMFYW**	**6**
Hydrophilic	**KRHDENQ**	**7**
Polar	**CDEHKNQRST**	**8**
Small	**ACDGNPSTV**	**9**
Tiny	**AGS**	**B**
Turnlike	**ACDEGHKNQRST**	**Z**
Any	**ACDEFGHIKLM NPQRSTVWY**	**X**

The following is a simple explanation of regular expressions.

Perl regular expressions are terms used for pattern matching in text strings, e.g. ***'aadgt', 'aa+dgt', 'a|d|c', '[mac]a'.***

Because nucleotide and amino acid sequences are text strings, regular expressions are very useful for finding motifs within sequences.

Motifs often include repetitive or ambiguous assignments at some locations. The rules and special characters used in regular expressions help define the full set of strings that match the motif pattern.

The following is a description of some of these characters and examples of how they are

used.
Although regular expressions seem complicated at first, they are very useful and easy to understand after going through some examples.

Special Characters

. Match any character.

+ Matches "one or more of the preceding characters".

******* Matches "any number of occurrences of the preceding character", including 0.

? Matches "zero or one occurrences of the preceding character".

[ ] Matches any character contained in the brackets.

[^] Match any character *except* those in the brackets.

{n} Matches when the preceding character, or character range, occurs exactly n times.

{n,} Matches when the preceding character occurs at least n times.

{n,m} Matches when the preceding character occurs at least n times, but no more than m times.

Here are some examples of searches.

ad+f (1 or more occurences of 'd') would match any of the following:
adf
addf
addd
f
addd
dddf
...

ad*f (0 or more occurences of 'd') would match:
a
f
a
d
f
a
d
d
f
adddf
...

ad?f (0 or 1 occurence of 'd') would match:
a
f
a

d
f

a[yst]c would match:

a
t
c
a
s
c
a
y
c

Specify the number of occurrences of a residue.

P{1,5} would match P from 1 to 5 times.

.{1,30} would match any amino acid 1 to 30 times so you could find a motif within 30 amino acids of something like the beginning.

Pattern Anchors

^	Match only at the beginning of the string.
$	Match only at the end of the string.

Here are examples of expressions using pattern anchors.

^*mdef* (e.g. a protein sequence **starting with** 'mdef') would match:

- mdef
- mdefab
- mdefared

fadfk but

not match:

- edefa
- emdefa
- eeeemdef

kdel$ (searches for proteins **ending with** 'kdel', a standard ER retention signal) would match:

- eeeekdel

- kdel

but not match :

- edefkdell
- akdeleef

Appendix B

Answers to exercise 2:

A: YXXΦ in the terminal 10 amino acids à ReEX = Y..[FTY].{0,6}$
B: YXXΦ in the terminal 20 amino acids à ReEX = Y..[FTY].{0,16}$

Variant calling in VEuPathDB galaxy (Part 1)

Learning objectives:

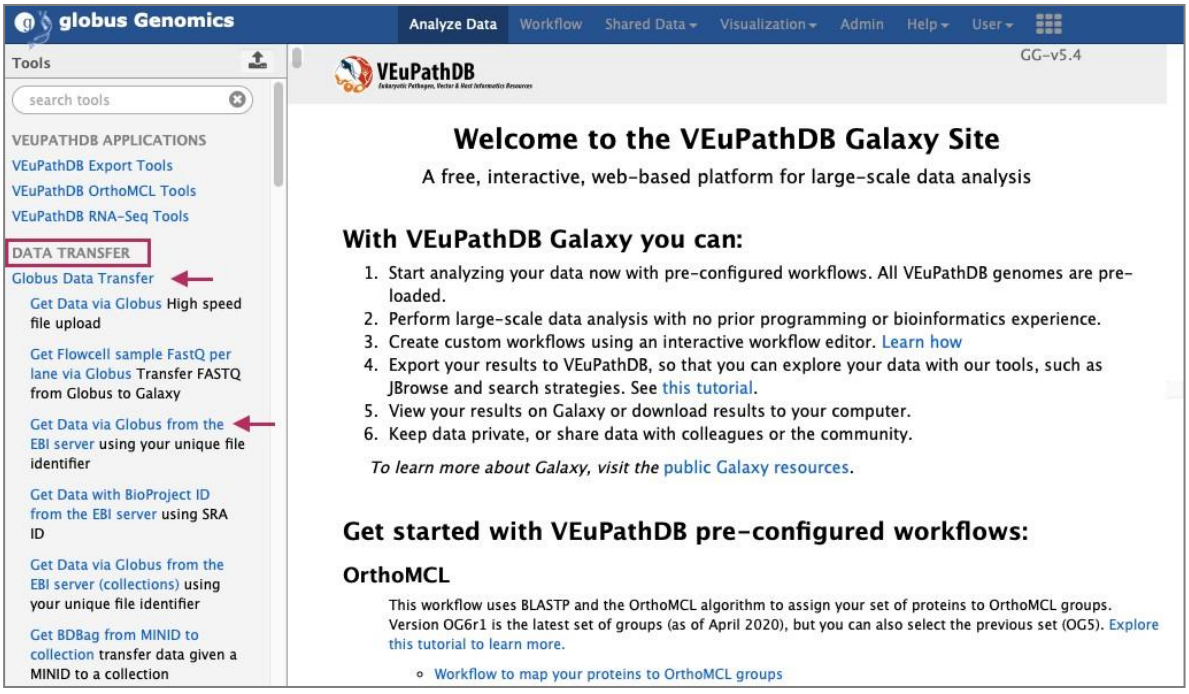
1. Retrieve DNA sequence data from the sequence repository EBI and upload data to VEuPathDB Galaxy using Globus Data Transfer;
2. Name a new project/history;
3. Deploy a Variant calling workflow in the VEuPathDB Galaxy.

Galaxy is an open, web-based platform for data-intensive biomedical research. Galaxy allows you to perform, reproduce, and share complete analyses without the use of command-line scripting. VEuPathDB developed its Galaxy instance in collaboration with Globus Genomics (VEuPathDB Galaxy). To learn how to use Galaxy, follow this link to access tutorials prepared by the Galaxy Training Network:
https://wiki.galaxyproject.org/Learn#Galaxy_101

There are different ways to get data into Galaxy. In this exercise we will use Globus Data Transfer to get data from the EBI server using a unique project ID.

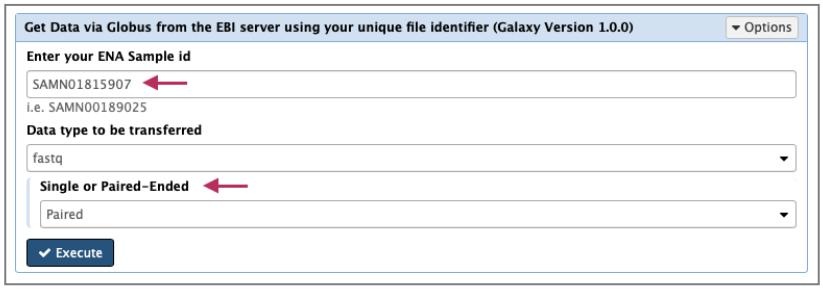
1. Retrieve DNA sequence data from the sequence repository and upload data to VEuPathDB Galaxy using Globus Data Transfer option.

a. Click on the “Globus Data Transfer” menu on the left to expand the Data Transfer section.
b. Click on the “Get Data via Globus from the EBI server” link.



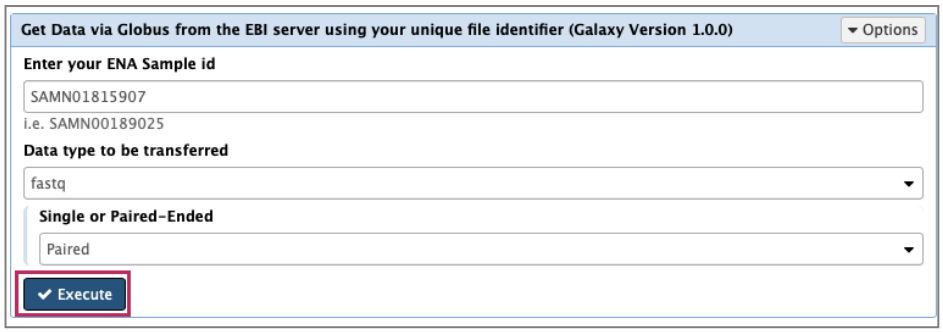
c. Enter ENA sample ID and define the dataset type to be transferred into the VEuPathDB Galaxy workspace.

The ENA ID should start with the letters 'SAM'. For this exercise, we will use SAMN01815907, which is a paired-ended dataset. Take care to specify



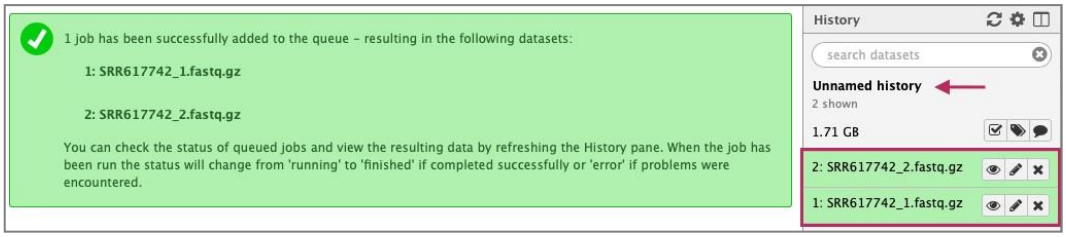
whether a dataset is a single or paired-ended as incorrect selection will cause the upload to fail.

d. Once the form is properly filled, click on the "Execute" button to start the data transfer process.



e. When the job has been successfully deployed and added to the queue, the screen will refresh, and the added job will appear in the history on the right.

Note: new jobs are highlighted in grey, in progress – yellow, and those completed are in green.

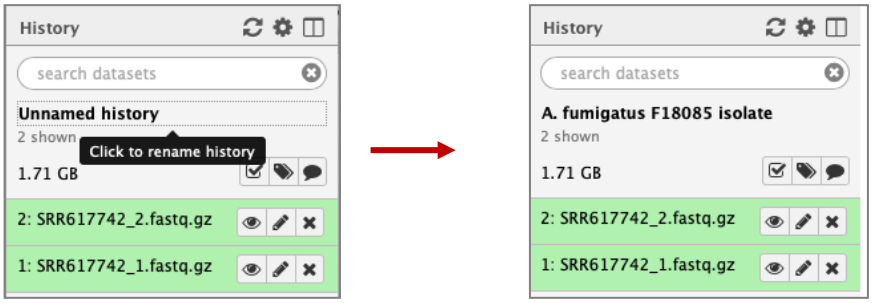


Notice that there are two files appearing in the history on the right. This is because the uploaded data is paired-ended.

2. **Rename your history.**

   By default, all new jobs will be added to the current history on the right. Unless renamed, the history will show up in your history as “Unnamed history”. Let’s rename the history to help us track this project in the future.

   f. Click on the “Unnamed history” and type “A. fumigatus F18085 isolate”, and then press “enter” to rename this history.

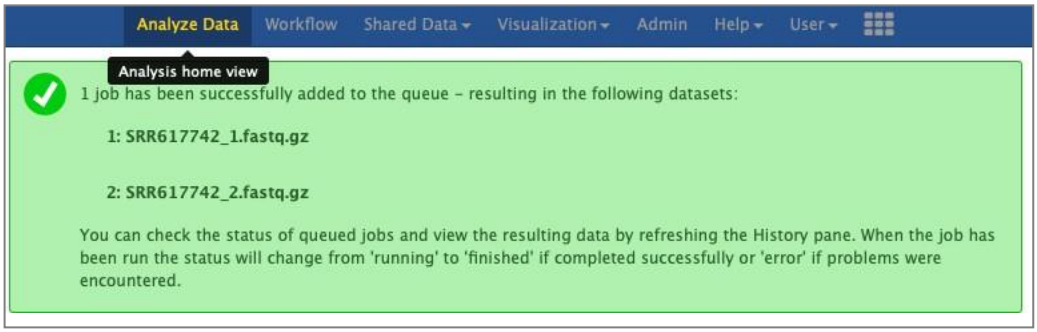


Note: if you would like to start a new project/history, click on the wheel button at the top of the history section and select “Create a new history”.

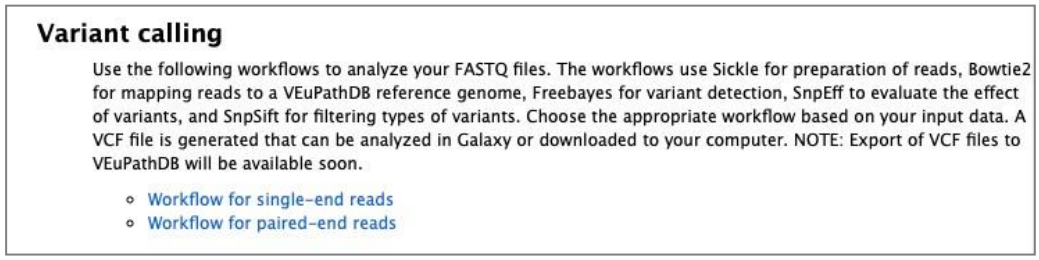
3. **Deploy a Variant calling workflow.**

   VEuPathDB Galaxy main landing page has several workflows for variant calling.

   g. To navigate to the main page, click on the “Analyze Data”, which is located in the main menu at the top.

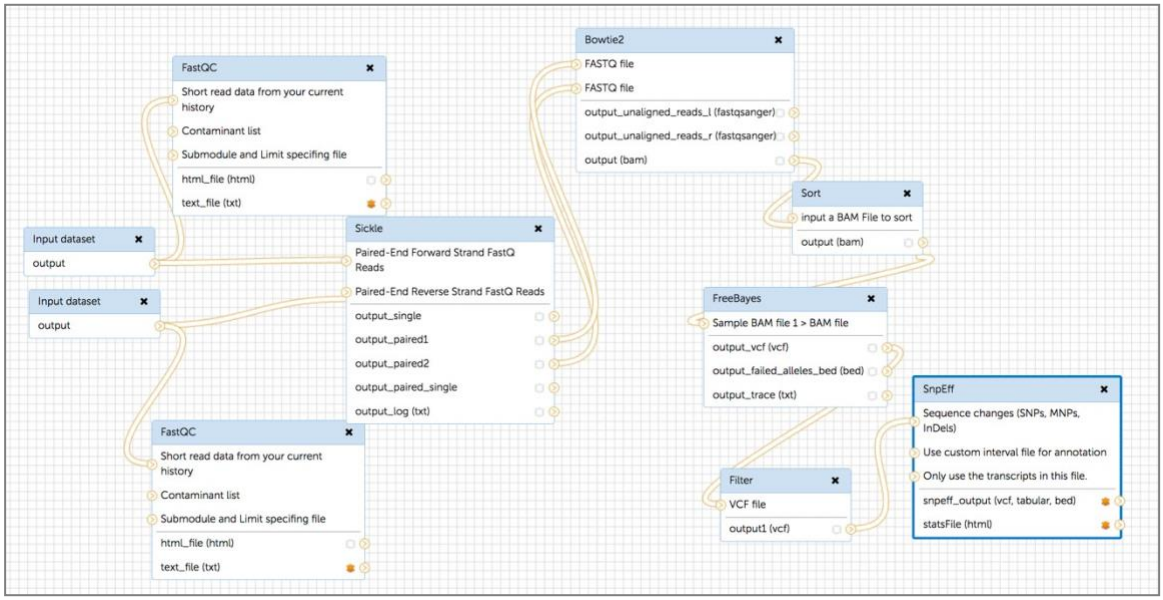


   h. Scroll down to the Variant calling section and choose the workflow for paired-end reads.



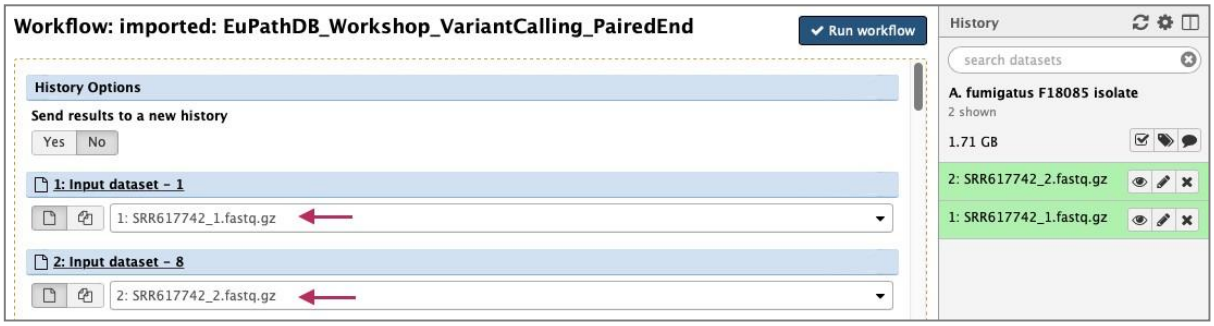
The pre-configured variant calling workflows include the following steps:

- Determine quality of the reads and generate reports (FastQC);
- Trim reads based on their quality scores (Sickle);
- Align reads to a reference genome using Bowtie2 and generate coverage plots ;
- Sort alignments with respect to their chromosomal positions (Sort);
- Detect variants (FreeBayes);
- Filter SNP candidates (Filter);
- Analyze and annotate variants, and calculate the effects of SNPs via SnpEff.

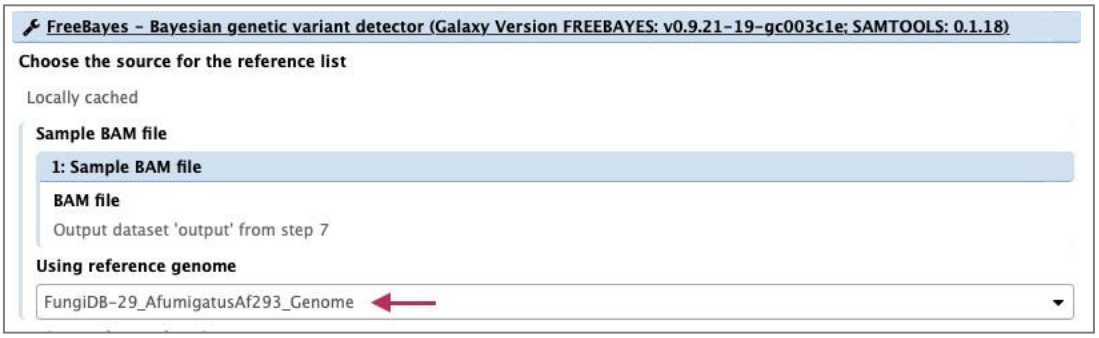


i. Click on the workflow for paired-end reads and set workflow parameters.

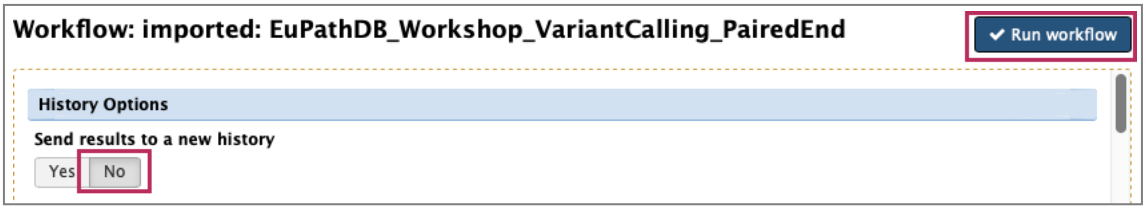
   - Make sure that the input steps for paired-end data are set to the xxxx_1.fastq.gz and xxxx_2.fastq.gz file (by default the same file will be selected in both fileds).



- Select the correct reference genome.
  - Select *Aspergillus fumigatus* Af293 as a reference genome (steps: Bowtie2, FreeBayes, SnpEff).

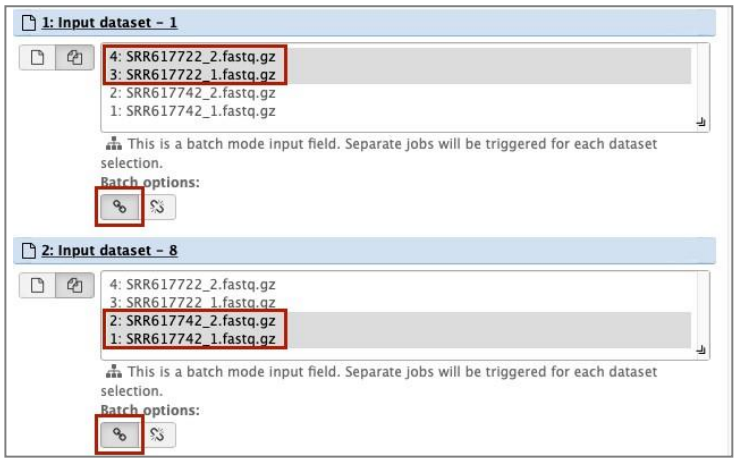


- Choose to deploy the analysis within the same history and click on the Run workflow button.



Note: You can use the same workflow to analyze multiple samples in batches. The. Upload steps remain the same, however, when setting up the workflow, click on multiple dataset button within the input dataset section.


1: Input dataset - 1
No fastq, fa
Multiple datasets

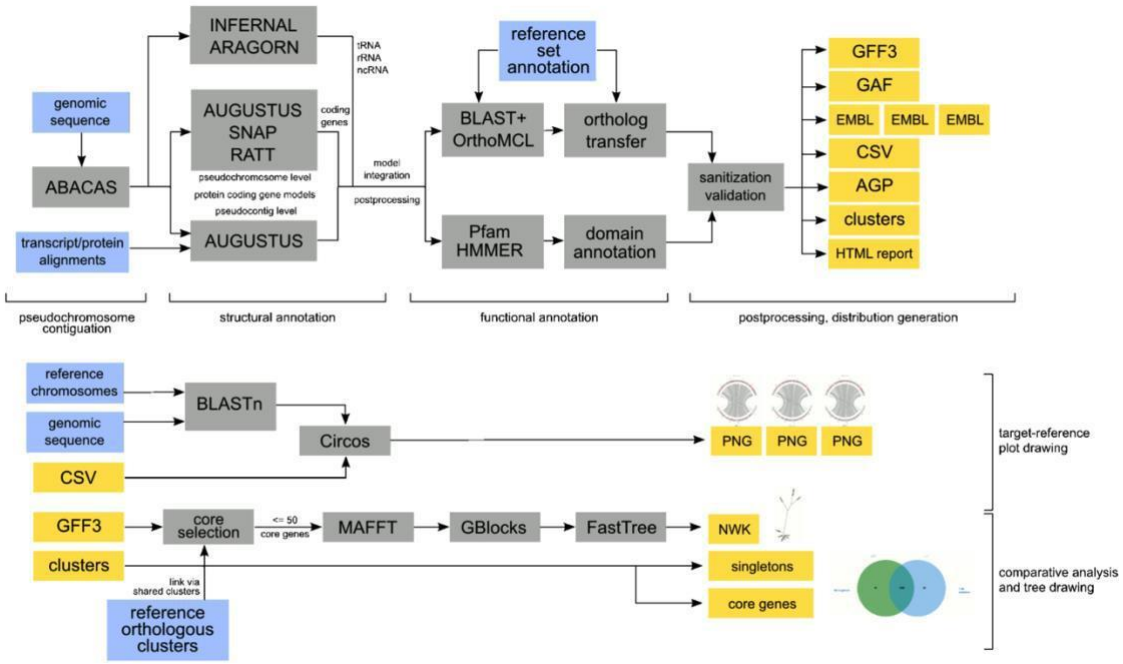

1: Input dataset - 1
4: SRR617722_2.fastq.gz
3: SRR617722_1.fastq.gz
2: SRR617742_2.fastq.gz
1: SRR617742_1.fastq.gz
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
Batch options:
2: Input dataset - 8
4: SRR617722_2.fastq.gz
3: SRR617722_1.fastq.gz
2: SRR617742_2.fastq.gz
1: SRR617742_1.fastq.gz
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
Batch options:

Genome Annotation with Companion (Part 1)

Learning objectives:

- Download genomes and chromosomes from VEuPathDB
- Annotate a genome with Companion
- Interpreting the Companion result
- Download the Companion output and load it into Artemis
- Learn how to use ACT to compare genomes and look at the Companion output

Companion, is an online pipeline that employs different software to annotate and compare an assembled sequence to a reference-annotated genome. The figure below illustrates the Companion pipeline, the software used and the expected output.



For this exercise, we will start with an assembled genome that is unannotated. We will obtain the assembled FASTA files from VEuPathDB sites. Companion can be accessed here: http://companion.gla.ac.uk/

Each group will download one of the following genomes (the tinyURL links will initiate the download) and will use Companion to compare with the specified genome as reference.

Group 1 – *Plasmodium coatneyi* Hackeri using *Plasmodium knowlesi* as reference
https://tinyurl.com/y47vvsoj

Group 2 - *Plasmodium coatneyi* Hackeri using *Plasmodium falciparum* as reference
https://tinyurl.com/y47vvsoj

Group 3 – *Cryptosporidium meleagridis* using *Cryptosporidium parvum* as reference
https://tinyurl.com/y4fgc3p5

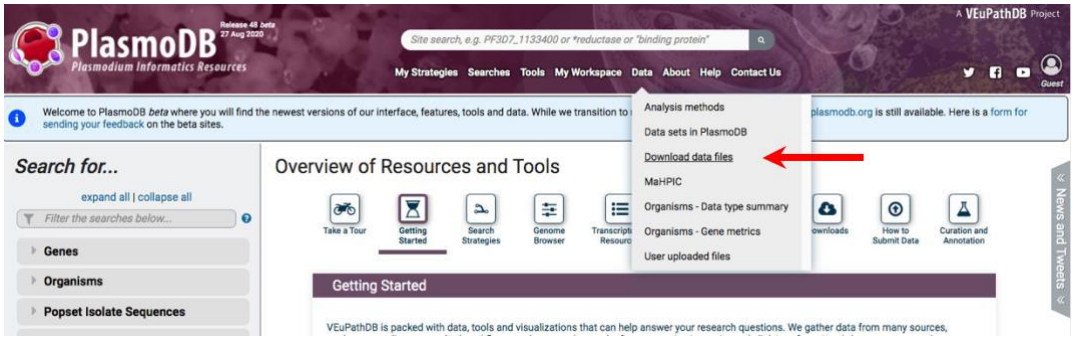
Group 4 *Cryptosporidium baileyi* using *Cryptosporidium parvum* as reference
https://tinyurl.com/y44ucs5t

Group 5 *Trypanosoma congolense* using *Trypanosoma brucei* 927 as reference.
https://tinyurl.com/yxausbhg

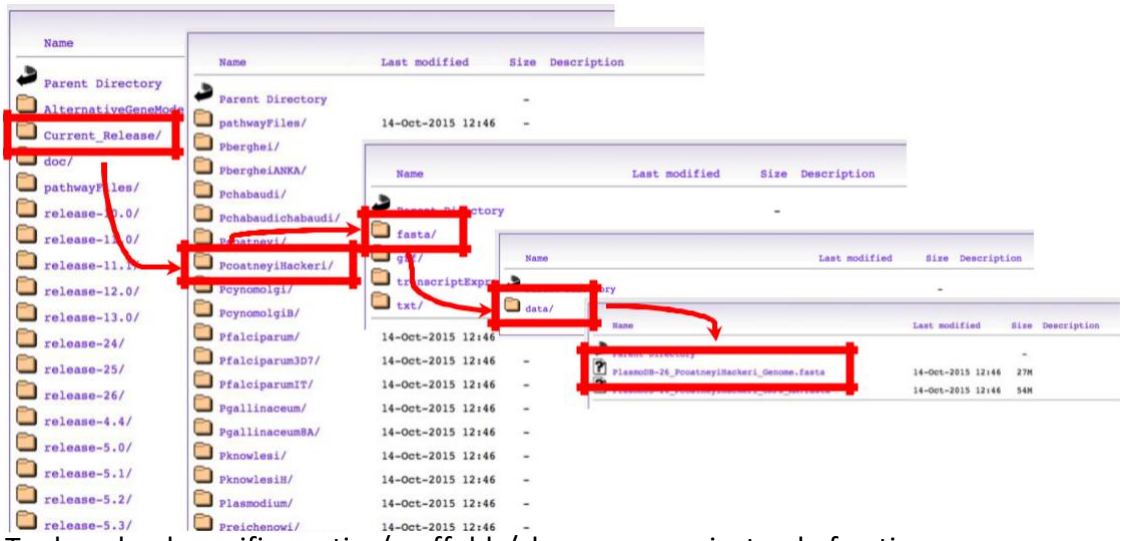
Group 6 *Trypanosoma congolense* 2019 using *Trypanosoma brucei* 927 as reference.
https://tinyurl.com/y4pqscrm

A word about downloads:

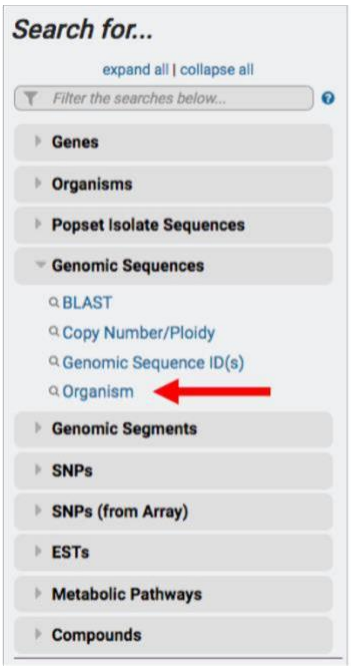
TinyUrls above are direct links to our genome FASTA files in the corresponding VEuPathDB site downloads section. All genomes in VEuPathDB sites are available for download from the "Data File" download section, which you can access from the top menu by clicking on "Data".

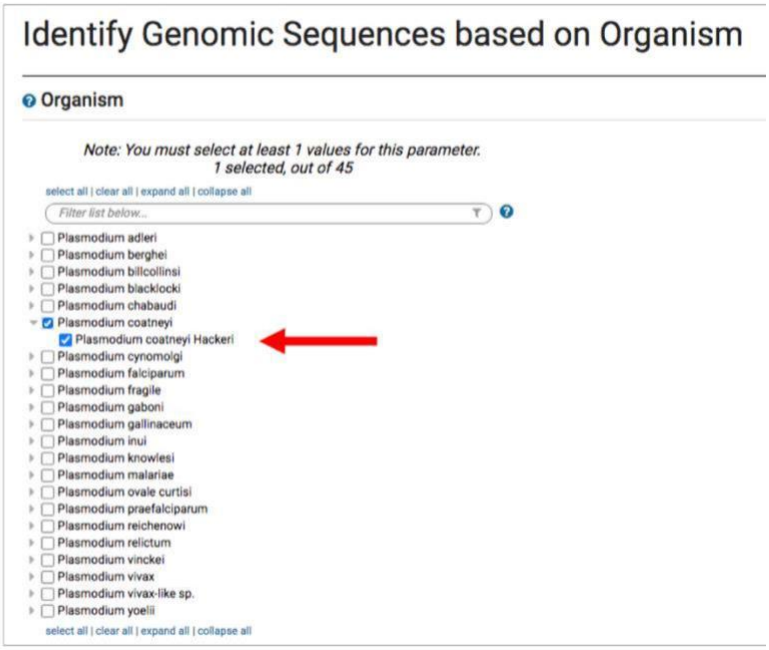


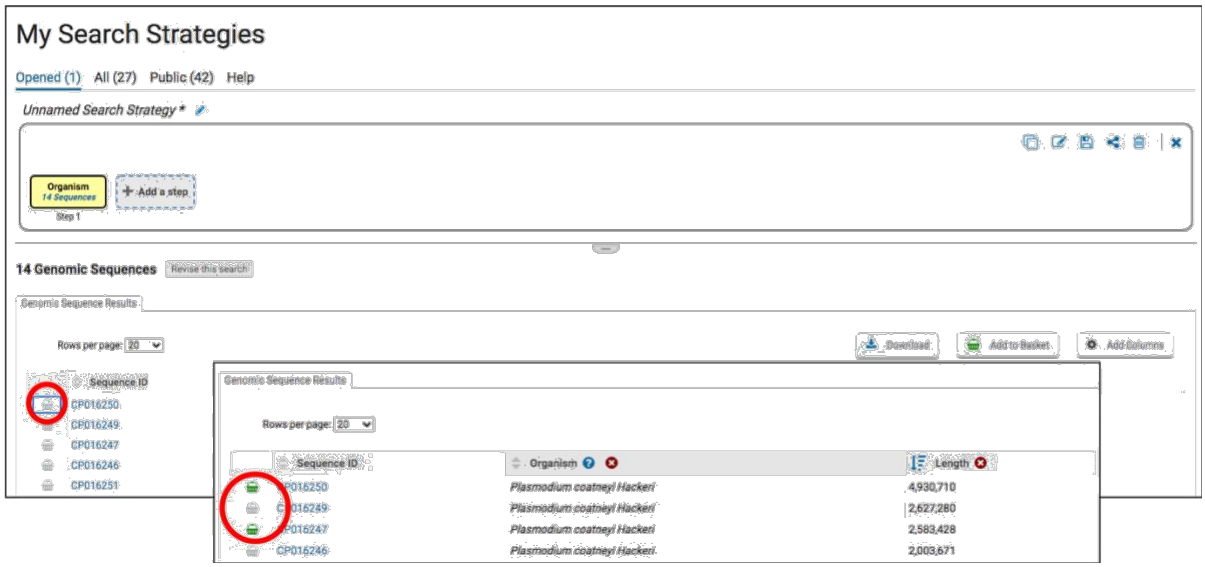
Selecting the option "Download data files" takes you to the download directories where you can navigate to the genome and data type you are looking for.



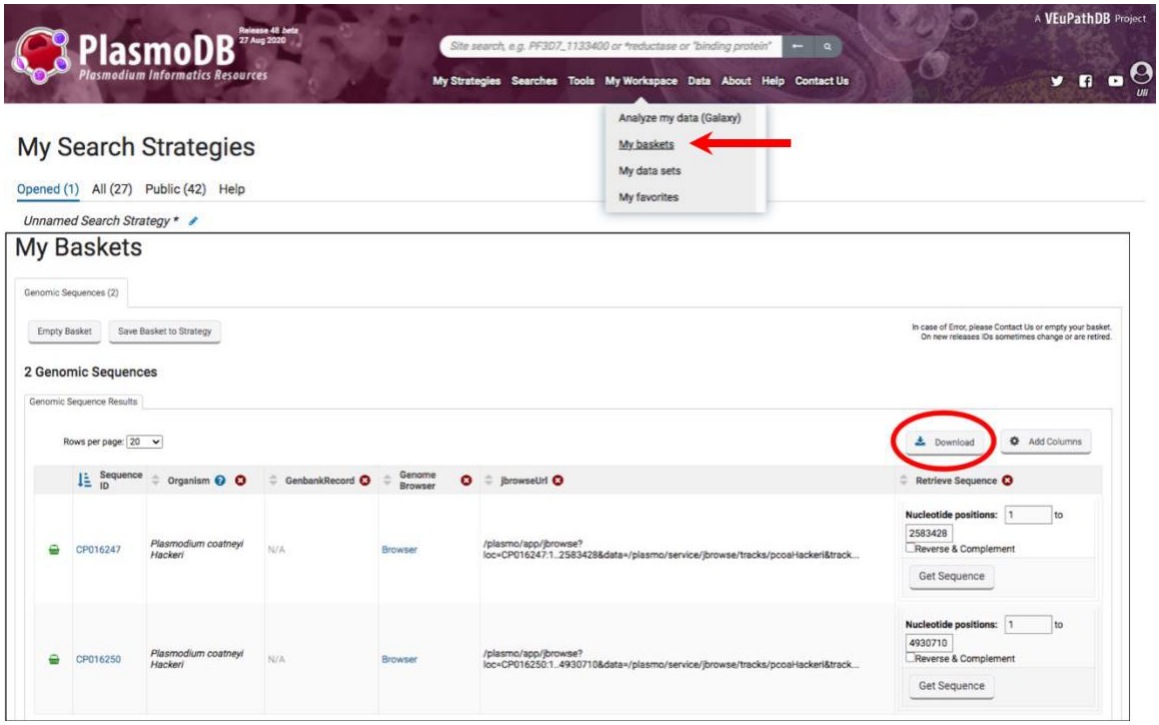
To download specific contigs/scaffolds/chromosomes instead of entire genomes, use a genomic sequence search and place the desired sequences into your basket.

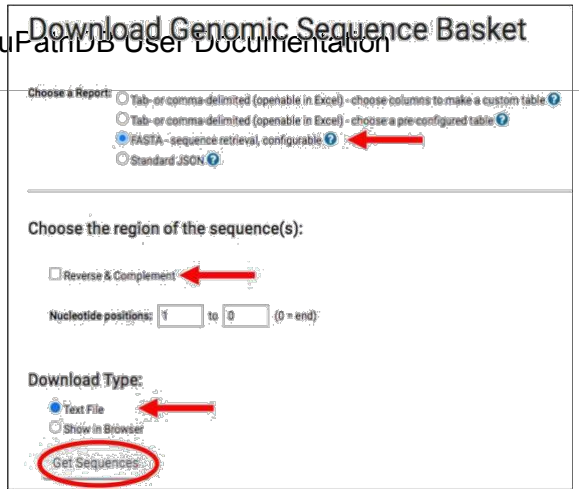






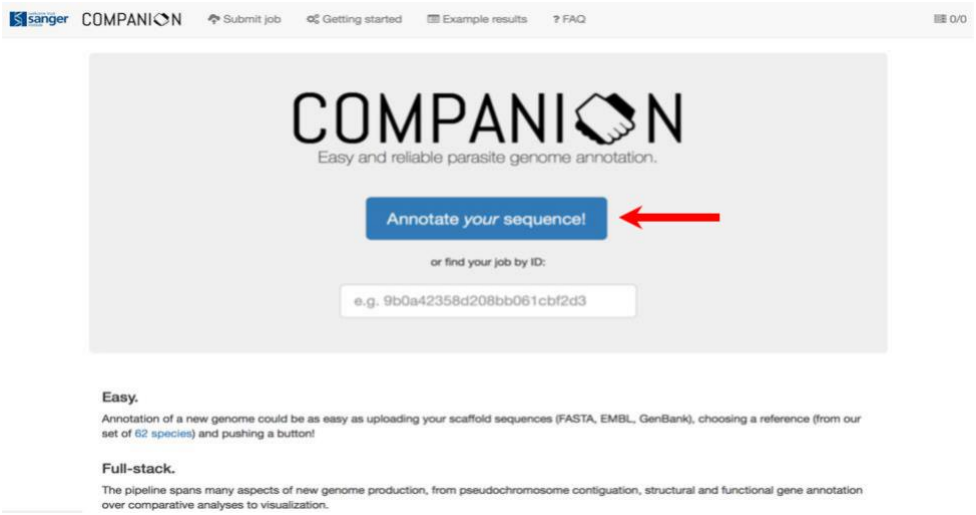
You can access the basket from the top menu.





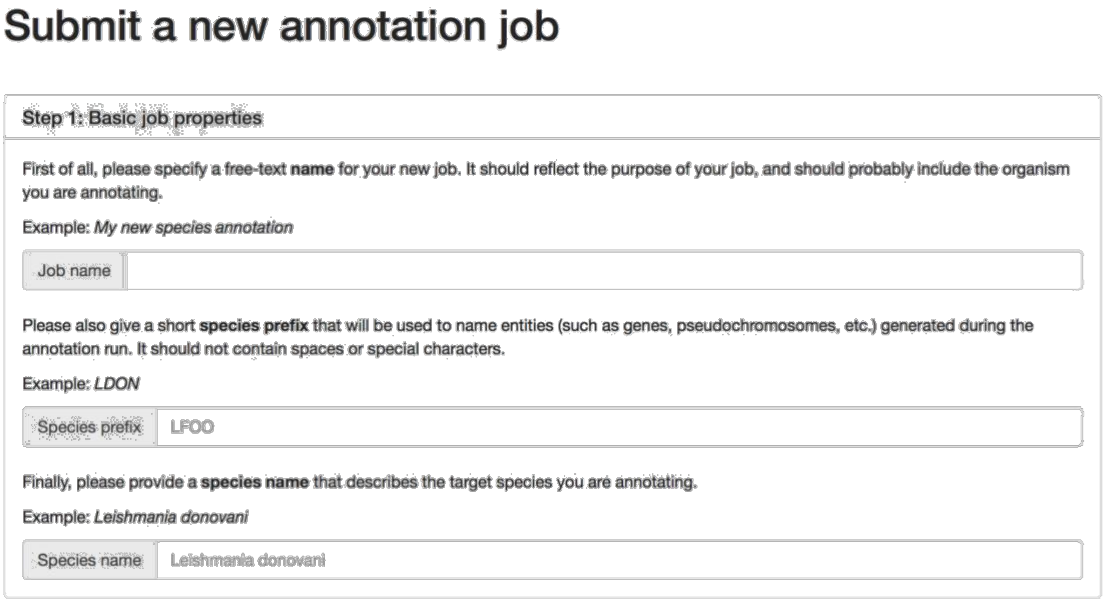
-**Back to the Annotation**: Once you have downloaded your sequence file, go to the Companion site:
http://companion.gla.ac.uk/

- Click on the “Annotate your sequence” link.



-Follow the instructions as described on the Companion website:

1. Provide basic information about the job you are about to submit. This includes a job name, species prefix (usually the first letter of the genus and the first three letters of the species: *Cryptosporidium parvum* = CPAR).

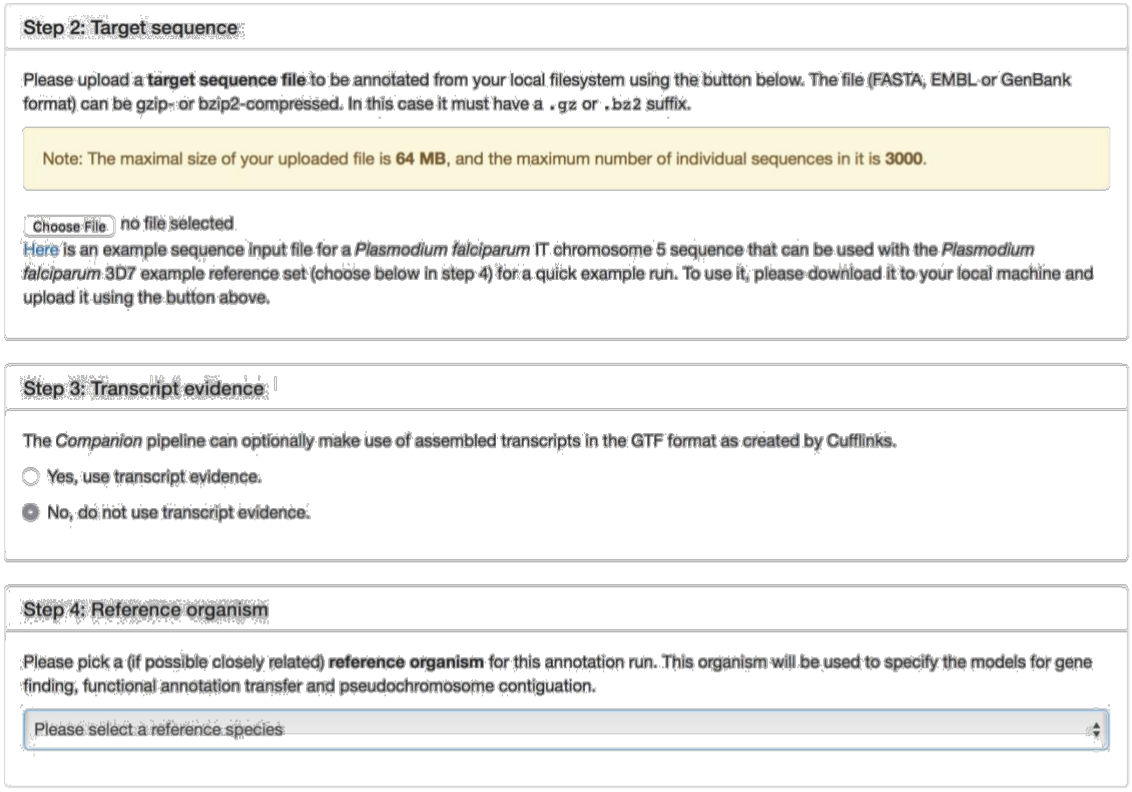


2. In step 2, choose the assembly file that you downloaded.
2. In step 3, indicate if you will be using RNAseq evidence to guide the annotation – in this exercise we will **not** use any RNAseq data.
3. In step 4, select the reference sequence you would like to use to transfer the annotation and to compare your sequence to. Typically, you would like to use a reference that is closely related, so a phylogenetic tree might be useful to look at. Here are examples of phylogenies for *Plasmodium* and *Cryptosporidium*.

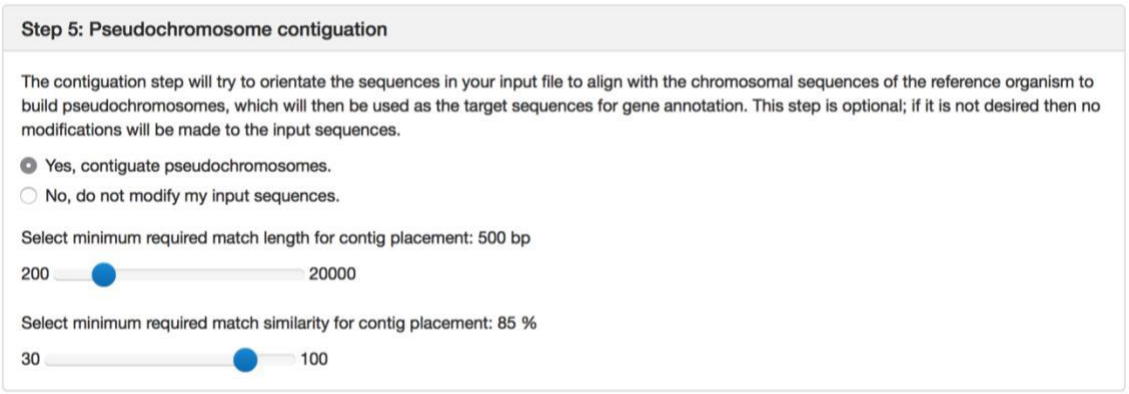
   http://tolweb.org/Cryptosporidium/124803

   http://tolweb.org/Plasmodium/68071

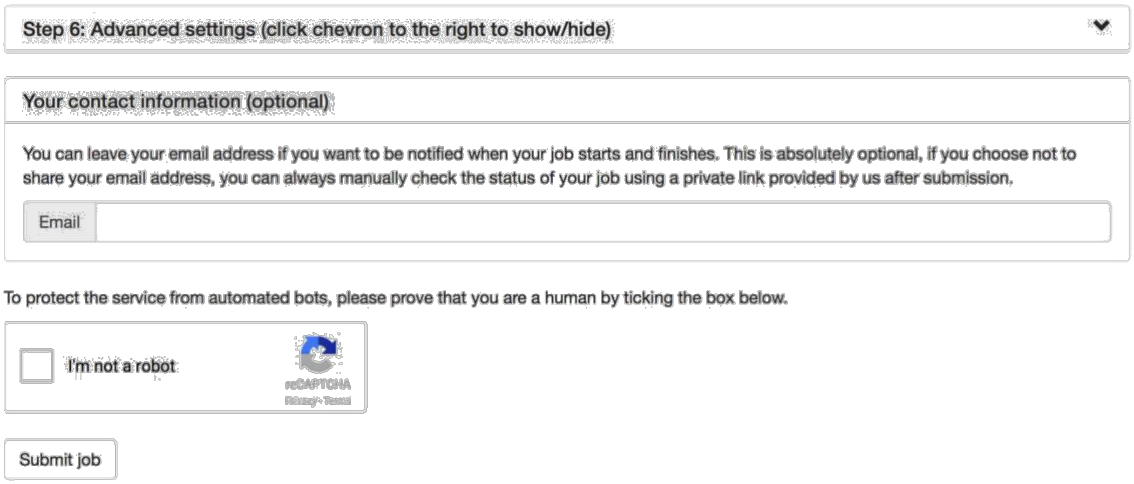
   Trypanosoma phylogenetic tree
   https://projects.exeter.ac.uk/meeg/sites/default/files/pictures/tryp_tree.jpg



5. In step 5, there are a few more parameters you may want to examine. For the purpose of our exercise we will keep these at the default values.

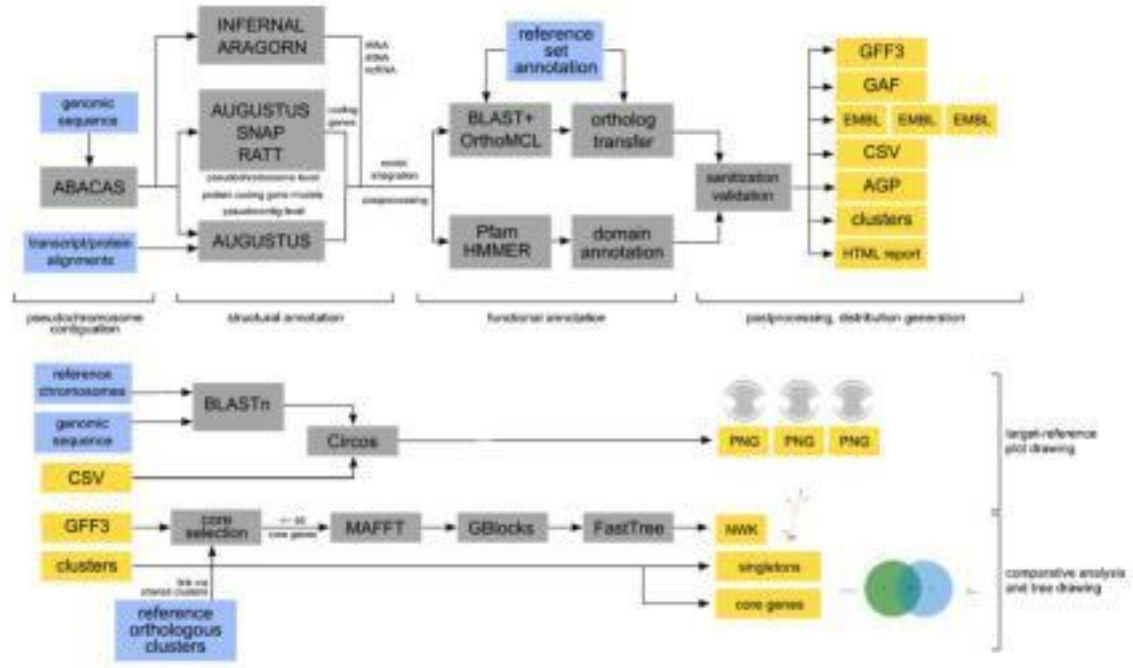


6. Enter your email address to get an update when your job starts running and when it is complete. Next, click on the “I’m not a robot” captcha (Completely Automated Public Turing test to tell Computers and Humans Apart). Finally, click on the “Submit Job” link.

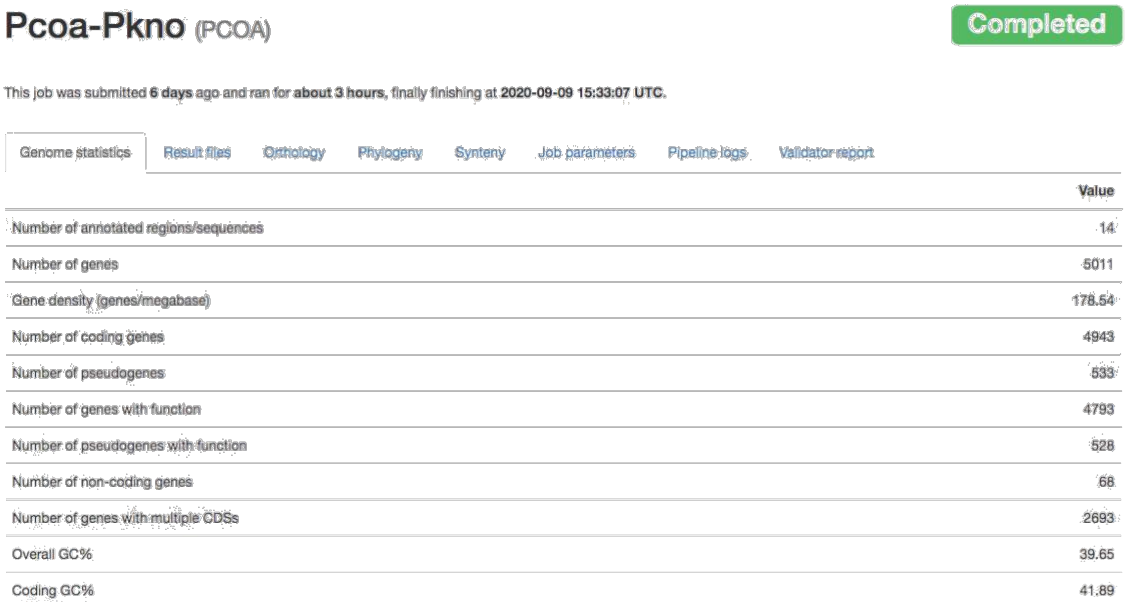


Genome Annotation with Companion (Part 2)

You should have gotten an email indicating the status of your annotation (ie. job started and job complete). The email contains a link to the annotation output.



- Explore your results with your group and discuss the annotation findings:
  - What does the genome statics tab tell you about your annotation? Are the results surprising? You can explore the reference genome you used in VEuPathDB to help you assess the results. (For example, are you getting a reasonable number of genes? What about the GC content? Number of non-coding genes?)



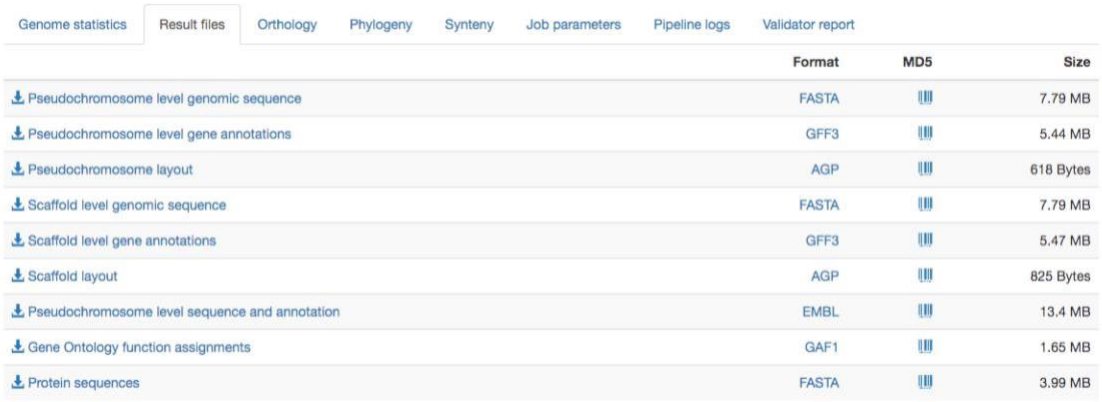
Pcoa-Pkno (PCOA) **Completed**

This job was submitted **6 days** ago and ran for **about 3 hours**, finally finishing at **2020-09-09 15:33:07 UTC**.

Genome statistics | Result files | Orthology | Phylogeny | Synteny | Job parameters | Pipeline logs | Validator report

	Value
Number of annotated regions/sequences	14
Number of genes	5011
Gene density (genes/megabase)	178.54
Number of coding genes	4943
Number of pseudogenes	533
Number of genes with function	4793
Number of pseudogenes with function	528
Number of non-coding genes	68
Number of genes with multiple CDSs	2693
Overall GC%	39.65
Coding GC%	41.89

- What does the “Result files” tab contain? What is an AGP file? What is a GFF3 file?

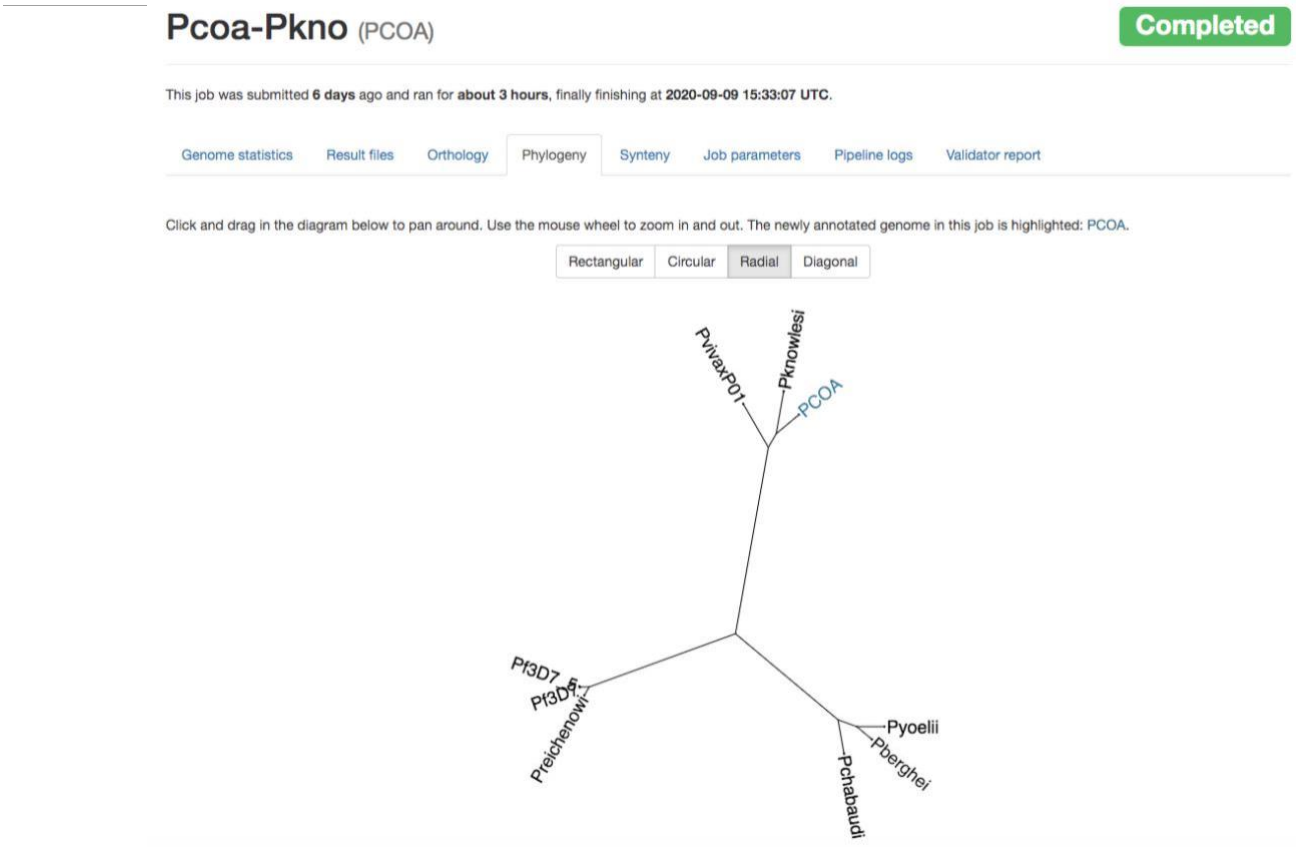


Genome statistics | Result files | Orthology | Phylogeny | Synteny | Job parameters | Pipeline logs | Validator report

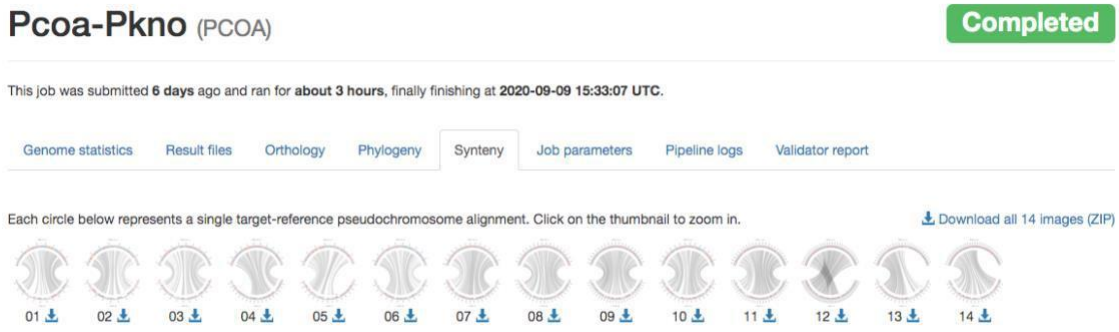
	Format	MD5	Size
Pseudochromosome level genomic sequence	FASTA		7.79 MB
Pseudochromosome level gene annotations	GFF3		5.44 MB
Pseudochromosome layout	AGP		618 Bytes
Scaffold level genomic sequence	FASTA		7.79 MB
Scaffold level gene annotations	GFF3		5.47 MB
Scaffold layout	AGP		825 Bytes
Pseudochromosome level sequence and annotation	EMBL		13.4 MB
Gene Ontology function assignments	GAF1		1.65 MB
Protein sequences	FASTA		3.99 MB

- What does the “orthology” tab display? How many predicted proteins from your new genome are in common with ones from the reference genome? How many are unique to yours? What do singletons represent (click on the singleton number to see what these genes are?

- What does the phylogeny tab represent? Does it make sense?



- Examine the Synteny tab – are these genomes syntenic?



Genetic Variation Exercises

SNPs and CNVs

Single Nucleotide Polymorphisms (SNPs): single nucleotide changes between isolates or strains. SNPs have different functional effects with most having no consequential effect on gene function. SNPs may directly affect protein function when they are non-synonymous (results in a change in the amino acid; missense) or when they are cause a premature stop codon (nonsense). SNPs that do not fall within genes are non-coding (between genes or intronic). These types of SNPs may still affect splicing, mRNA stability, transcription, etc.
Copy number variation (CNV): variation in copy number of genes or regions of a genome. CNVs may be result of deletions or duplications.

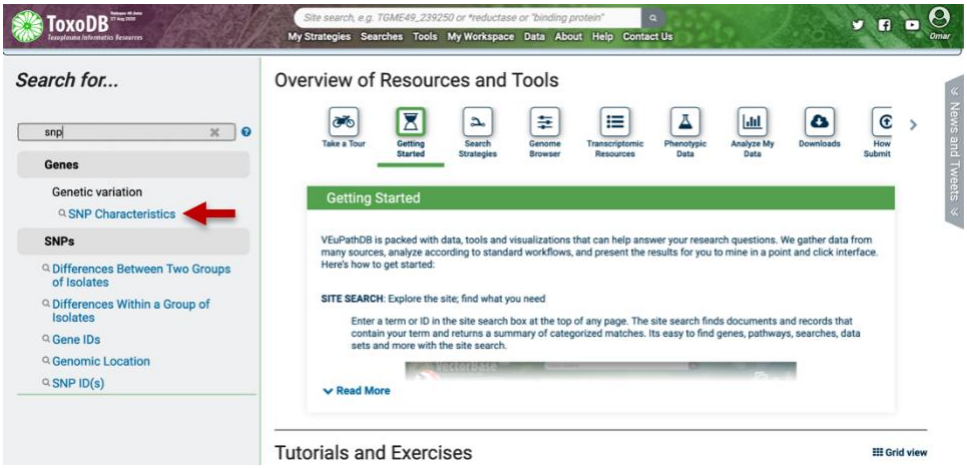
Learning Objective:

3. Run SNP searches in VEuPathDB
4. Explore SNP search parameters and their effect on search results
5. Use SNP searches to identify genes that are under diversifying or stabilizing selection
6. Run CNV searches in VEuPathDB
7. Explore CNV search parameters
8. Use CNV searches to identify regions of a genome that exhibit duplications or deletions.

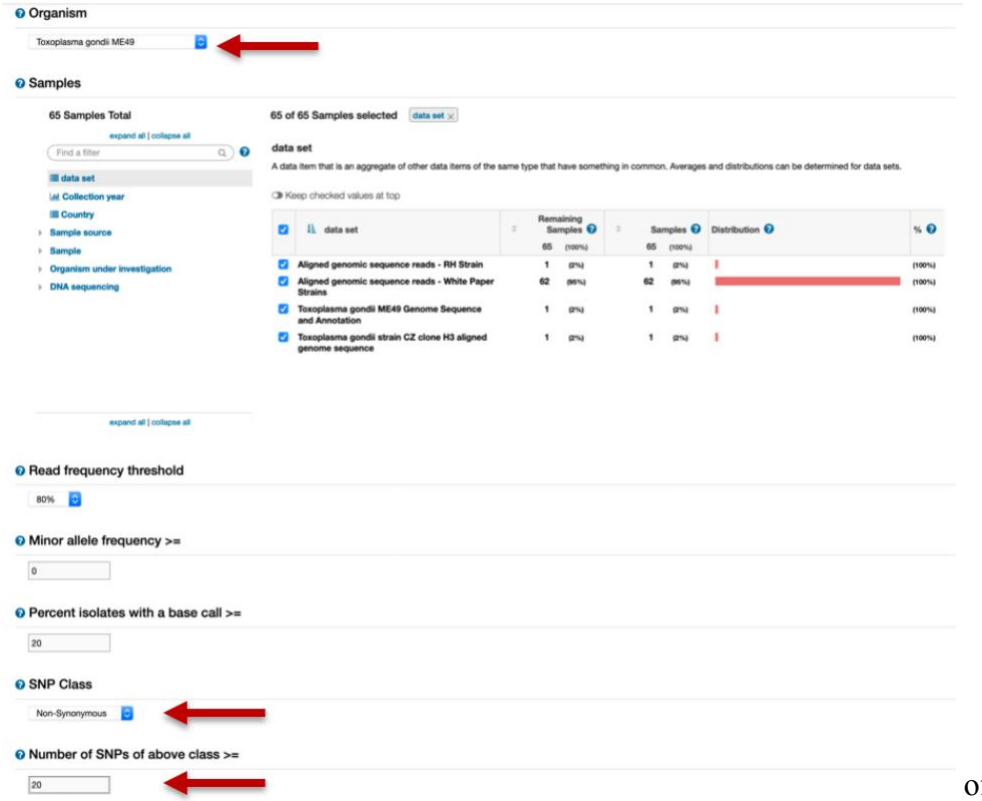
SNP Searches

In VEuPathDB SNPs can be used to characterize similarities and differences within a group of isolates or that distinguish between two groups of isolates. They can also be utilized to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection). Isolates are assayed for SNPs in VEuPathDB by two basic methods; re-sequencing and then alignment of sequence reads to a reference genome or DNA hybridization to a SNP-chip array (available in PlasmoDB only). In these exercises we'll explore both of these methods and ask a variety of questions to identify SNPs or genes of interest. If you do not understand the purpose of a parameter, please remember to mouse over the "?" icon and/or read the more detailed description at the bottom of the question page.

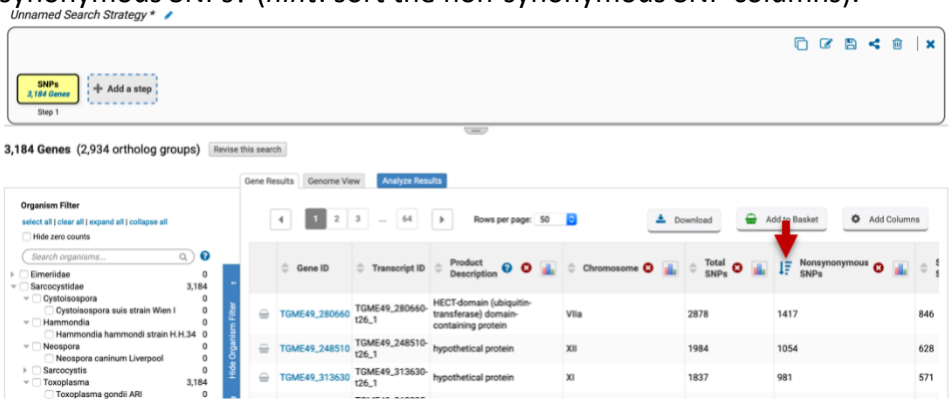
1. Identify *T. gondii* genes that contain at least 20 nonsynonymous SNPs.
   - ***a.*** Start by running a search for genes based on SNP characteristics – this search can be found under the 'Genetic Variation' category.



   - ***b.*** Select Toxoplasma gondii ME49 from the drop-down list. Notice how the sample information changes when you change organism.
   - ***c.*** In the sample section, select all available samples.
   - ***d.*** Change the SNP class to Non-synonymous and the 'number of SNPs of above class' field to 20.



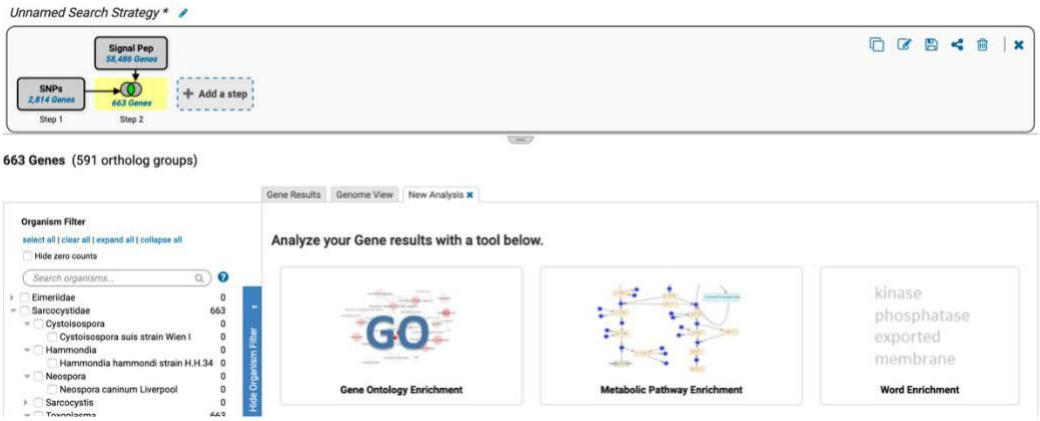
e. How many genes did you return? Which gene has the highest number of non-synonymous SNPs? (*hint*: sort the non-synonymous SNP columns).



f. What happens if you revise this search and change the "Percent isolates with a base call >=" field to 100?

g. How many of these genes have a predicted secretory signal peptide? (*hint*: add a step that identifies all genes with a signal peptide).

h. What kinds of genes are in this result list? One way to determine if you have anything enriched in your results is to run an enrichment analysis. Click on the "Analyze Results" tab then compare the results you get from the GO enrichment and from the Word enrichment, we will discuss these results.

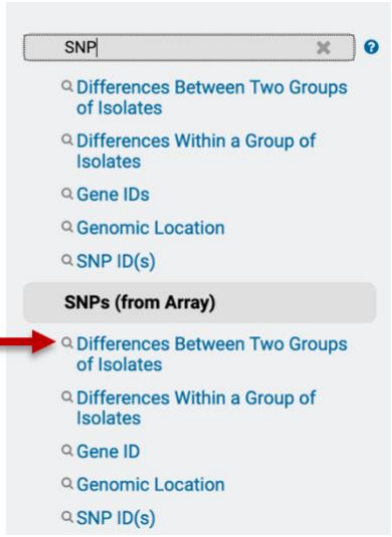


2. **Identify SNPs that distinguish parasites with rapid clearance times following treatment with the anti-malarial drug Artesunate vs. those that have delayed clearance times.** We have a published study in PlasmoDB (Takala-Harrison et. al.) with sufficient meta-data about the samples to ask this interesting question.

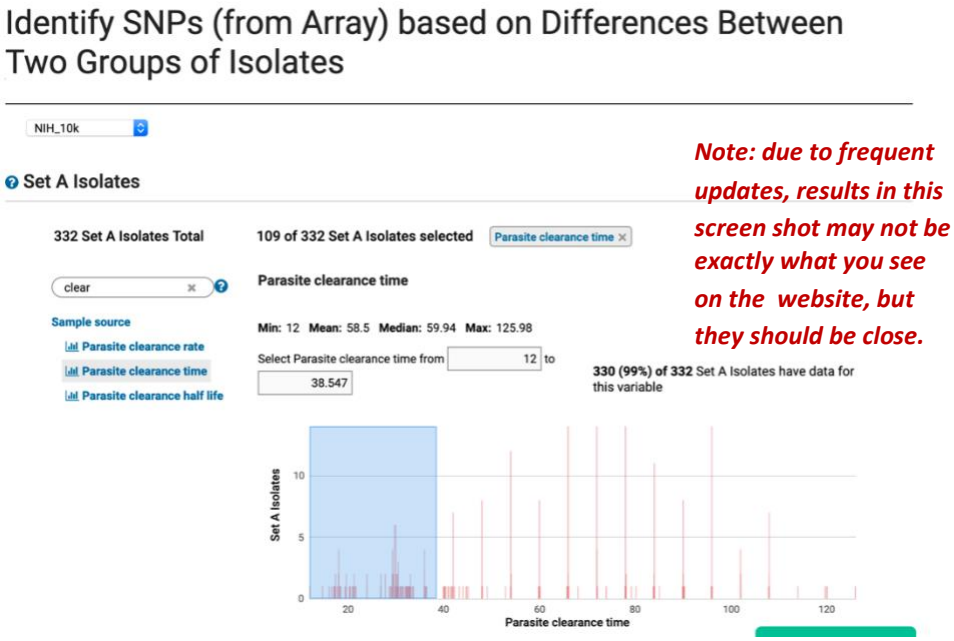
   For this exercise use http://PlasmoDB.org

Navigate to the "Differences between two groups of isolates" search under "Search for SNPs (from Array).

a. Unlike re-sequencing experiments that can identify any SNPs in the sequence, SNP-Chips have a pre-determined set of SNPs that are assayed and there are multiple different Chips on which these assays can be run. For this study, the authors used the NIH_10K Chip, an array with approximately 10,000 SNPs of which ~8000 can be assayed. Choose this in the Isolate assay type parameter.

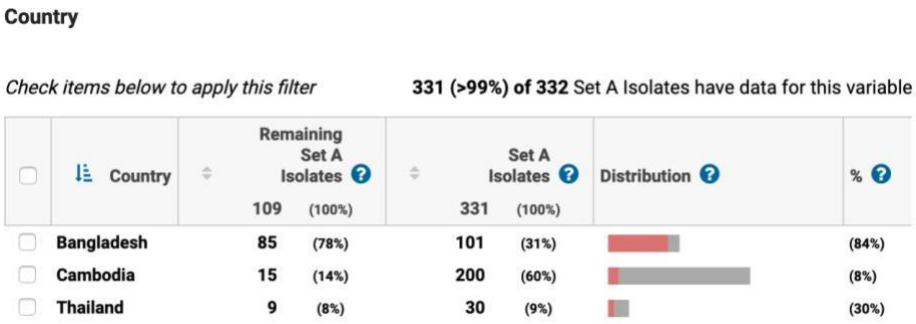


b. Once this is done, an interesting set of characteristics are seen in the parameters to choose isolates. In addition to geographic location, there are clinical parameters like Clearance Time, Parasitemia levels, etc.
   In this exercise we want to identify SNPs that distinguish parasites with rapid clearance times from those with delayed clearance times but you could try other possibilities once you are finished. In Set A Isolates, click on some of the characteristics to explore the data. Then choose

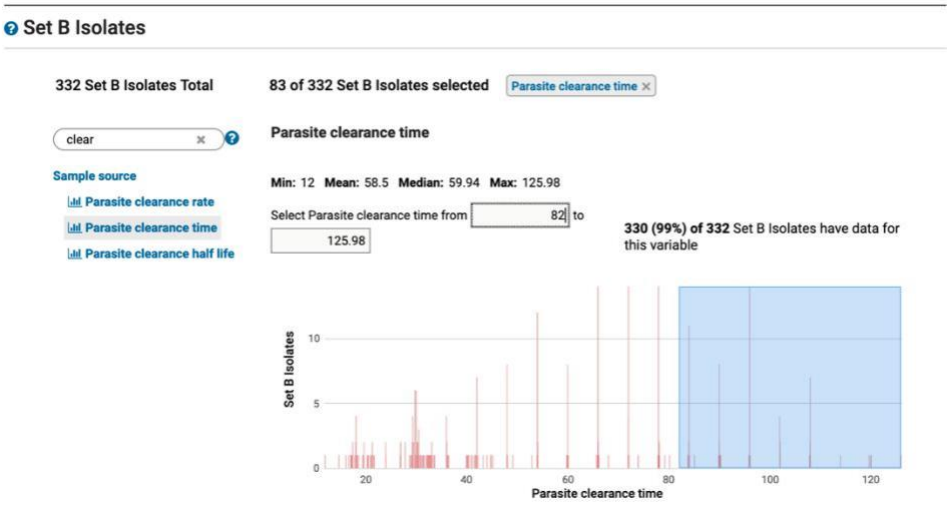


Note: due to frequent updates, results in this screen shot may not be exactly what you see on the website, but they should be close.

Clearance Time and select 0 – 38 or 39 minutes. Do these rapid clearance samples appear to be evenly distributed geographically? *Hint: click on Geographic Location to view the distribution of these selected samples (pink section of histogram).*



c. We'll keep the defaults of 80 for both Major Allele Frequency and Percent Isolates with Call for this exercise.
d. Now select Clearance times of 82 – end for Set B Isolates. Are these isolates geographically biased?



e. Keep defaults for Major Allele and Percent with call and run the search. How many SNPs did you find?

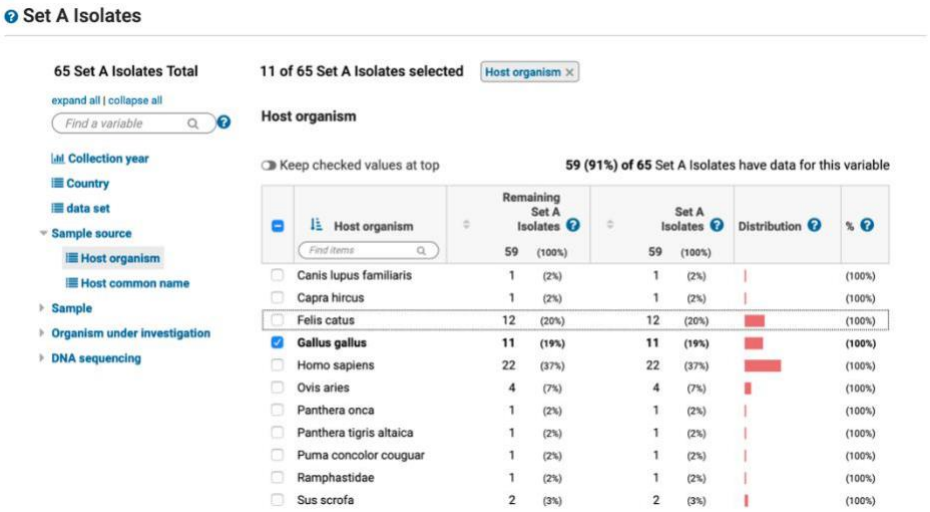
A gene (Kelch13) has been identified that is involved in Artemesinin resistance in South East Asia. Is one or more of your SNPs in the region (+/- 10 KB) of the kelch13 gene? Note that we are not expecting that the SNP would be within the gene as this is a Chip experiment where the SNPs were pre-determined and there may not be a SNP on the array within a particular gene that we care about. However, if there is a haplotype that is being selected for in the presence of artemesinin, any SNPs within that haplotype (region of the genome) should likewise be selected.

Hint: add a step to search for genes by text and search for kelch13. This will require you to use the genomic co-location operation as outlined in exercise 3. Set it up the same way except choose custom and start – 10000, stop + 10000 to define the region.

3. **Find SNPs that distinguish *Toxoplasma gondii* strains isolated from chickens as compared to those isolated from cats. *NOTE: This exercise in ToxoDB explores the hypothesis that we can identify SNPs/genes involved in T. gondii host preference.***

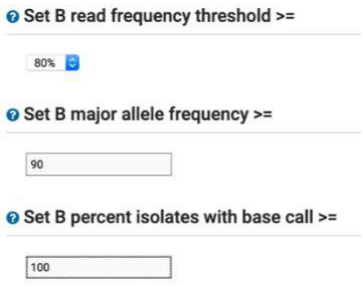
Navigate to "Identify SNPs based on Differences Between Two Groups of Isolates".

a. Click select set A isolates and select hosts from the left column. Check the chicken (*Gallus gallus*) box to select the 11 chicken isolates.



b. Click select set B isolates and select hosts from the left column. Check the cat (*Felis catus*) box to select the 12 cat isolates.

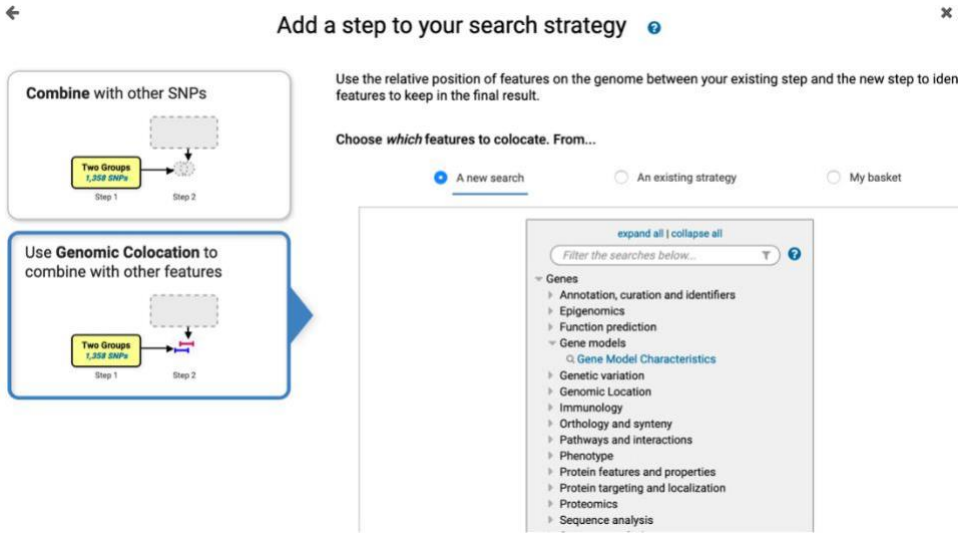
c. Let's run a very stringent search and change the "major allele frequency" parameters for both sets to 90. (*What does that mean?)*. Also, set the isolates with base call parameter to 100 for both sets A and B.



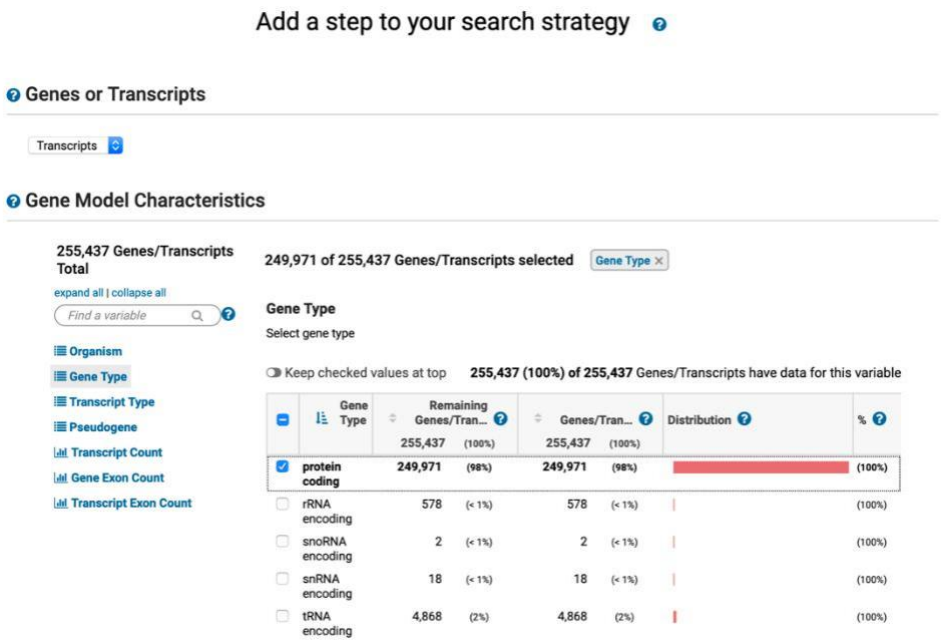
1. How many SNPs did your search return? Does this large number that distinguish these two fairly large groups of isolates surprise you?

You want to identify genes that could potentially be involved in host preference in *Toxoplasma gondii* and you expect that the SNPs from this search you just ran may be in protein coding regions of genes involved in this preference. How might you identify genes containing these SNPs?

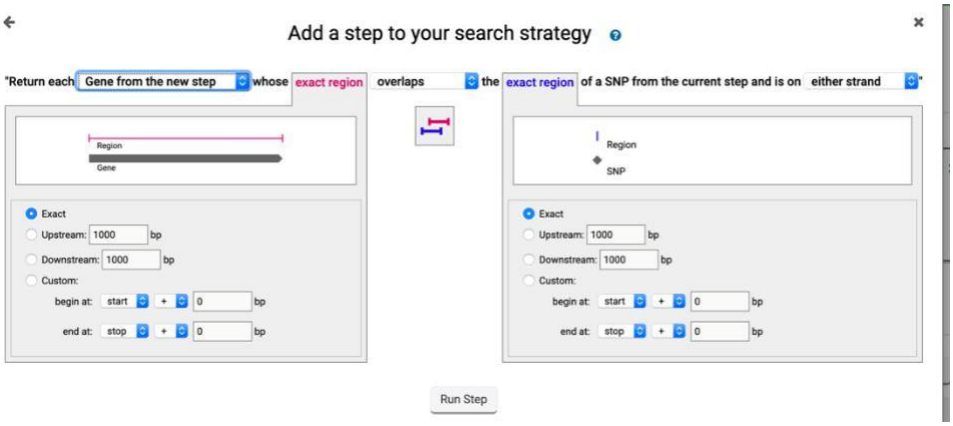
d. Add a step to identify *protein-coding genes* in *Toxoplasma gondii ME49*. Select the "Use Genomic Colocation…" option. Then select the gene search called "Gene Model Characteristics".



e. Configure the gene model characteristics search to find protein coding genes only.



f. Configure the genome colocation page to return "Gene from Step 2 whose exact region overlaps the exact region of a SNP in Step 1 and is on either strand"



- How many genes are returned?
- What is the gene that contains the most SNPs on your list? *Hint: sort the list high to low by match count.*
- Does this gene have orthologs in other species from ToxoDB? *Hint: go to the gene page and look at the genomic context and orthologs/paralogs in ToxoDB table.*
- Does it have orthology in any other species? *Hint: click on the link under the orthologs table and look at in OrthoMCL.*
- What does this say about this gene? How can you follow up on what role this gene may be playing for the organism? *Hint: you are a biologist and will need to look at the data on the gene record page and interpret it based on your experience and intuition.*
- Do these genes appear to be randomly distributed along the genome? *Hint: click the "Genome View" tab to view the distribution.* If you are a *Toxoplasma* biologist, do you have any hypotheses why the distribution may be skewed?

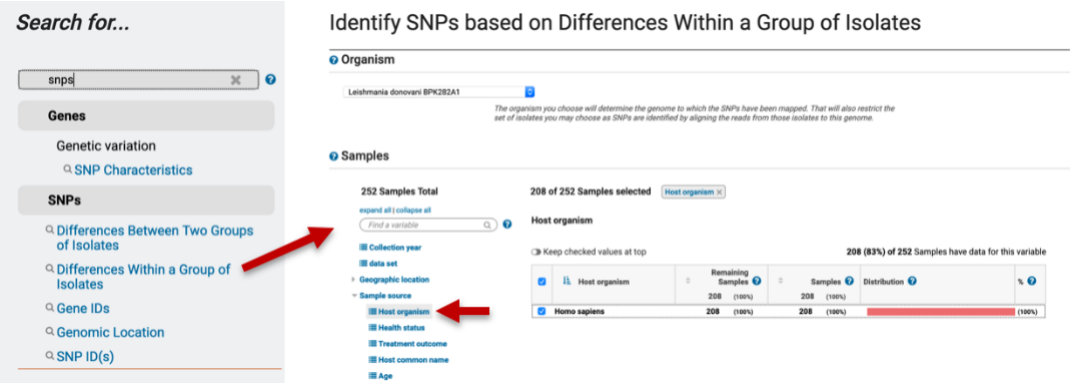
As a last resort: https://toxodb.org/toxo/im.do?s=4fe2f7409d4ba4d6

4. Identifying SNPs within a group of isolates

For this exercise use http://TriTrypDB.org

a. Go to the "Differences Within a Group of Isolates" search.

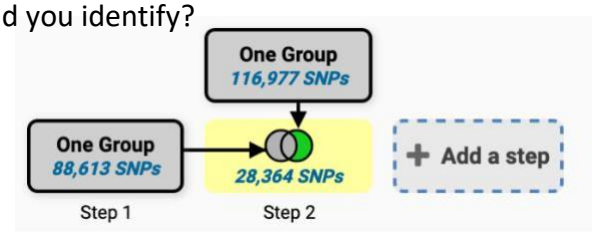
Hint: you can find this under the "SNPs" category (remember you can filter the searches by typing a key word like "snps" in the filter box.



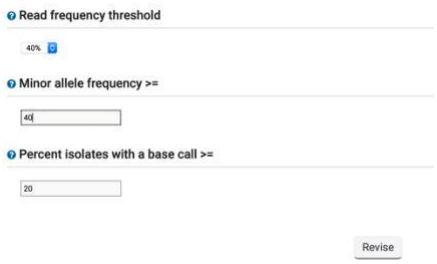
b. What does this search do? Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters.

Run the query and look at your results.

- How many SNPs were returned?
- Are any of these heterozygous SNPs?
- How would you identify heterozygous SNPs? Add a step to your strategy to identify SNPs from these isolates that may be heterozygous. *Hint: choose a read frequency threshold of 40% and select the 2 minus 1 operation.*
- How many SNPs did you identify?



- Click on the second step results to view them. What do you notice about the %minor alleles? (*many are quite low … i.e. in one or two of the isolates*). How can you remove these from your search results? *Hint: revise this search and increase the minor allele frequency threshold (try 20 and 40 and compare results).*



- Why might you want to increase the minor allele threshold when you run SNP searches?
- Try increasing / decreasing the “Percent isolates with base call”. How does this impact your results? Why might you want to change this parameter?
- Go to a record page for a SNP with a high minor allele frequency. What do you see in the Strains table? Why are many of the strains repeated?

CNV Searches

1. **Using resequencing data to identify regions of copy number variation (CNV)**

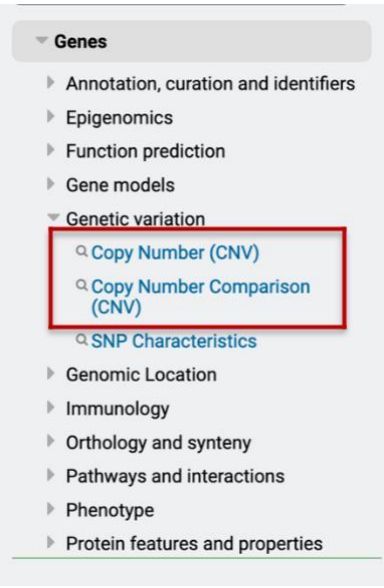
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in ToxoDB are mapped to the same reference strain ME49, as a result we can estimate a gene’s copy number in each of the aligned strains.

The goal of this exercise is to identify

Gene searches taking advantage of sequence alignment data can be found under the under the “Genetic Variation” category. Two available searches that define regions of CNV are:

Copy number: This search returns genes that are present at copy numbers (haploid number or gene dose) within a range that you specify.

Copy number comparison: This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor.

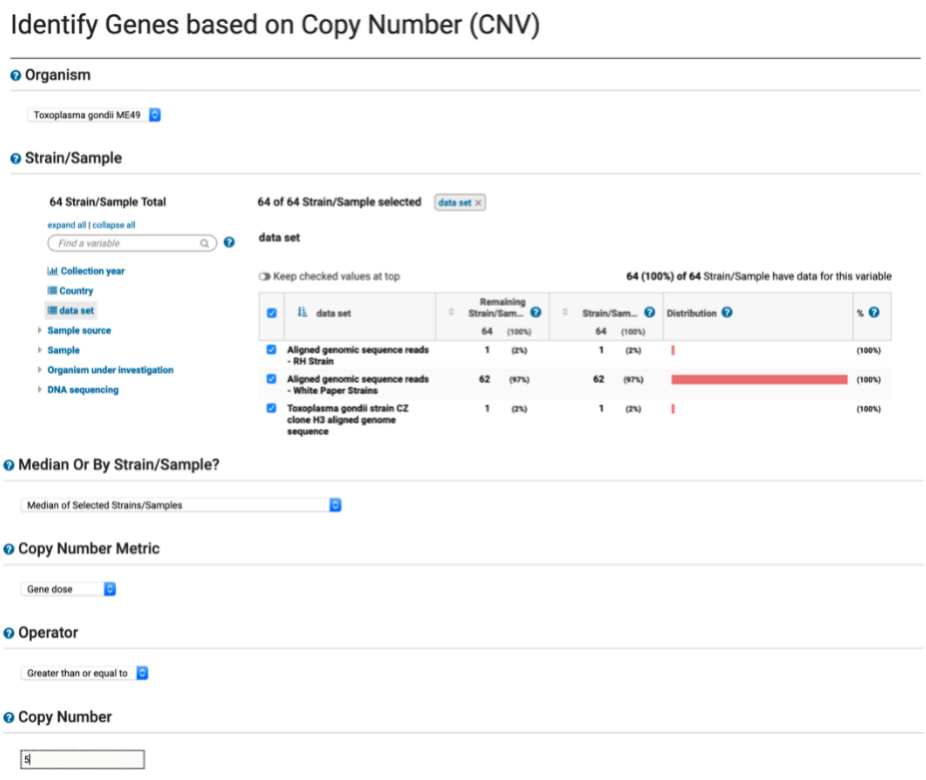


You have the choice between two different metrics for defining copy number: ***haploid number or gene dose***. Haploid number is the number of genes on an individual chromosome. Gene dose is the total number of genes in an organism, accounting for copy number of the chromosome.

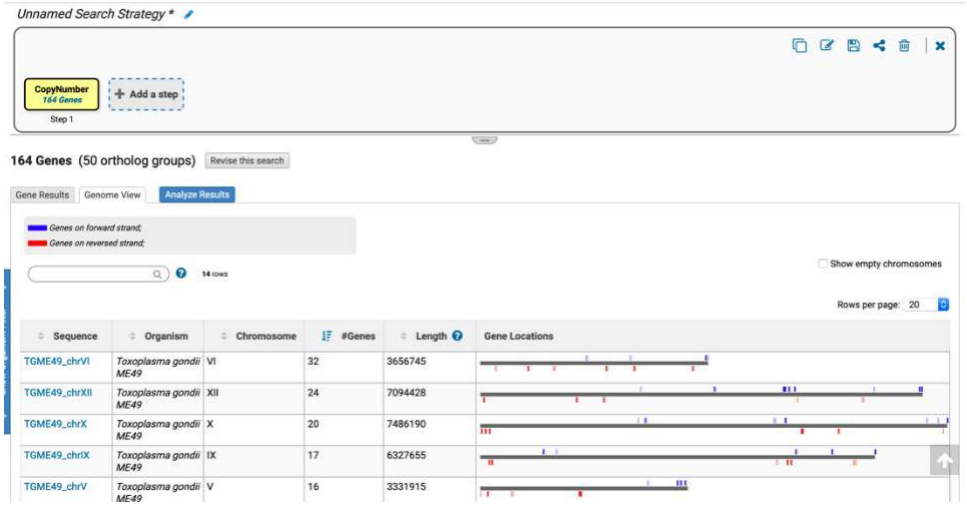
For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2. You can choose to search for genes where at least one of your selected isolates meets your cutoff criteria for the chosen metric (By Strain/Sample), or where the median of the chosen metric across all the selected isolates meets the cutoff (Median of Selected Strains/Samples)

Begin by choosing an Organism (reference genome) and one or more re-sequenced isolates. Choose whether you want to apply your search criteria to individual samples or to the median of your chosen samples. Then choose your Metric, Operator and Copy Number, and initiate the search by clicking the GET ANSWER button. Genes returned by the search will have a copy number based on your chosen metric within the range that you specified. For example, searching with the haploid number equal to 4 will return genes with 4 copies on a chromosome.

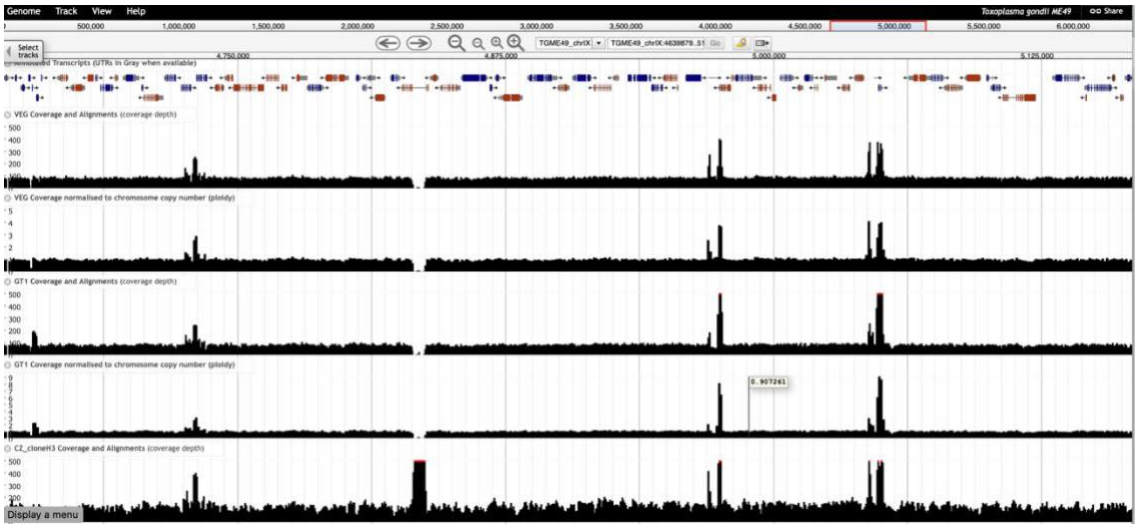
a. Use the copy number search to identify genes that are present at a copy number great than 5. Set up the copy number search to include all available isolates/strains, select the median of selected strains/samples, use Gene Dose for copy number metric and set the copy number to 5.



How many genes did you get? Are any of these genes clustered in the same location? (*hint*: click on the “Genome view” tab and examine the red and blue lines in the gene location column – wider lines indicate more than one gene in that location, click on the line to view what is there).



What happens if you edit this step and change the “Median Or By Strain/Sample” parameter to “By Strain/Sample (at least one selected strain/sample meets criteria)”? Do you get more or less genes? Which genes have the highest CNV? (*hint* : sort the median gene dose column from highest to lowest). Is this what you expected? Does the coverage of reads from resequenced strains aligned to the reference support this conclusion? Here is a link to a JBrowse view with some of the reseqeunced strain coverage data turned on: **https://tinyurl.com/y3mc53zm**



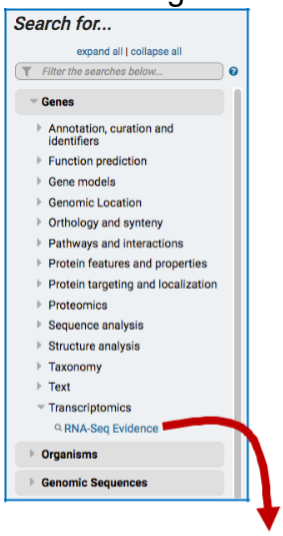
Host Response

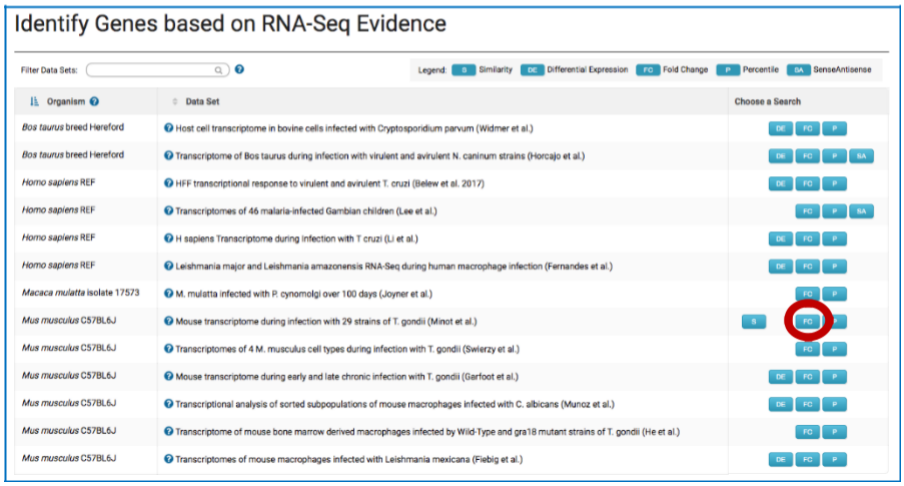
Learning objectives:

- Exploring host responses by running a search strategy in HostDB
- Add steps in a search strategy
- Performing a GO enrichment analysis
- Revising steps in a search strategy

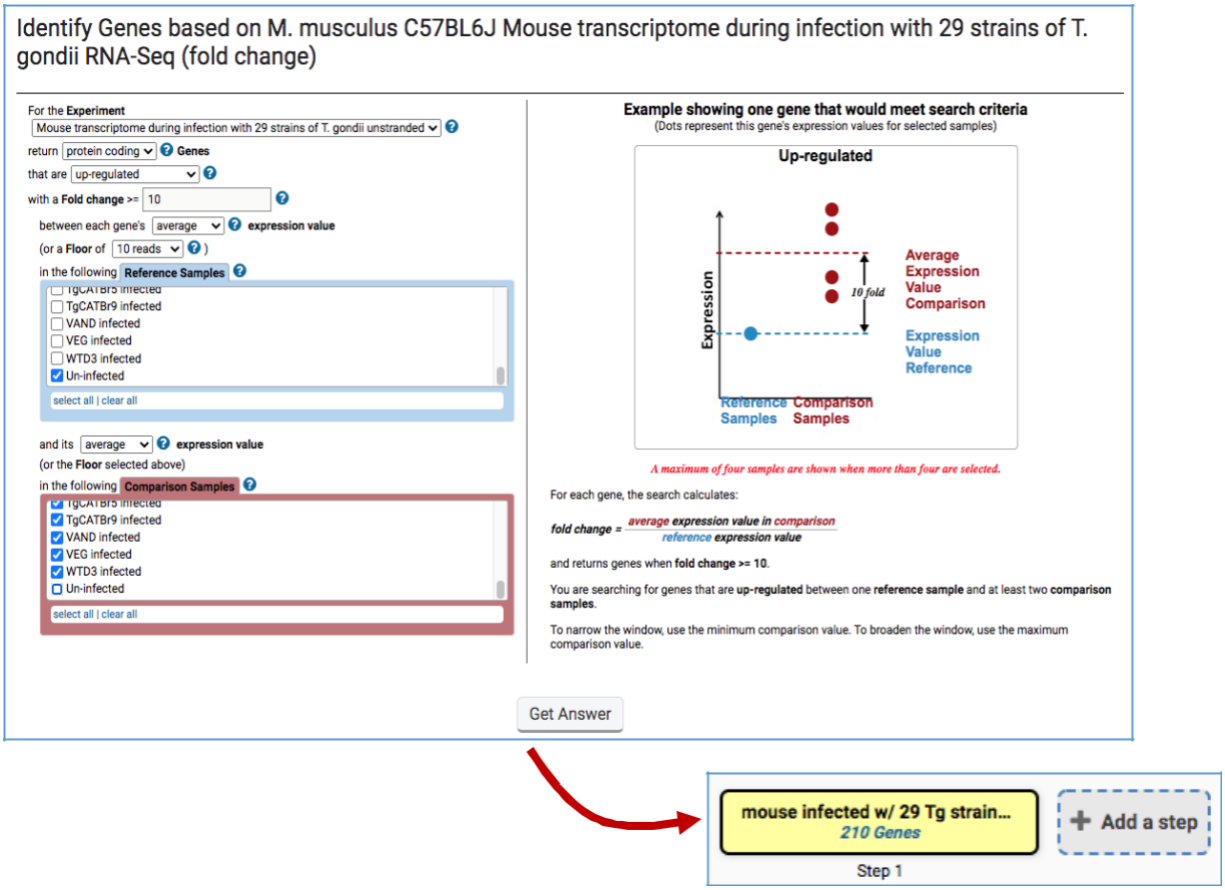
1. **Find host genes that are upregulated in infected mouse cells compared to uninfected ones. For this exercise use http://hostdb.org**

a. HostDB has data from a published study that performed a comparative transcriptome analysis of 29 different strains of *Toxoplasma gondii* and the murine macrophages infected with them. We loaded the parasite component of the data in ToxoDB and the host component in HostDB. Go to HostDB.org and navigate to the “Transcriptomics” section then select “RNA Seq Evidence”. Select the fold change query for the “Mouse transcriptomes during infection with 29 strains of T gondii (Minot et al.)” experiment.

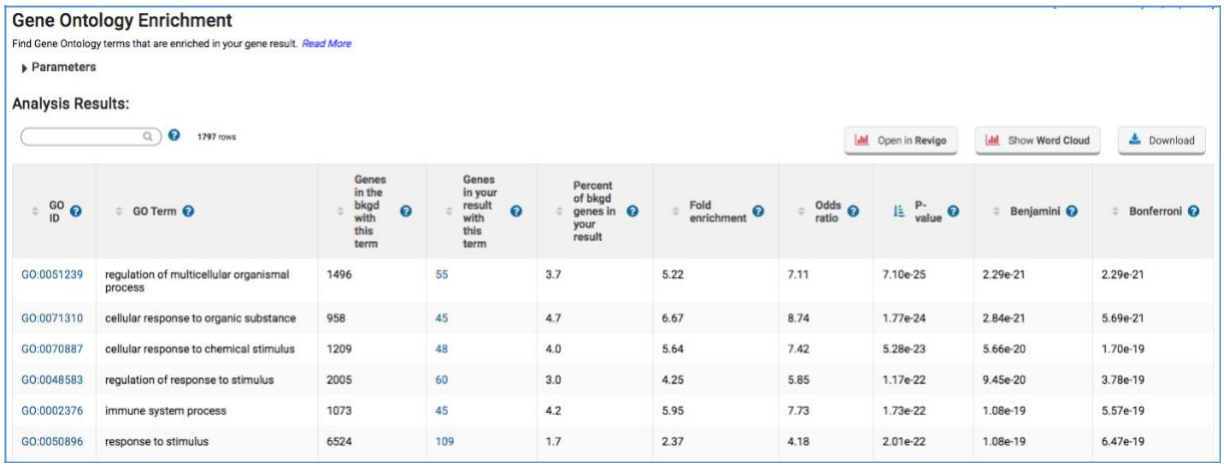




b. Configure the search to return genes that are up-regulated at least 10-fold across all strains in the experiment compared to the uninfected control. Make sure to select upregulated. In the example below a fold change of 10 was selected and the “average” operation was applied on the comparison samples.



c. What are the functional characteristics of the genes in this result? What kinds of GO terms are enriched? Does the host immune response appear to be turned on? Is there a particular cellular location that is common in this group of genes?
Hint: click on the “Analyze Results” tab and perform a GO enrichment analysis for the biological process ontology.



Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. *Read More*

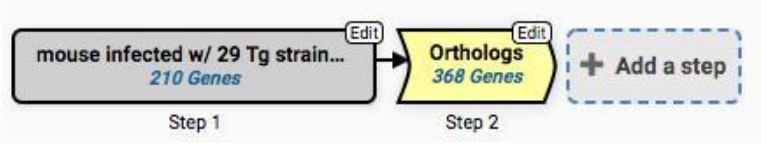
▸ Parameters

Analysis Results:

1797 rows — Open in Revigo | Show Word Cloud | Download

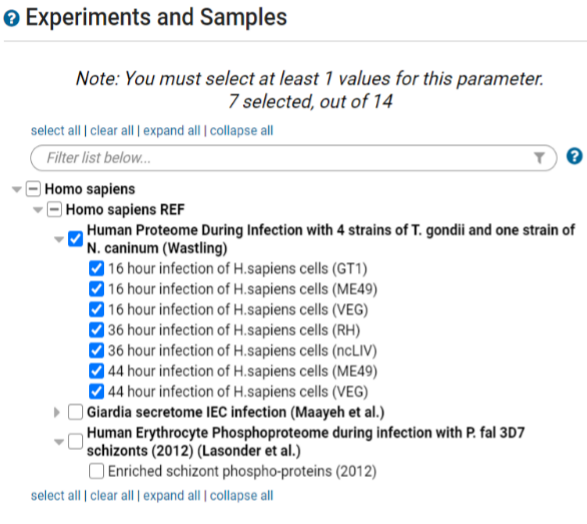
GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0051239	regulation of multicellular organismal process	1496	55	3.7	5.22	7.11	7.10e-25	2.29e-21	2.29e-21
GO:0071310	cellular response to organic substance	958	45	4.7	6.67	8.74	1.77e-24	2.84e-21	5.69e-21
GO:0070887	cellular response to chemical stimulus	1209	48	4.0	5.64	7.42	5.28e-23	5.66e-20	1.70e-19
GO:0048583	regulation of response to stimulus	2005	60	3.0	4.25	5.85	1.17e-22	9.45e-20	3.78e-19
GO:0002376	immune system process	1073	45	4.2	5.95	7.73	1.73e-22	1.08e-19	5.57e-19
GO:0050896	response to stimulus	6524	109	1.7	2.37	4.18	2.01e-22	1.08e-19	6.47e-19

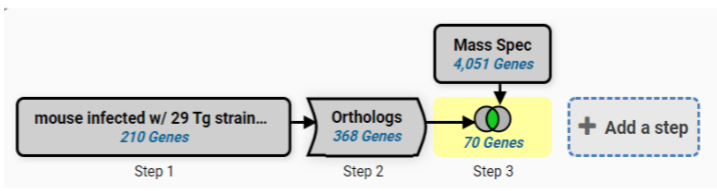
d. Expand the result set to include human orthologs/paralogs of these genes. *Hint:* add a “Transform by Orthology” step choosing Homo sapiens.



e. Does this set of human genes also show enriched GO terms? What, if any, are the enriched GO terms?

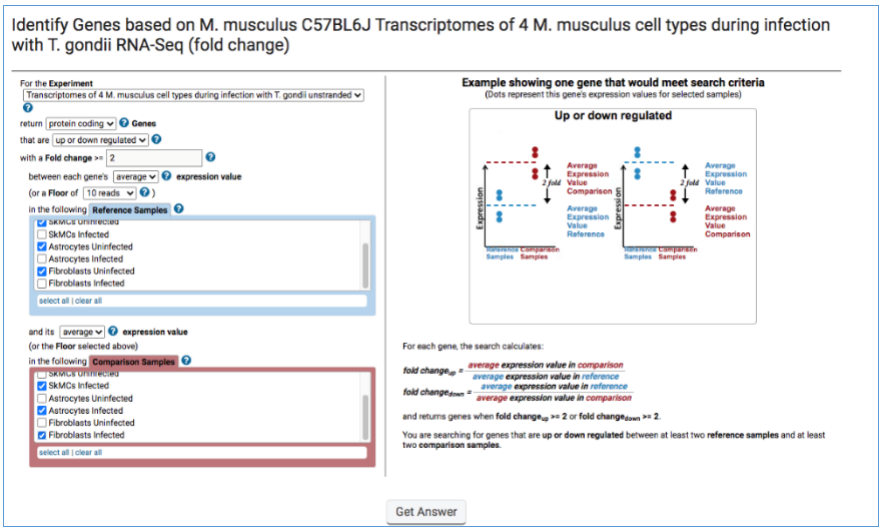
f. Do any of these human genes also have peptide evidence for their expression during infection? *Hint:* add a step and explore the proteomics data “Human Proteome During T. gondii infection”



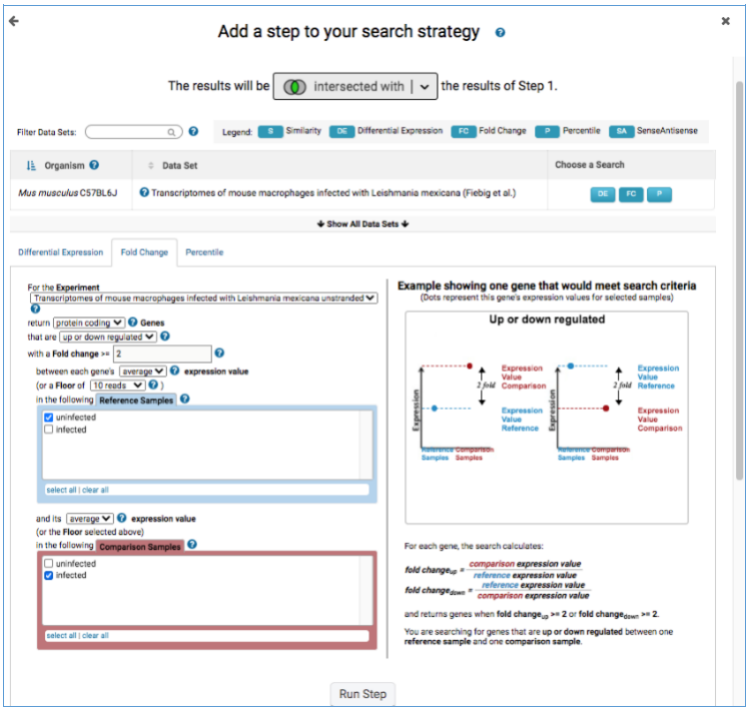


2. **Identify host genes that are differentially regulated in multiple infection models (e.g. different hosts and parasites).**

   Go to the RNAseq searches in HostDB. How many different pathogen infections are available? How many host organisms?
   Start by running a search using the “M. musculus C57BL6J Transcriptomes of 4 M. musculus cell types during infection with T. gondii (Swierzy et al.)”. Identify all genes that are differentially regulated (up or down) in all infected cell types compared to uninfected cell types.

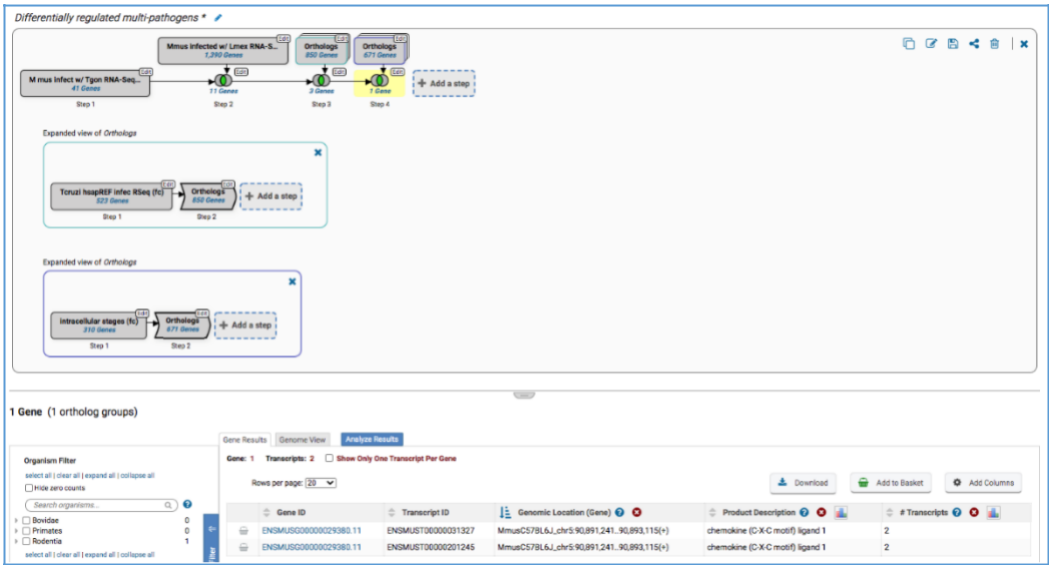


How many of these results are also differentially regulated in a *Leishmania* infection model? Try adding a step and running a search for genes that are differentially regulated in “M. musculus C57BL6J Transcriptomes of mouse macrophages infected with Leishmania mexicana (Fiebig et al.)”.



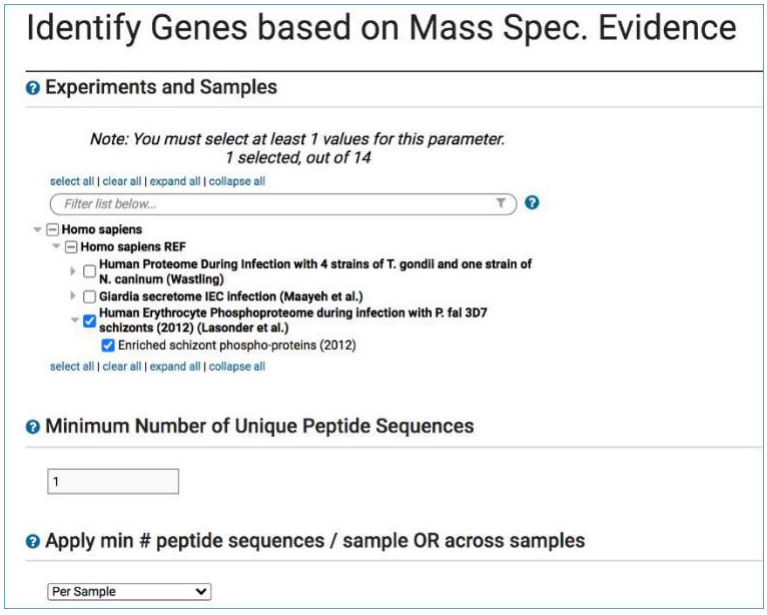
Continue adding steps from other pathogen infections. For example, try the "Host cell transcriptome in bovine cells infected with Cryptosporidium parvum". How many results in common did you get? If your answer is zero, did you remember to transform these results from bovine to mouse?

Here is an example search strategy to explore:
https://hostdb.org/hostdb/im.do?s=4d0a7299510641cf

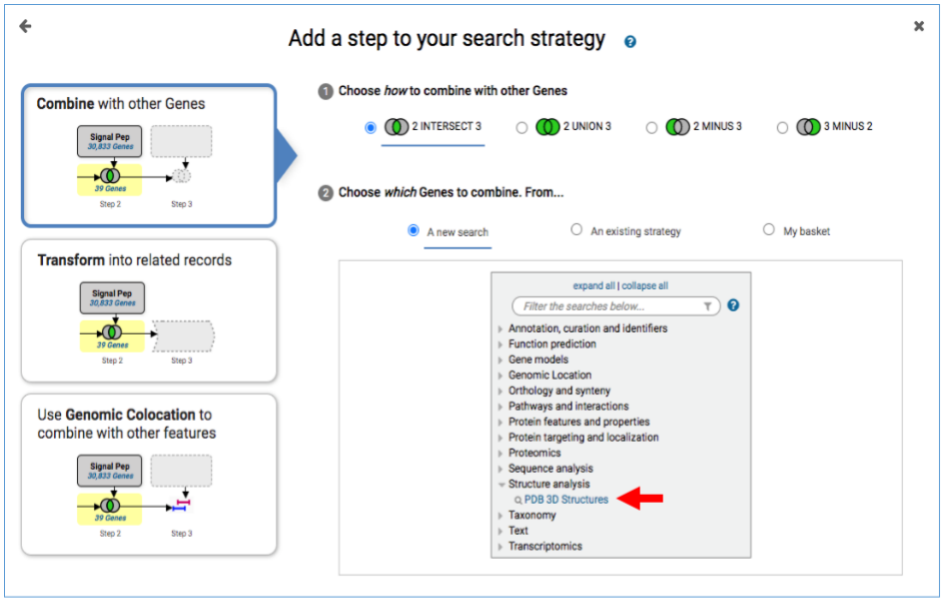


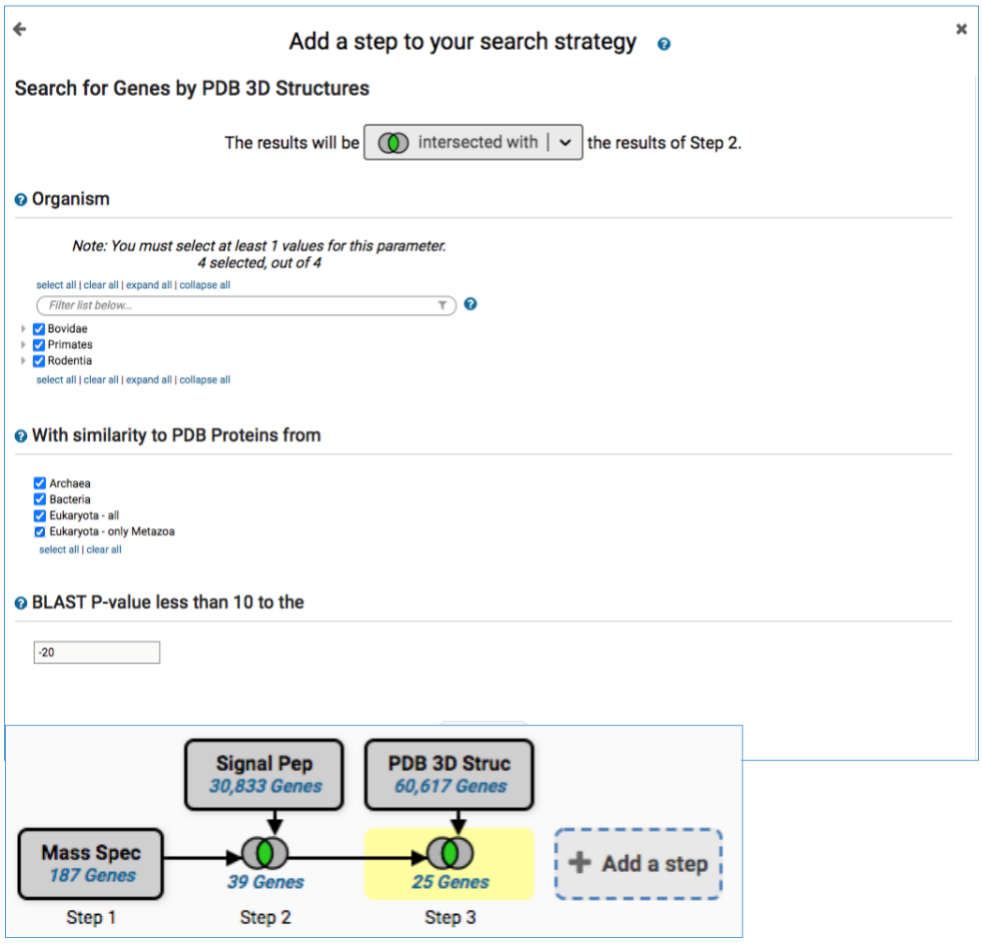
3. Identify host genes from a Plasmodium infection that are phosphorylated, secreted and have similarity to a 3D structure in the PDB database.

To do this you can start by looking at the available proteomics data. Can you find an experiment that identifies phosphoproteins?



How many of these gene are also predicted be secreted? To figure this out add a step and search for genes that have a secretory signal peptide. Using the same logic as above, add another step to identify any gene with similarity to any structure in the PDB database.





Metabolic Pathways - Exploring pathways and compounds

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

Learning objectives:

Explore the metabolic pathways searches and visualization tools
Search for a pathway using the name or pathway identifier
Paint data onto pathway maps to explore:
- Which enzymes in a pathway are present in different genera
- How transcriptional abundance of enzymes in a pathway differs under experimental conditions

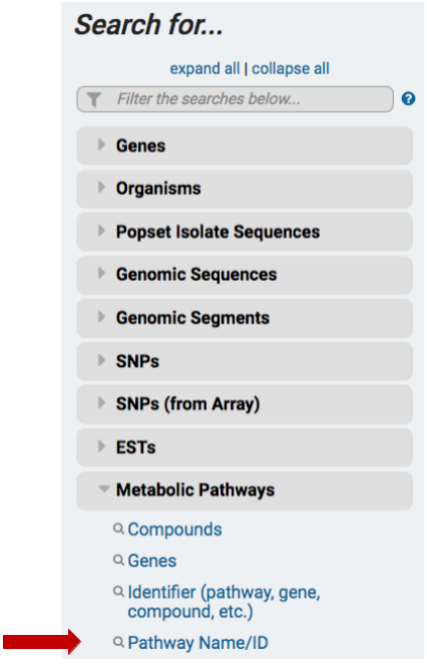
Explore the compound search options

1. **Find and explore the metabolic pathway for glycolysis.**
   For this exercise use http://plasmodb.org

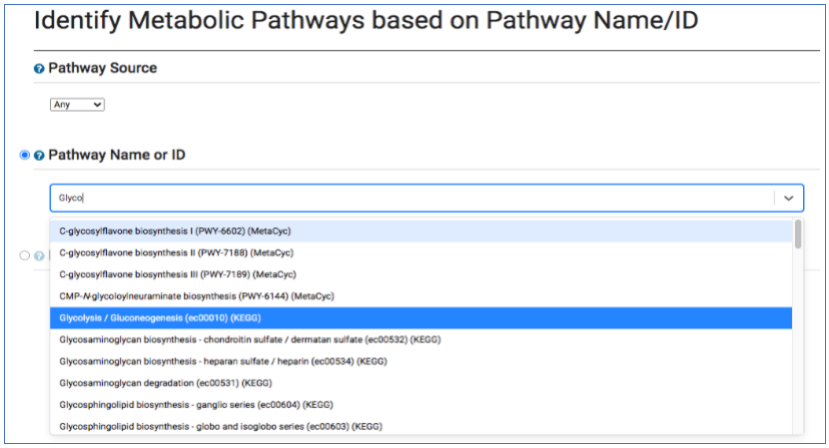
   Navigate to the search page for Identify Metabolic Pathways based on Pathway Name/ID.

   Find the metabolic pathway searches on the home page. You can look under “Metabolic Pathways” or use the search filter. You can find metabolic pathways based on the pathway name or identifier, or using genes or compounds involved in the pathway. Search for the **glycolysis** pathway using the Pathway Name/ID option.

   This search is equipped with a type-ahead function for finding the metabolic pathway name. Begin typing glycolysis and then choose the pathway name from the list that appears.

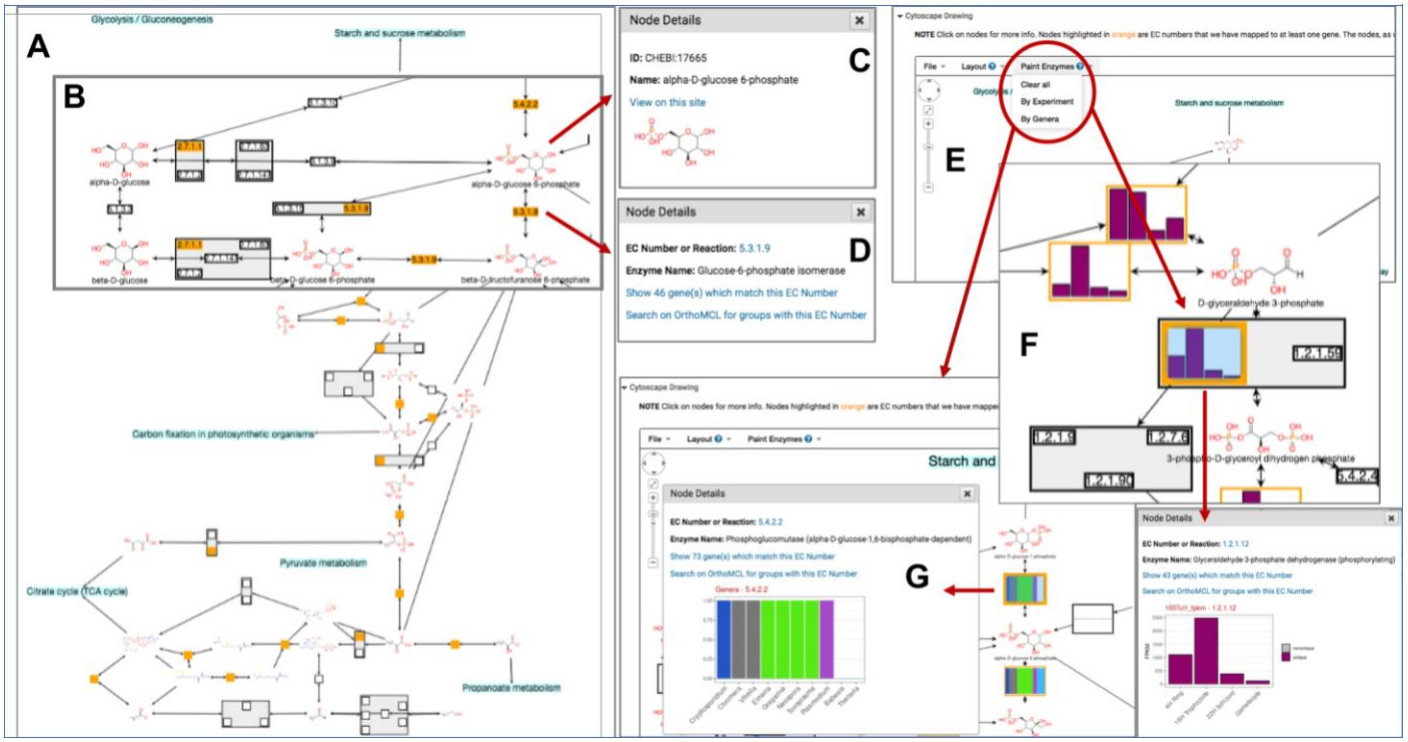


a. Examine the Glycolysis / Gluconeogenesis pathway.

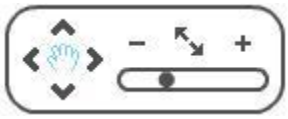


The search takes you straight to the record page for the Glycolysis / Gluconeogenesis (ec00010) metabolic pathway from KEGG. The overview section of the record page contains an interactive graphical representation of the pathway. The pathway map and the legend can be repositioned.

A. Initial pathway view is zoomed out.
B. Zoom in to see more details including EC numbers and metabolite structures.
C. Click on a metabolite structure to get additional information.
D. Click on the EC number to get more info about the enzyme including links to retrieve all genes in the database assigned to this EC number.

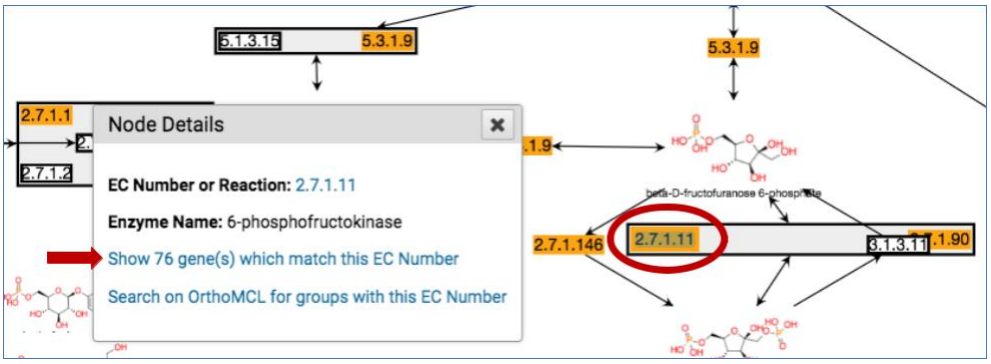


E. The drop-down menu under the heading “Paint Enzymes” allows you paint the pathway based on experimental data or phyletic pattern.
F. Painting pathway by experiment provides a graphical representation of experimental results. Click on the graph to see more details.
G. Painting pathway based on phyletic pattern provides a graphical representation of phyletic distribution. Clicking on the phyletic pattern graphic provides additional information.



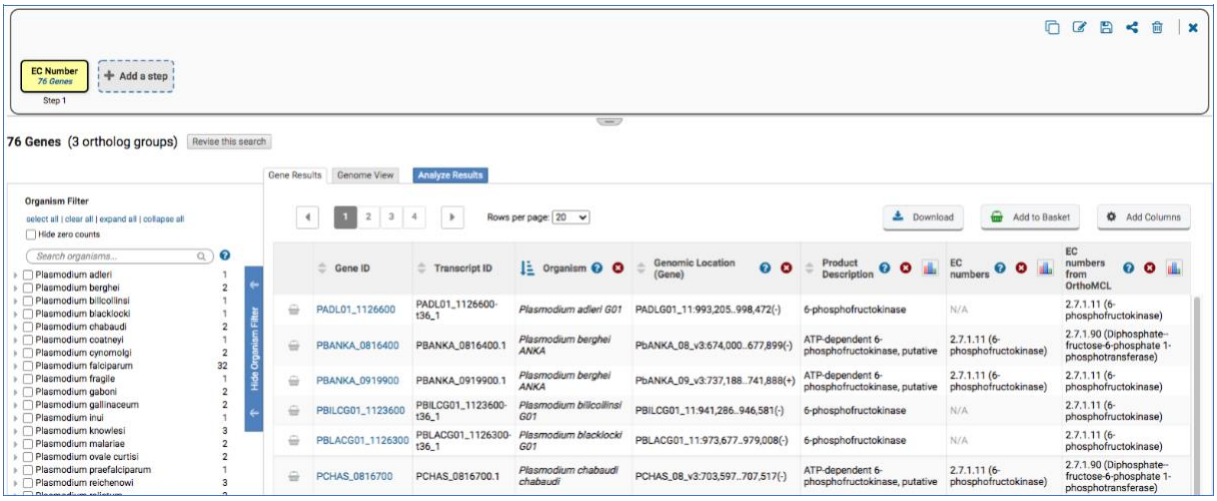
Use the Tool Box to move (drag) the map and individual nodes. Zoom in and out to help explore the map.
What do the rectangles with numbers like 2.7.1.11 represent?
What is the difference between the rectangular nodes that are orange and those that are not?
Why are some enzymes grouped?
Find the node representing 6-phosphofructokinase (EC number = 2.7.1.11). You may need to zoom and reposition the map to find the node.

Click on the 2.7.1.11 node to open a popup with information about this enzyme.



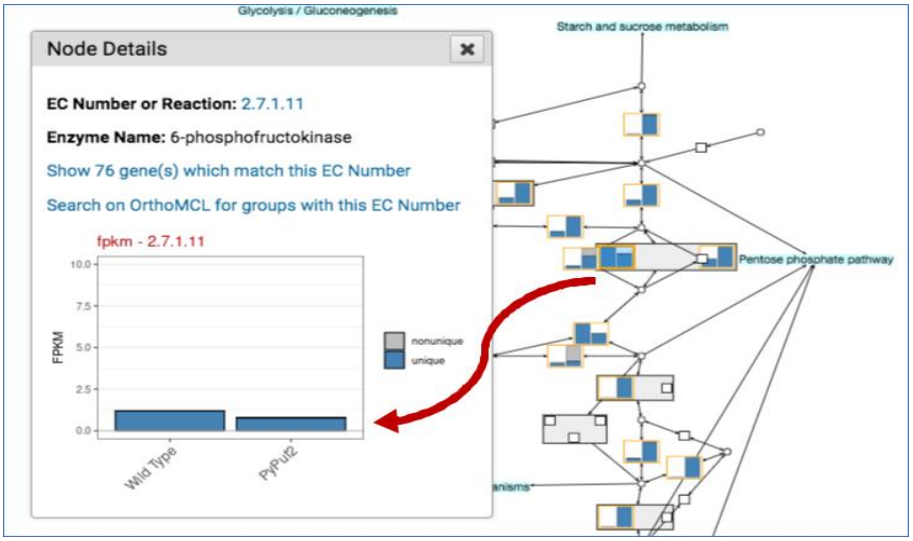
How many genes in the database matched this EC number?

Try the link ‘Search for Gene(s) by EC Number’. Where did you end up? What do the 76 genes in the result list represent? Is 6-phosphofructokinase unique to P. falciparum? Notice the two columns called “EC numbers” and “EC numbers from OrthoMCL”. What do these columns represent?

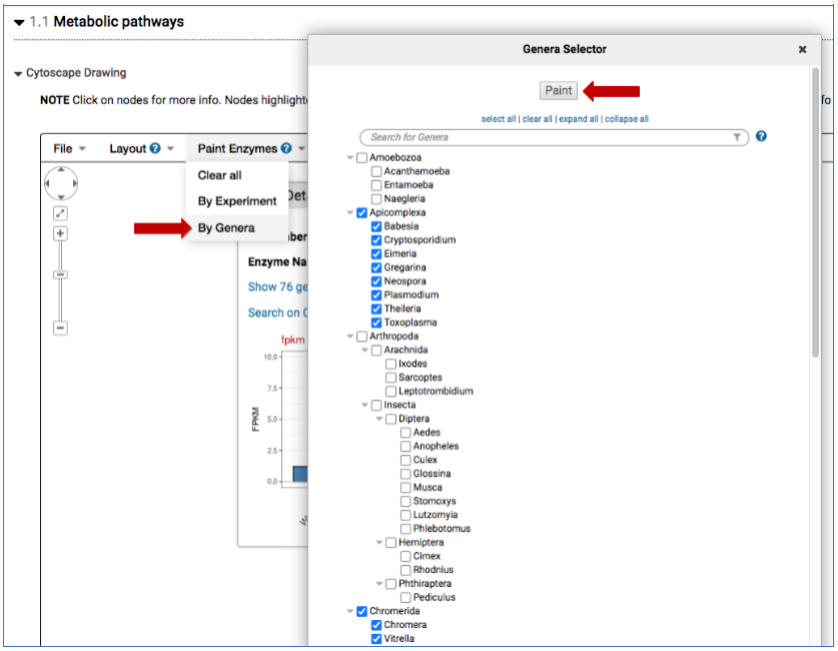


Use your Browser's back button to return to the glycolysis pathway record page and open the Paint Experiment menu. Choose the experiment "Salivary gland sporozoite transcriptomes: WT vs Puf2-KO (Lindner et al)". Be patient while the graphs appear in place of the EC numbers.

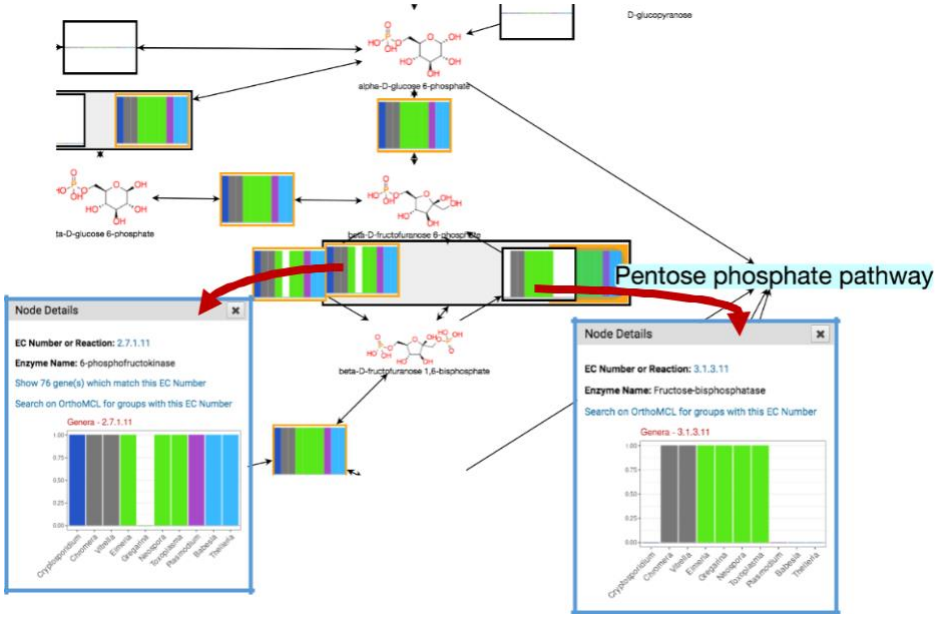
Does 6-phosphofructokinase appear to be expressed in salivary gland sporozoites? What enzymes in this pathway are affected in knockouts of Puf2?



Use the Paint Genera option to determine whether 6-phosphofructokinase has orthologs across Apicomplexa and Chromerida.

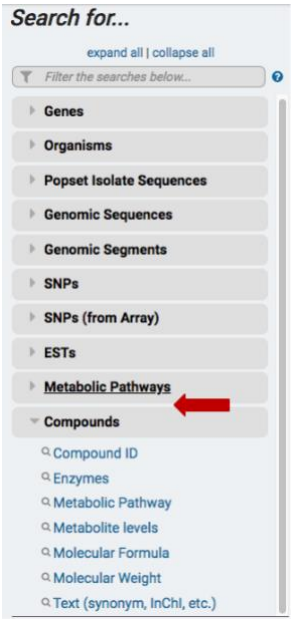


- What about the enzyme that catalyzes the reverse reaction (Fructose-bisphosphatase)?



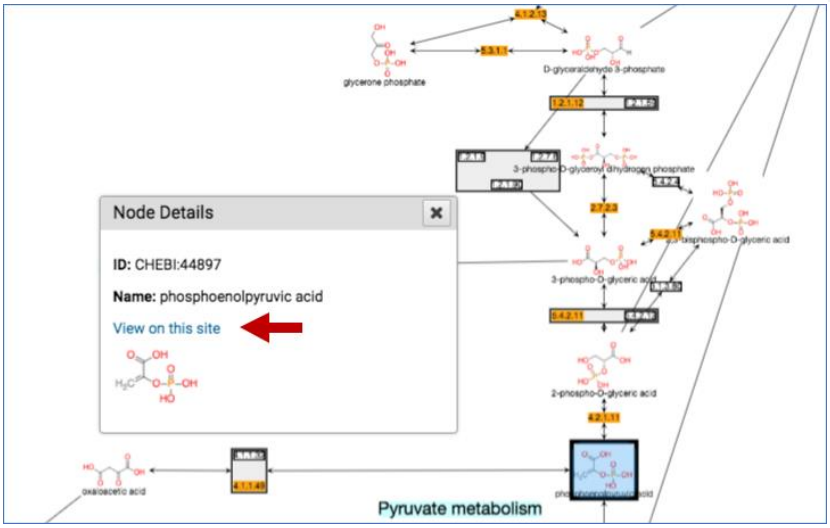
2. Find and explore the compound record page for phosphoenolpyruvate (phosphoenolpyruvic acid or PEP).

Compound records are accessed by running one of the compound searches available under the “Compounds” heading. Compounds may be retrieved by ID, text, metabolic pathway, molecular formula, molecular weight and metabolite levels. Compound records can also be accessed from the metabolic pathway legend after clicking on a compound (blue circle) in the map.



Choose one of these searches and retrieve the PEP record page.

Alternatively, you can reach the PEP record page via a metabolic pathway where it is present as a substrate or a product of an enzymatic reaction (ie. glycolysis). Click on the node representing a compound

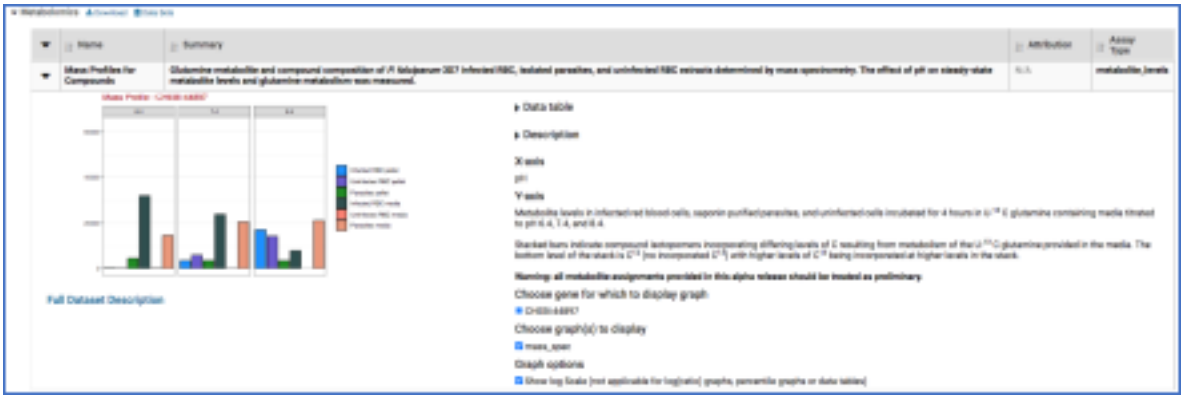


Which method did you use to get to the PEP record page? What compound name worked the best?

Examine the PEP record page.

What data sections do you see?

Under which conditions is PEP present at highest concentrations? (Hint: navigate to the Metabolomics section)

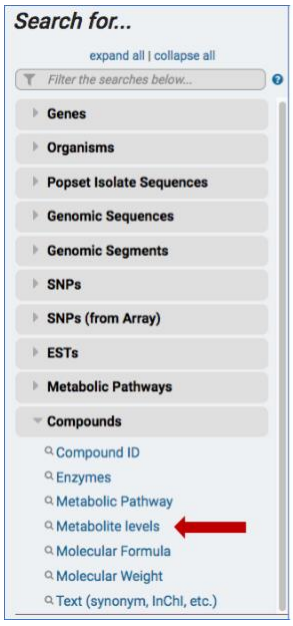


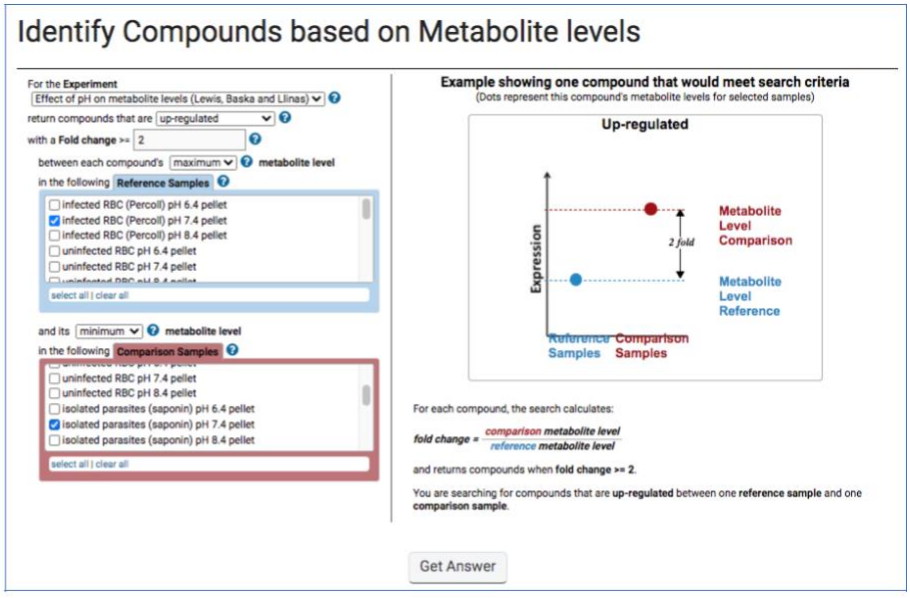
3. **The metabolite abundance experiment in PlasmoDB compares the following conditions at 3 pH levels:**

a. Parasites isolated from infected red blood cells using saponin lysis
b. Whole infected red blood cells isolated with Percoll
c. Whole uninfected red blood cells
d. For both conditions, data was collected from the cell pellet and the media supernatant.

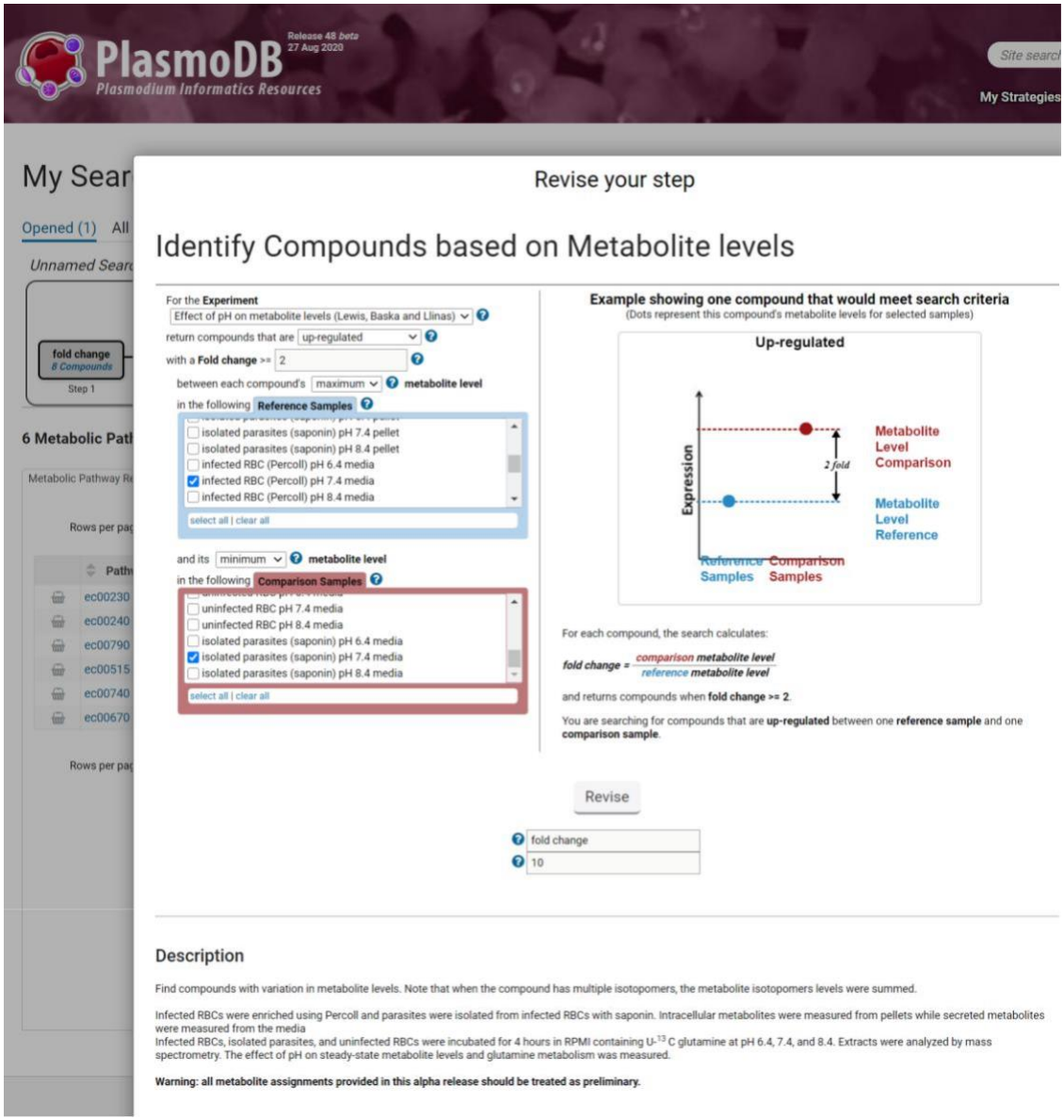
Find metabolites that are enriched in the isolated parasites (saponin) compared to infected red blood cells (Percoll) in the cell pellet at pH 7.4.

This can be done using the metabolite levels search, which looks a lot like the fold-change searches you have previously seen for transcriptomics data.

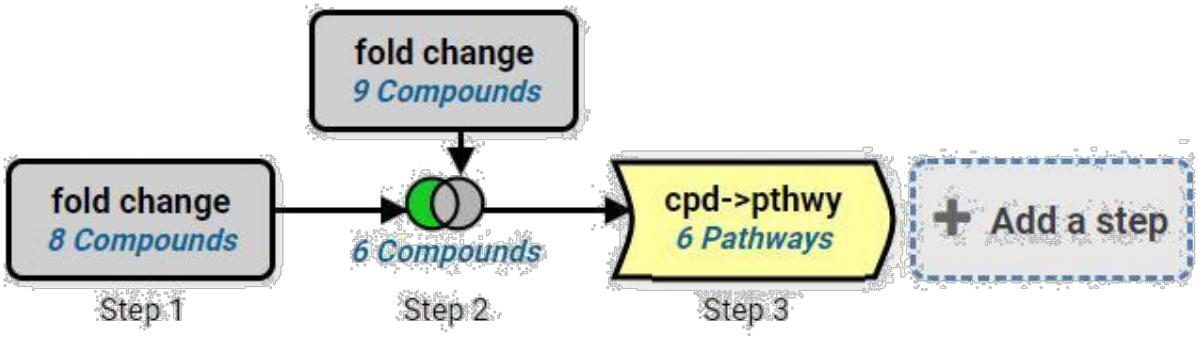




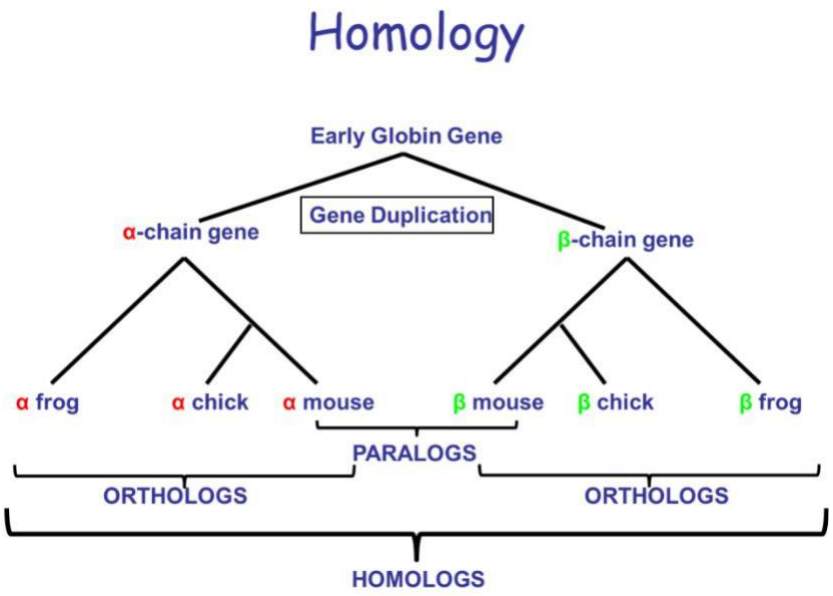
How many compounds did you get?
Add a step and use the same search to find out how many of these compounds (metabolites) are **NOT** enriched by 2-fold in isolated parasites (saponin) compared to the infected red blood cells (Percoll) in the media supernatant at pH 7.4. Make sure to use the correct operator!



How many compounds do you have now? Which metabolic pathways do these compounds belong to? Click Add a Step and transform the results to metabolic pathways.


fold change
9 Compounds
fold change
8 Compounds
6 Compounds
cpd->pthwy
6 Pathways
Add a step
Step 1
Step 2
Step 3

Orthology and Phyletic Patterns



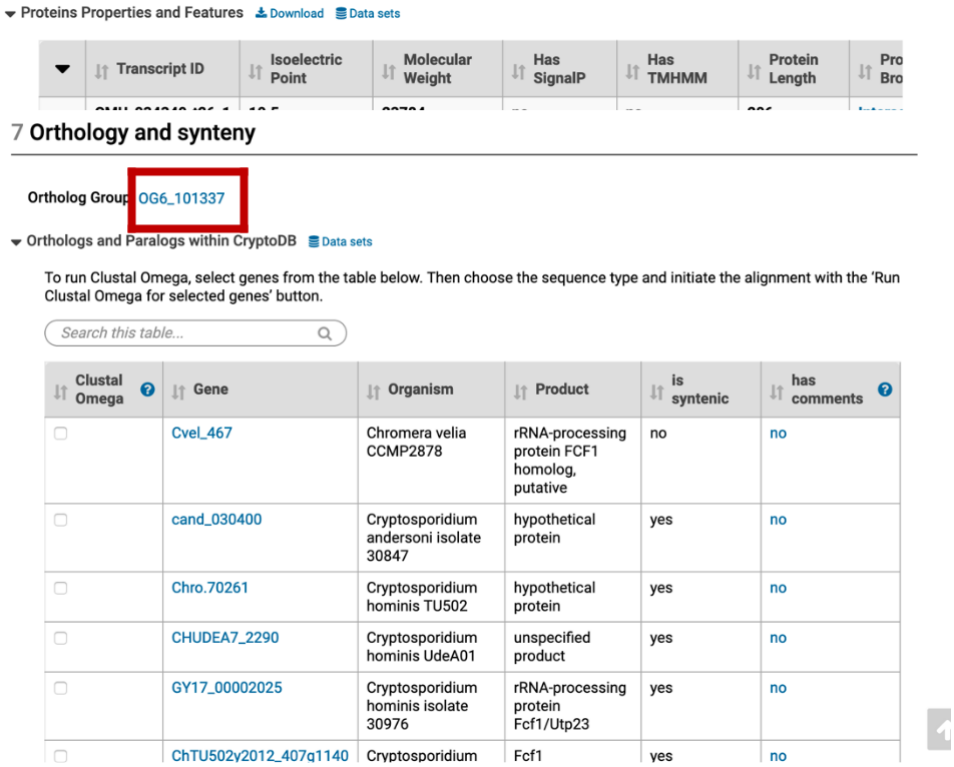
Learning objectives:

Explore the orthology table on VEuPathDB gene pages
Getting to OrthoMCL from VEuPathDB gene pages
Run searches in OrthoMCL
Explore the cluster graphs in OrthoMCL
Leverage the phyletic pattern search
Leverage the orthology transform tool

1. **Getting to OrthoMCL from VEuPathDB databases**
   Note: For this exercise use http://cryptodb.org and http://orthomcl.org/

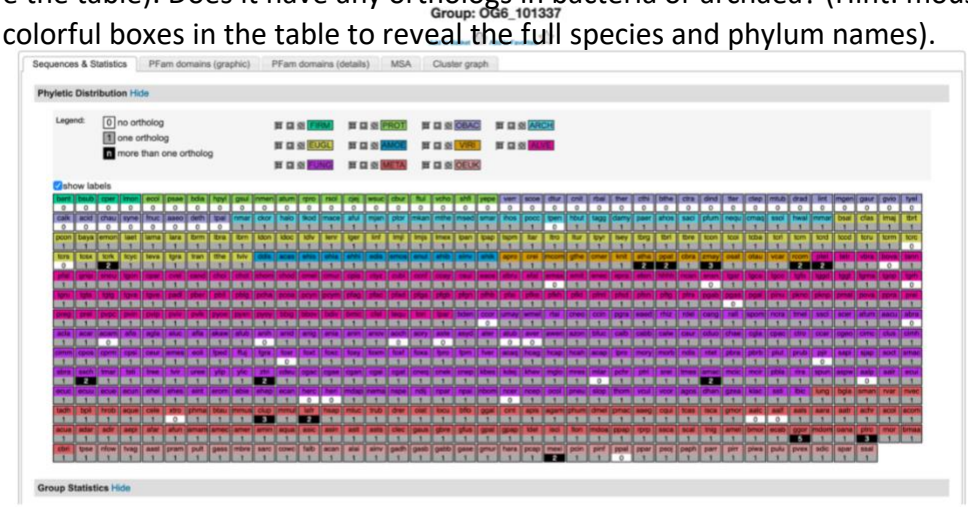
a. Go to the gene page for the *Cryptosporidium muris* gene with the ID: CMU_034340

b. What information on the gene page can you use to guess a function for this gene? It is annotated as a hypothetical protein! Hint: look at the orthologs table and the domains in the protein features graph. You may also want to visit some of the external links or take a look at InterPro domains.



c. Go to the Orthology and Synteny section and look at the table labeled "Orthologs and Paralogs within CryptoDB". Does this gene have orthologs in other *Cryptosporidium* species? What about other organisms? (hint: click on the Ortholog Group link above the table).

d. What about orthologs in organisms not in VEuPathDB? (hint: click on the Ortholog Group link above the table). Does it have any orthologs in bacteria or archaea? (Hint: mouse over the colorful boxes in the table to reveal the full species and phylum names).

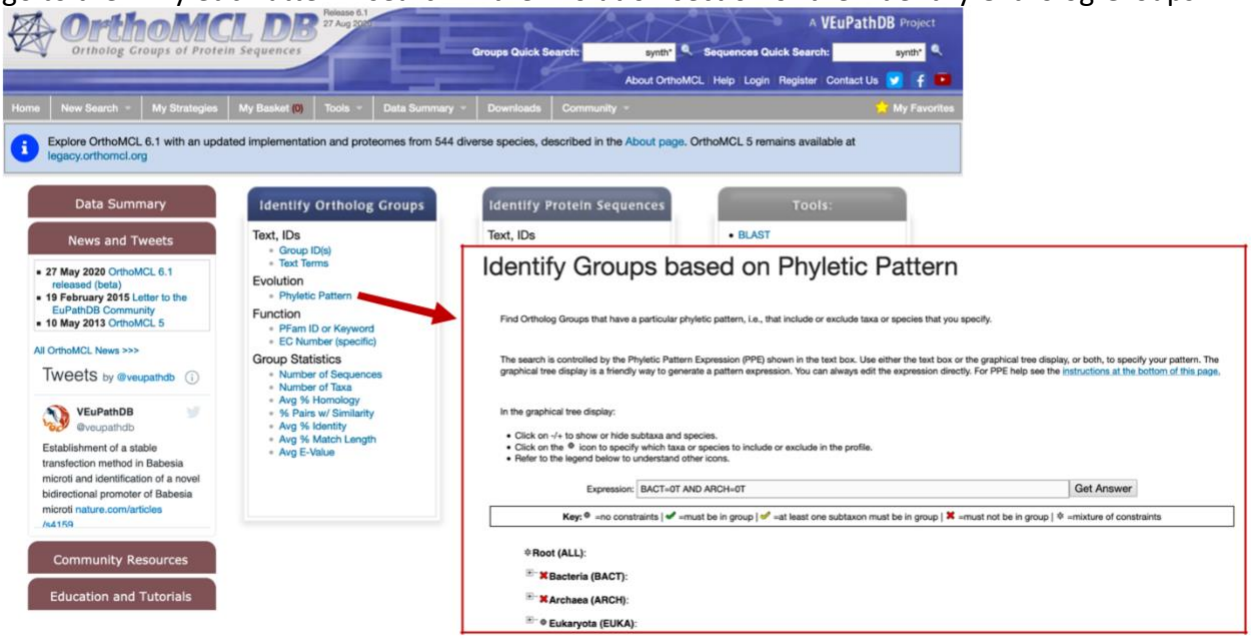


e. Look at the PFAM domain architectures found under the PFam domains (graphic) tab. Do all the proteins in this group have similar domain architecture?

f. Based on the orthologs, what do you think this protein might be doing? If you had to give this gene a name, what would you call it?

2. Using the phyletic pattern tool in OrthoMCL
Note: For this exercise use http://orthomcl.org/

How many protein groups in OrthoMCL do not have any orthologs in bacteria or archaea? (Hint: go to the "Phyletic Pattern" search in the Evolution section of the "Identify Ortholog Groups"

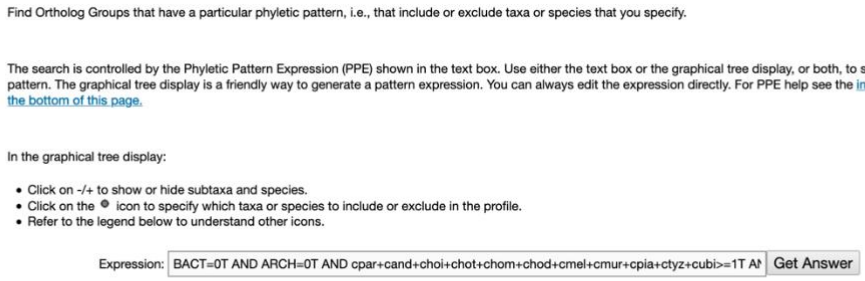


category). To specify a phyletic pattern click on the icon next to the taxonomic group or species to include or exclude it.

a. How many protein groups do not contain orthologs from bacteria and archaea?

b. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea. If you are getting frustrated trying to figure this one out, you have a right to be! You cannot answer this question by using the check boxes (we will discuss why). However, OrthoMCL has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Can you figure out what expression to use to answer this question? (hint: scroll down to the bottom of the page to find additional information about expression parameters.

Before looking at the answer below, try this on your own or with the people sitting next to you.



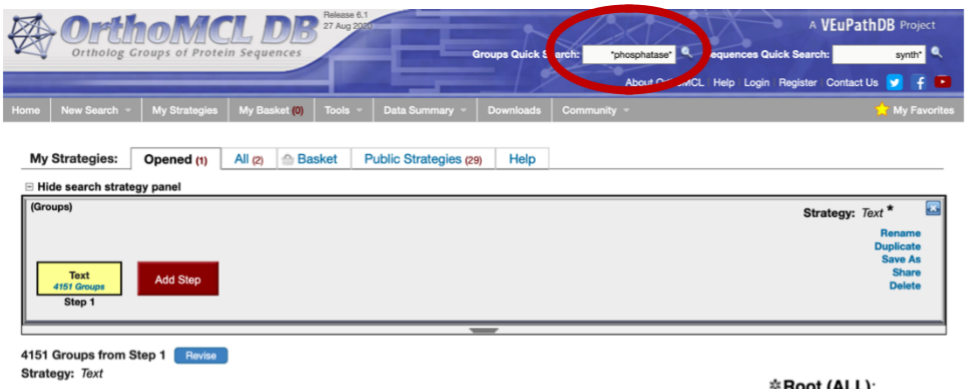
BACT=0T AND ARCH=0T AND cpar+cand+choi+chot+chom+chod+cmel+cmur+cpia+ctyz+cubi>=1T AND gass+gadh+gasb+gabb+gase+gmur>=1T

All VEuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Evolution -> Orthology Phylogenetic Profile. This search is very useful to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus but not present in the host as these genes may make good drug targets or vaccine candidates. Optional: go to your favorite VEuPathDB site and run this search to identify all genes that are not present in human or mouse.

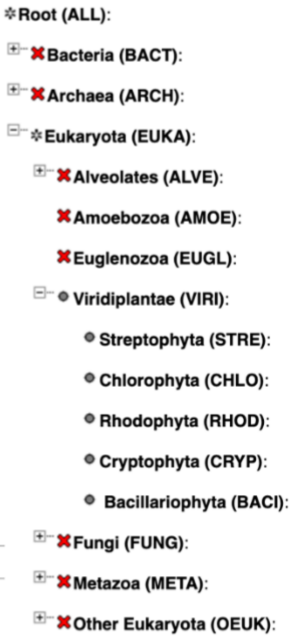
3. Combining searches in OrthoMCL (Use http://orthomcl.org for this exercise).

Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

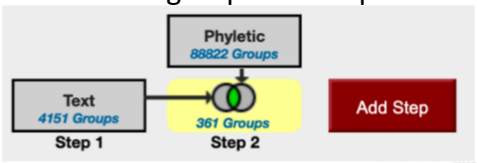
a. Use the text search **to find OrthoMCL groups** that contain the word "*phosphatase*" (note that the search should be run without the quotation marks but with the asterisks).



b. Add a step and run a phyletic pattern search for groups that contain any plant protein but do not contain any other organism outside plants. (hint: make sure everything has a red x on it except for plants (Viridiplantae (VIRI)), which should be a grey circle).



c. How many groups did you return? Explore the multiple sequence alignments from some of these groups. (Hint: click on a group ID and open the MSA tab).



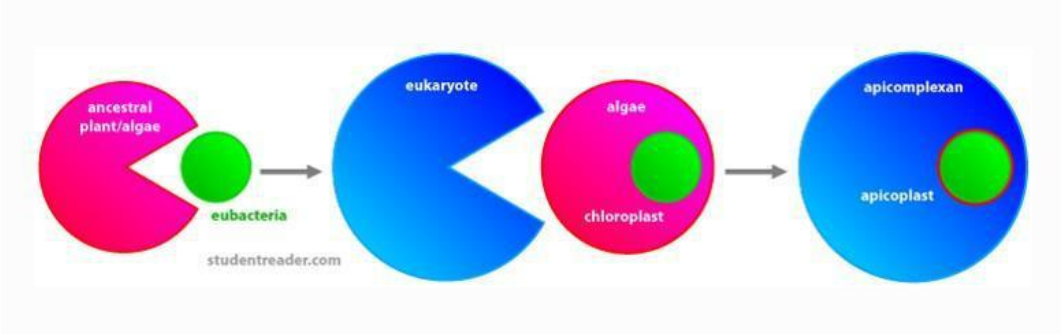
4. Exploring a specific OrthoMCL group - examining the cluster graph. (Use http://orthomcl.org for this exercise).

a. Visit the orthomcl group OG6_131670. You can either type the ID in the group quick search option at the top of the page of follow this link: http://orthomcl.org/group/OG6_131670

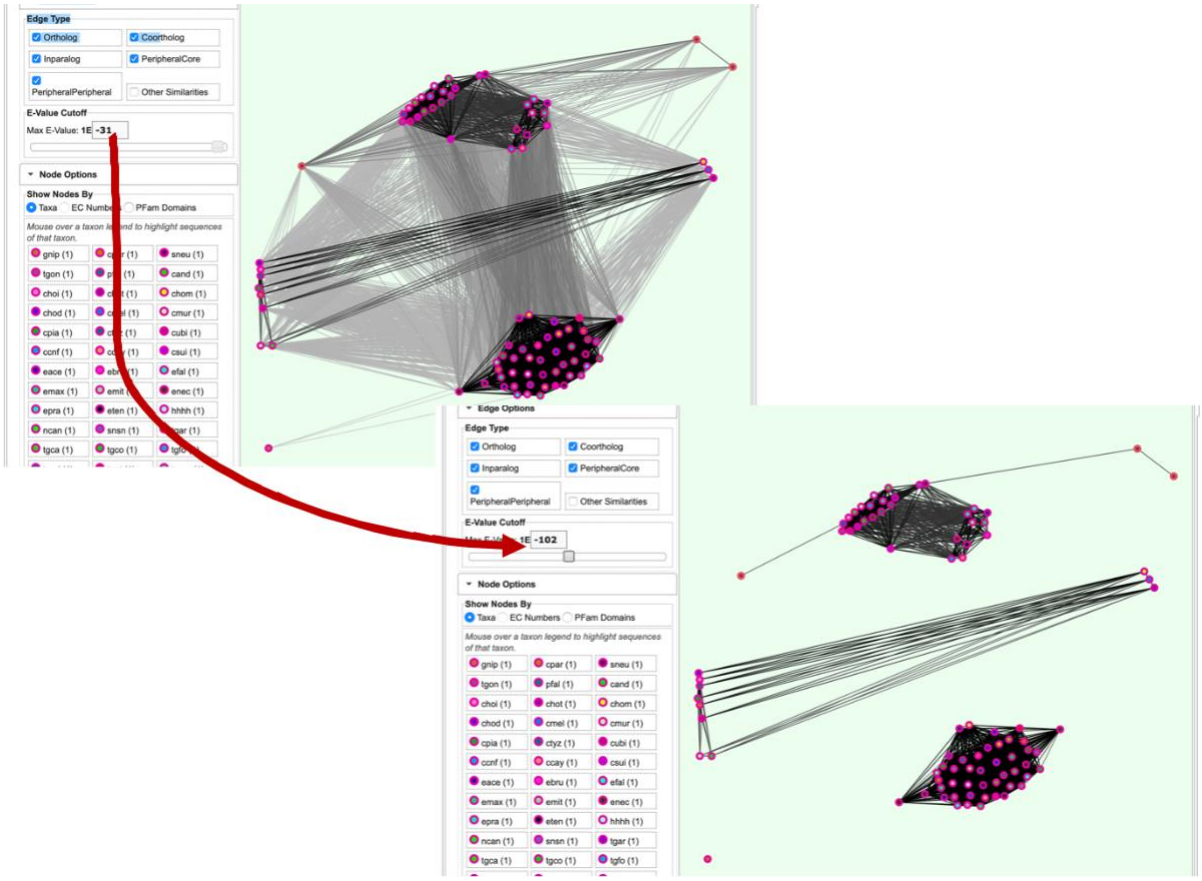
b. *Examine the "Sequences & Statistics" tab:* Based on the product descriptions of the members of this group, what kind of a proteins are in this group? What is the phylogenetic distribution of the members of this group?

c. *Examine the "Cluster Graph" tab:* Modify the E-value cutoff slider. What happens when you increase or decrease the E-value? Can you identify subclusters?

5. Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.
Note: For this exercise use http://veupathdb.org

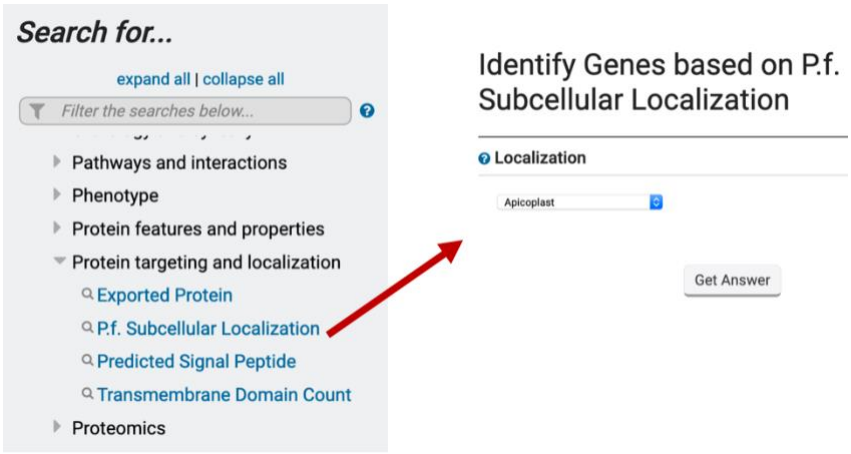


The apicoplast likely became encased in four membranes via a double endosymbiotic event.



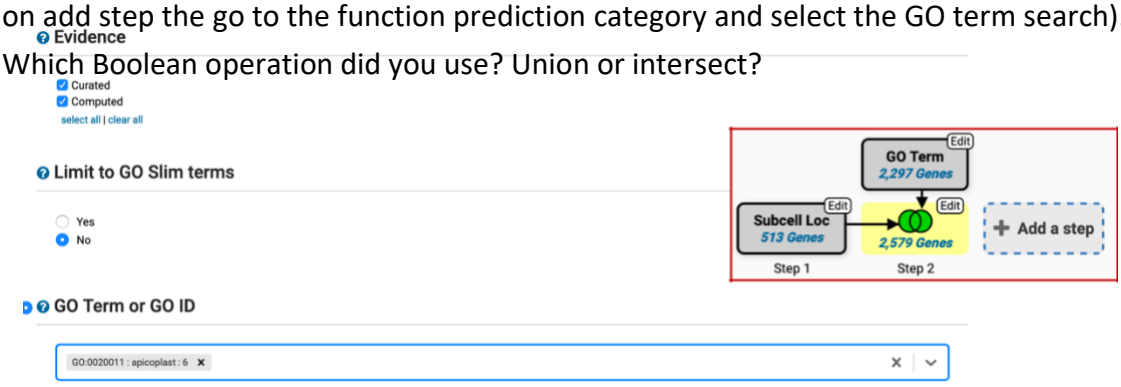
The chloroplast arose by engulfment of a cyanobacteria by a plant/algae ancestor. An algae was then engulfed by the ancestor of all apicomplexans. Thus, an apicoplast organelle arose with four membranes.

a. Start by finding genes in *Plasmodium* that are predicted to target to the apicoplast. Hint: click on “Protein targeting and localization” then on “P.f. Subcellular Localization”. You can further expand your list of potentially Apicoplast targeted proteins by running a GO terms search for the term “apicoplast” or the GO ID: GO:0020011 in *P falciparum* 3D7 (hint, click

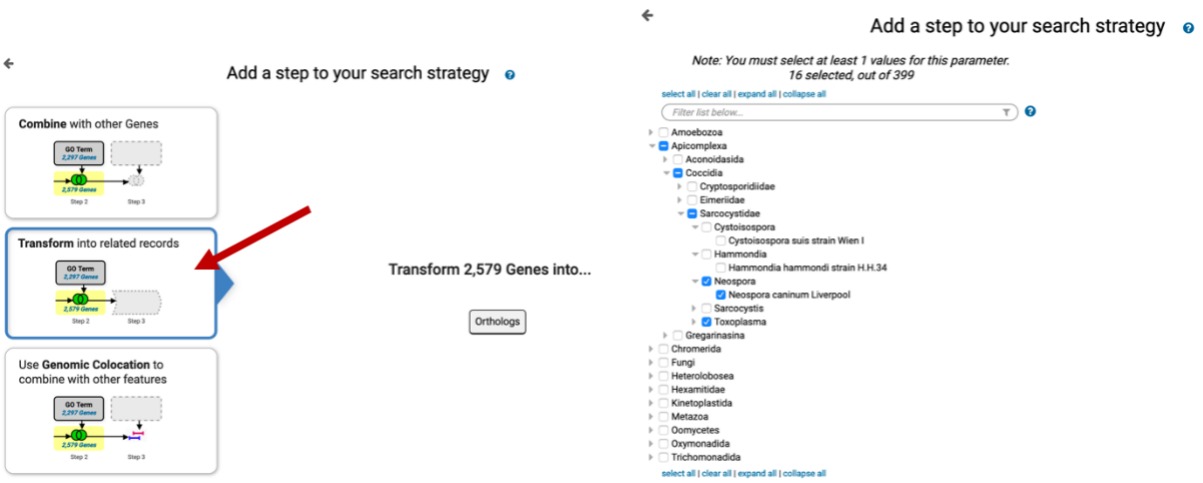


on add step the go to the function prediction category and select the GO term search).

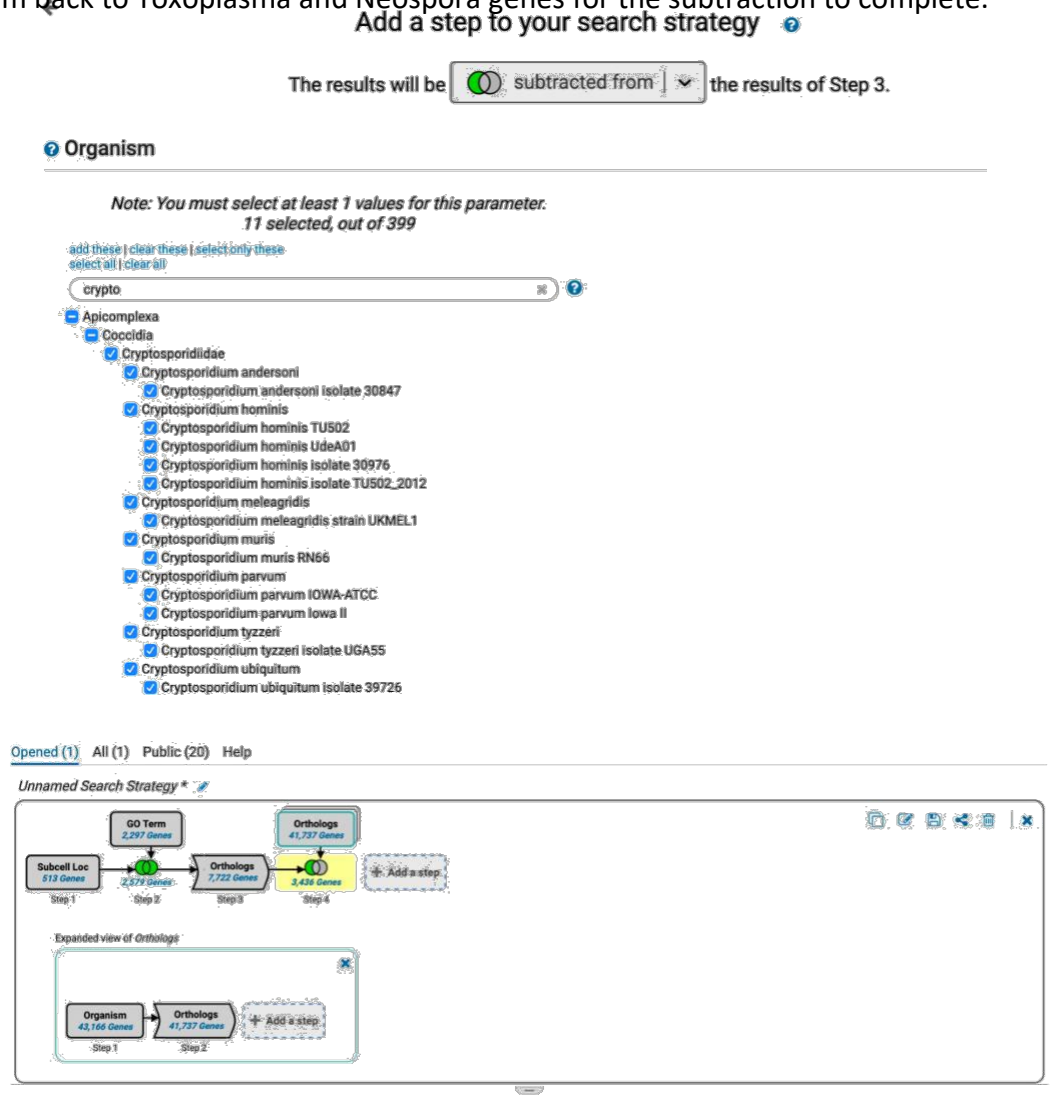
Which Boolean operation did you use? Union or intersect?



b. Transform the results of the above search to their *Toxoplasma* and *Neospora* orthologs. Hint: add a step, then select “Transform by Orthology”. On the search page , select all *Toxoplasma* and *Neospora.*



Although *Cryptosporidium* is an apicomplexan parasite it has lost its apicoplast! Can you use this fact to refine your results from the above search? Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy and use the ortholog transform back to Toxoplasma and Neospora genes for the subtraction to complete.

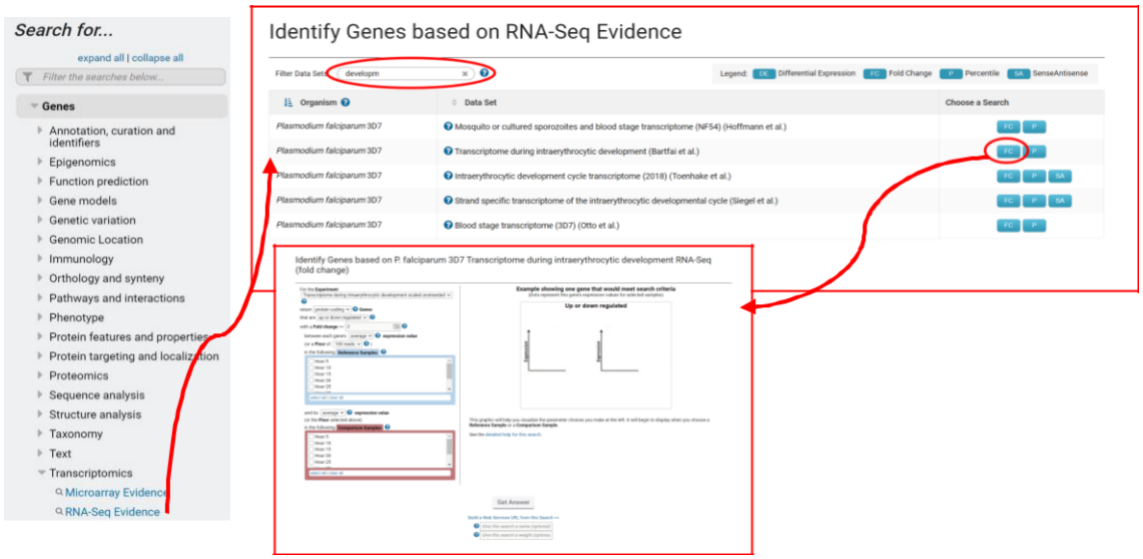


Exploring Transcriptomic data

1. Exploring RNA sequence data in *Plasmodium falciparum.*
   Note: For this exercise use http://www.plasmodb.org

   a. Find all genes in *P. falciparum* that are up-regulated during the later stages of the intraerythrocytic cycle.

   Use the fold change search for the data set "**Transcriptome during intraerythrocytic development (Bartfai *et al.*)".** For this data set, synchronized Pf3D7 parasites were assayed by RNA-seq at 8 time-points during the iRBC cycle. We want to find genes that are up-regulated in the later time points (30, 35, 40 hours) using the early time points (5, 10, 15, 20, 25 hours) as reference.



There are a number of parameters to manipulate in this search. As you modify parameters on the left side note the dynamic help on the right side. See screenshots.

Direction: the direction of change in expression. Choose up-regulated.

Fold Change>= the intensity of difference in expression needed before a gene is returned by the search. Choose 12 but feel free to modify this.

Reference Sample: the samples that will serve as the reference when comparing expression between samples. choose 5, 10, 15, 20, 25

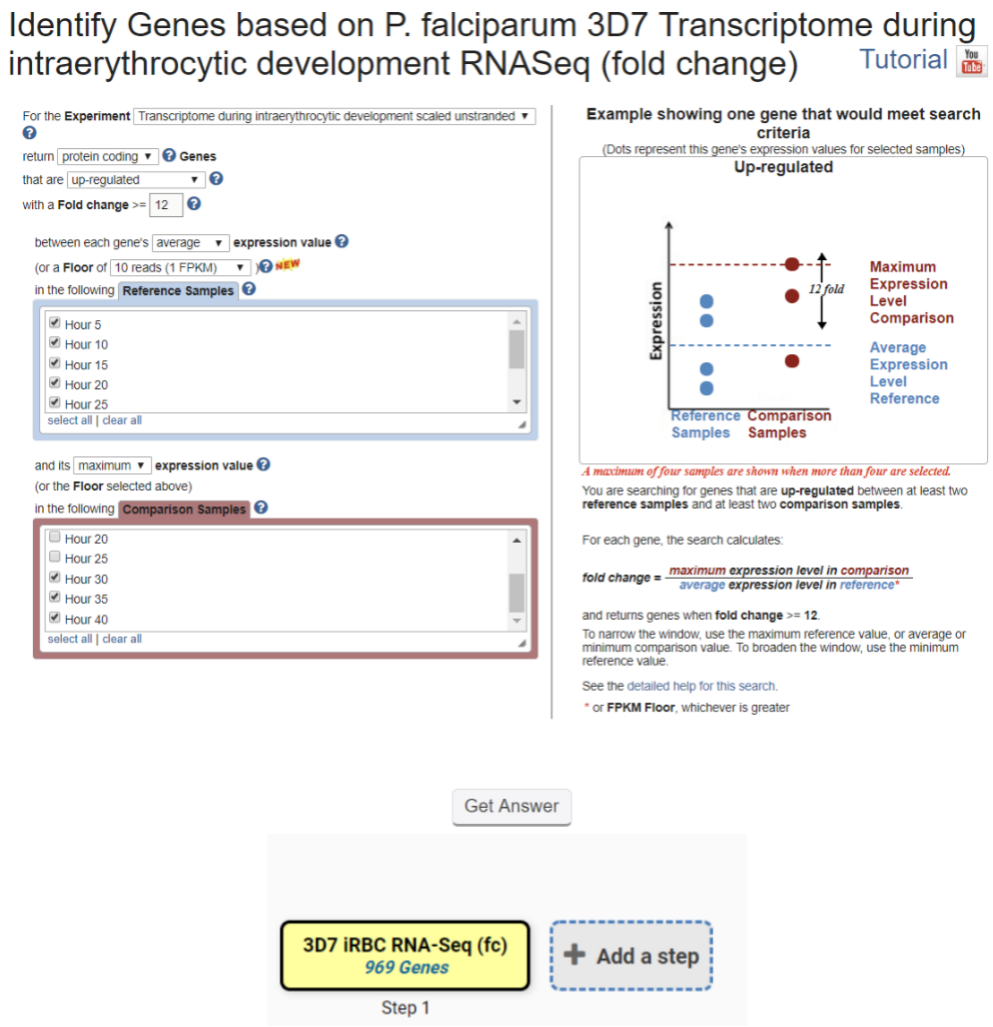
Between each gene's AVERAGE expression value: This parameter appears once you have chosen two Reference Samples and defines the operation applied to reference samples.

Fold change is calculated as the ratio of two values (upregulated ratio = expression in comparison)/(expression in reference). When you choose multiple samples to serve as reference, we generate one number for the fold change calculation by using the minimum, maximum, or average. Choose average

- (or a Floor of 10 reads): This parameter defines a lower limit of aligned reads for a gene to avoid unreliable fold change calculations. (Low numbers of aligned reads means low expression but the low values may be may be technically inaccurate. Dividing by small numbers creates large numbers. 2000FPKM/10 = 200; 2000/0.1 = 20,000) If a gene has fewer than 10 aligned reads, it is assigned 10 reads before the fold change calculation is made. Leave this as default at 10 reads.

- Comparison Sample: the sample that you are comparing to the reference. In this case you are interested in genes that are up-regulated in later time points choose 30, 35, 40

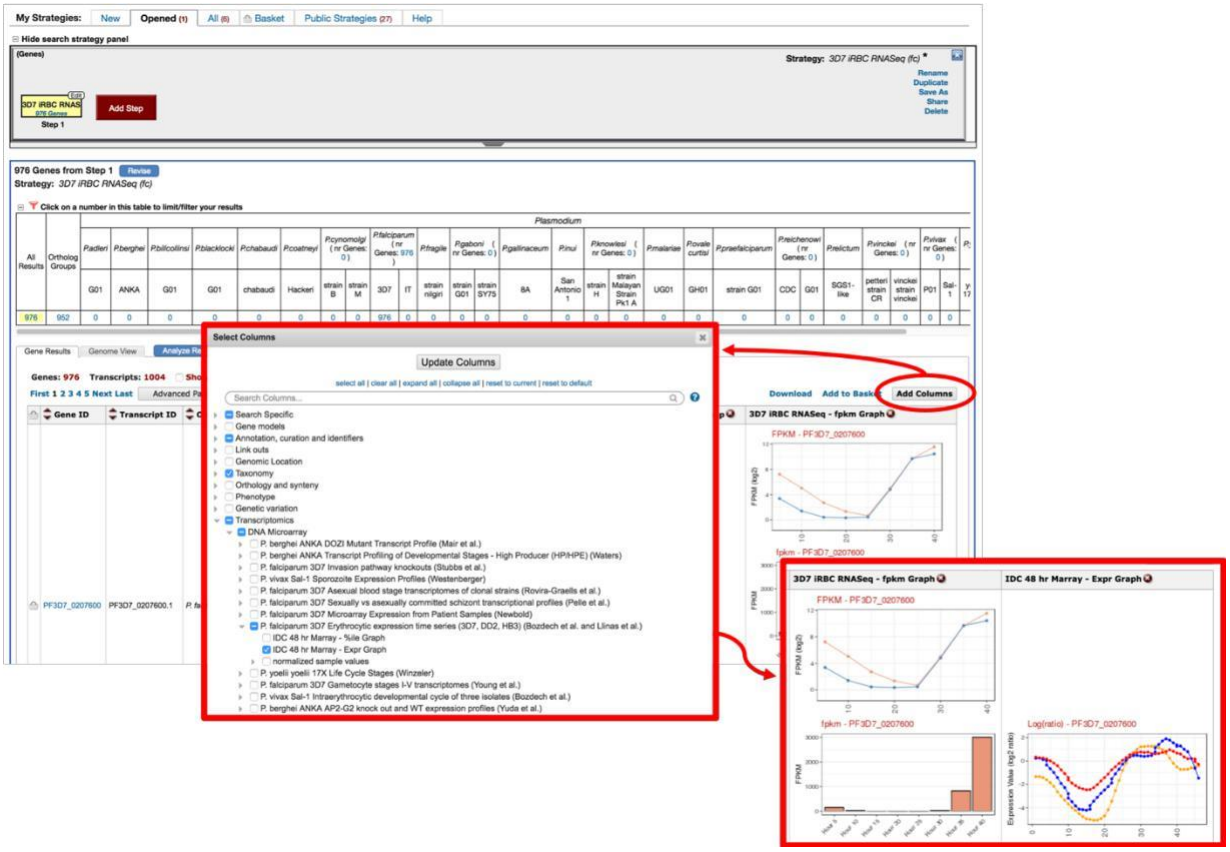
- And its AVERAGE expression value: This parameter appears once you have chosen two Comparison Samples and defines the operation applied to comparison samples. See explanation above. Choose average



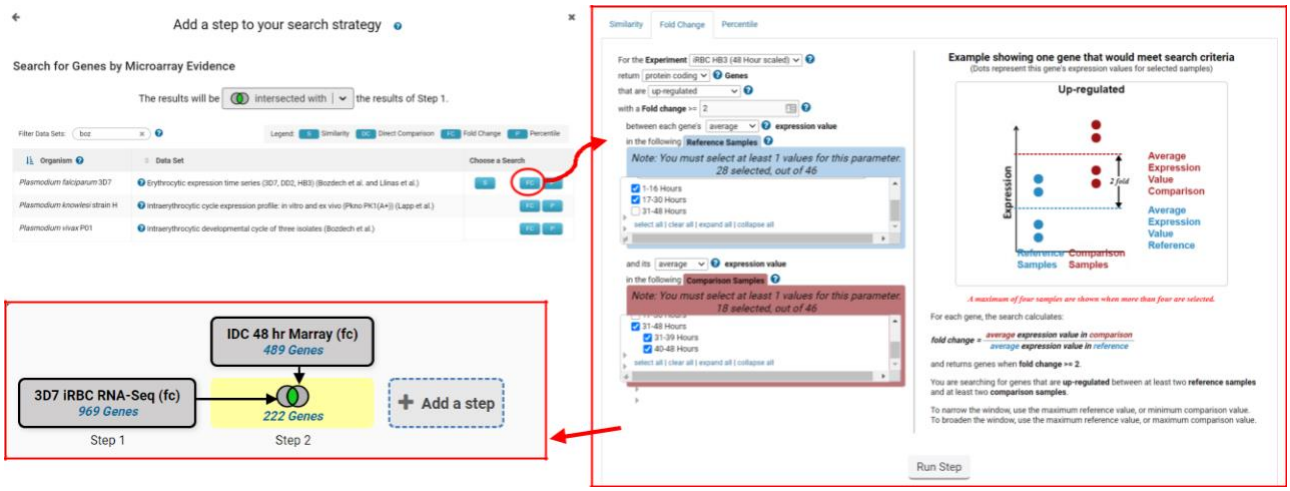
b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

Hint: PlasmoDB contains data from a similar experiment that was analyzed by microarray

instead of RNA sequencing. This experiment is called: **Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech et al. and Linas et al.). IDC 48 hr Marray – Expr Graph** shows normalized expression values. To directly compare the data for genes returned by the RNA-seq search that you just ran, add the column called "Pf-iRBC 48hr - Graph".

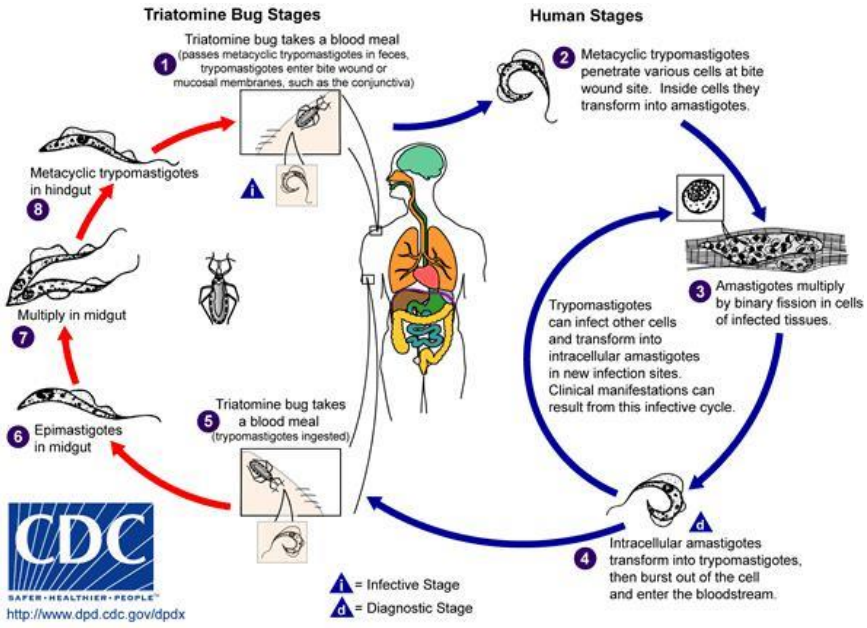


OPTIONAL: You can also run a fold change search using this experiment to compare results on a genome scale. Add a step to your strategy and intersect your current results (genes upregulated 12 fold in later IDC time periods) with a fold change search using the "Erythrocytic expression time series (3D7, Dd2, HB3) (Bozdech et al. and Linas et al.)" experiment (under microarray evidence). Configure it similarly to the RNA-seq experiment although you will probably need to make the fold change smaller (try 2 or 3) due to the decreased dynamic range of microarrays compared to RNA-seq.

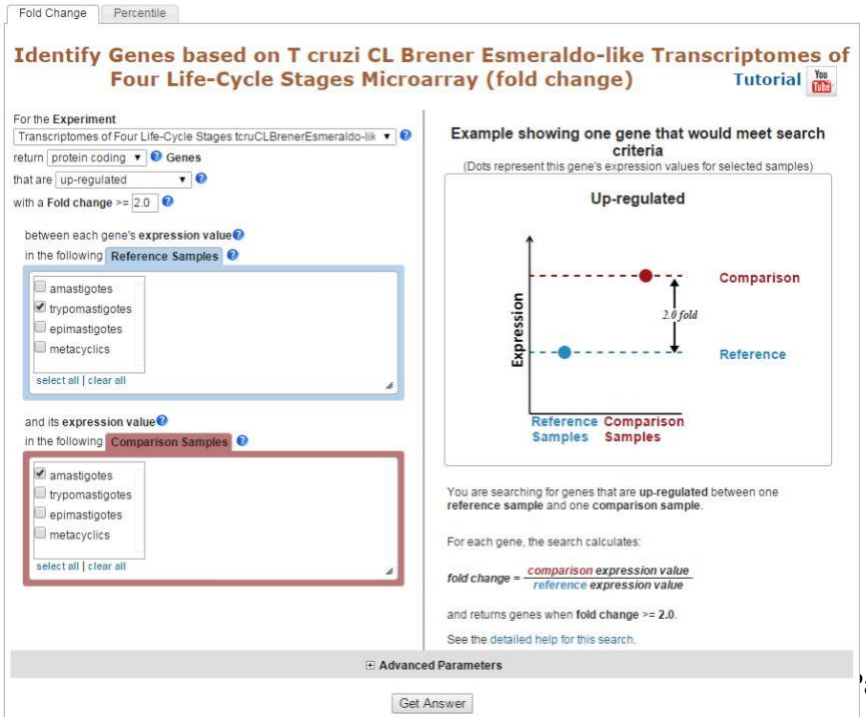

Add a step to your search strategy
Search for Genes by Microarray Evidence
The results will be intersected with the results of Step 1.
IDC 48 hr Marray (fc)
489 Genes
3D7 iRBC RNA-Seq (fc)
969 Genes
222 Genes
Add a step
Step 1
Step 2
Similarity
Fold Change
Percentile
Example showing one gene that would meet search criteria
Up-regulated
Average Expression Value Comparison
Average Expression Value Reference
Expression
2 fold
Reference Samples
Comparison Samples
Run Step

2. Exploring microarray data in TriTrypDB.

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



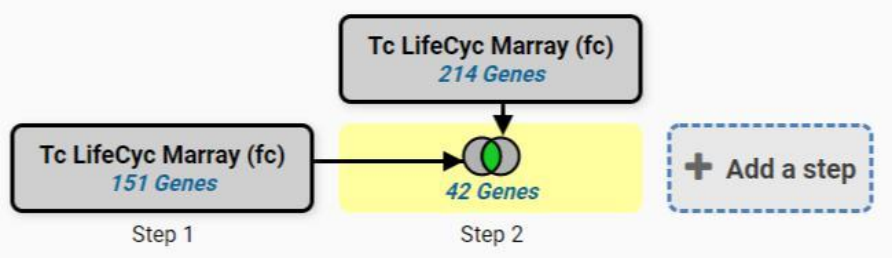
a. Find *T. cruzi* protein coding genes that are upregulated in amastigotes compared to trypomastigotes. Go to the transcript expression section then select **microarray**. Choose the fold change (FC) search for the data set called: **Transcriptomes of Four Life-Cycle Stages (Minning et al.)**.



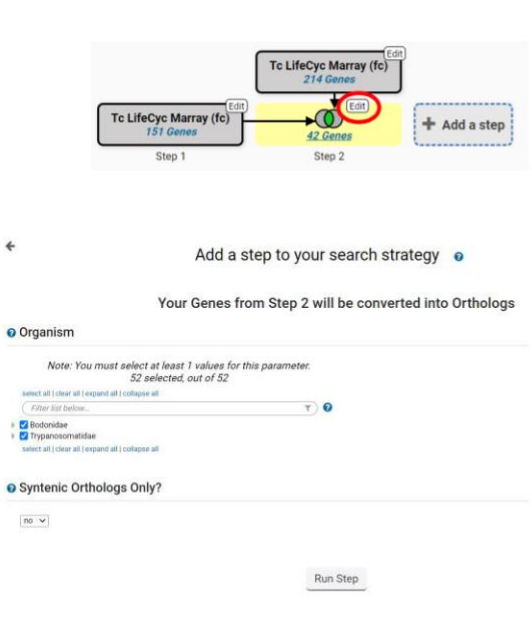
- Select the direction of regulation, your reference sample and your comparison sample. For the fold change keep the default value 2.

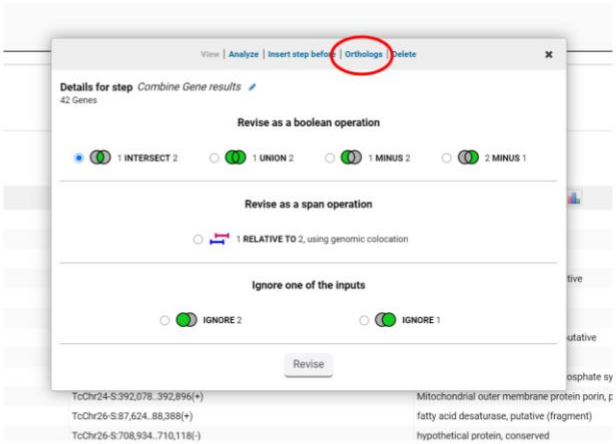
- How many genes did you find? Do the results seem plausible?

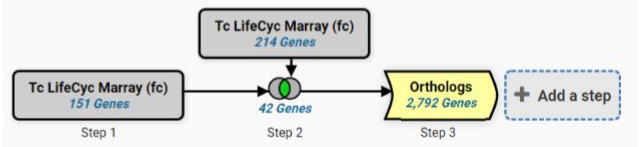
- Are any of these genes also up-regulated in the replicative insect stage compared to the transmissive insect stage? How can you find this out? (*Hint*: add a step and run a microarray search comparing expression of epimastigotes to metacyclics).



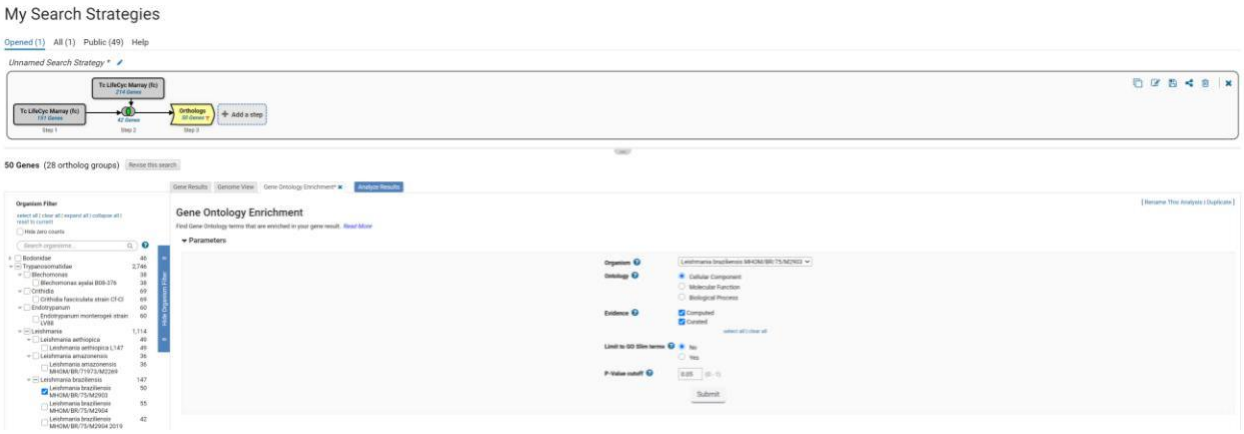
- Do these genes have orthologs in other kinetoplastids? Transform your results into orthologs in all other organisms in TriTrypDB. This can be done by adding a step, or by editing a step, as shown in the screenshots.





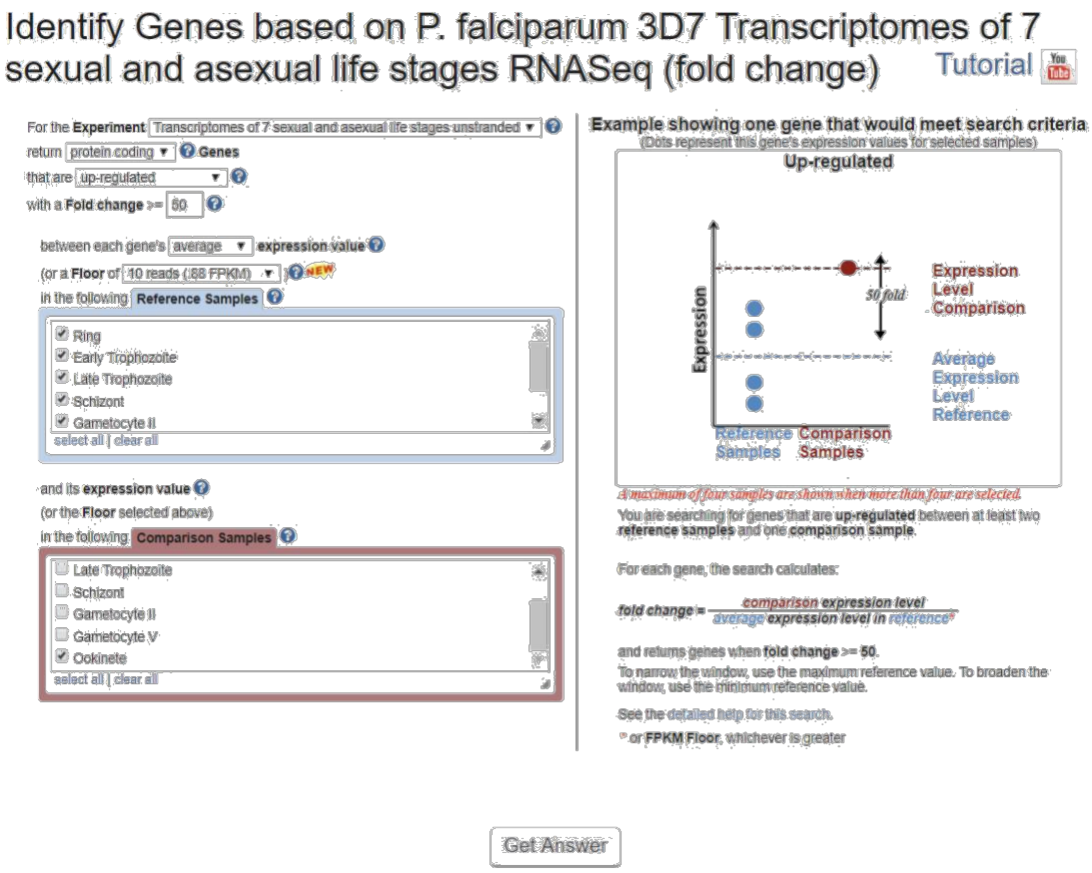


- How many orthologs exist in *L. braziliensis* MHOM/BR/75/M2903? (*Hint*: look at the organism filter to the left of your result list. Click on the check box next to a species, then click apply to view results from a specific species). Explore your results. Scan the product descriptions for this list of genes. Did you find anything interesting? Perhaps a GO enrichment analysis would support your ideas.



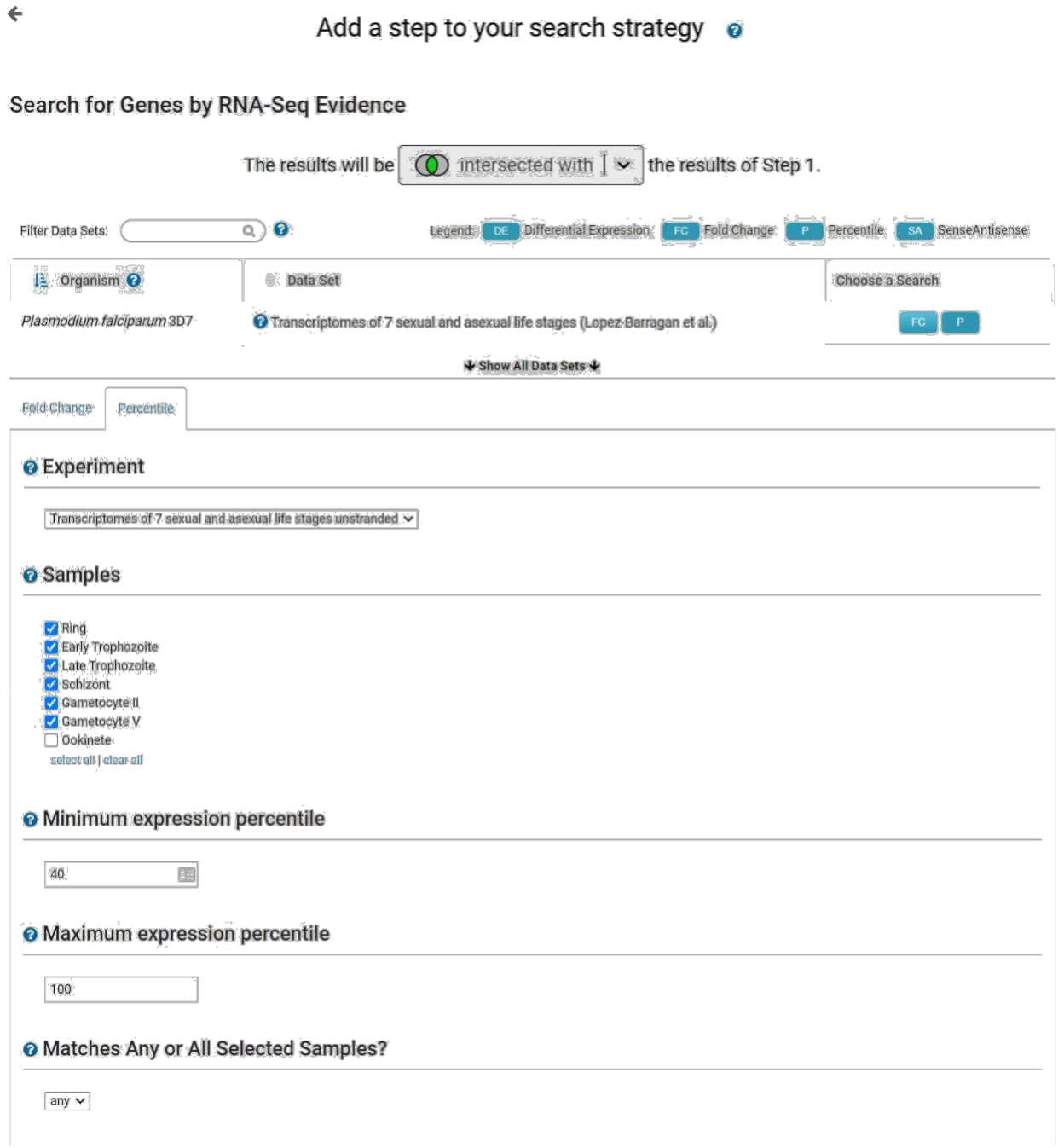
3. Finding genes based on RNAseq evidence and inferring function of hypothetical genes. Note: Use http://plasmodb.org for this exercise.

a. Find all genes in *P. falciparum* that are up-regulated at least 50-fold in ookinetes compared to other stages: "Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.)". For this search select "average" for the operation applied on the reference samples.

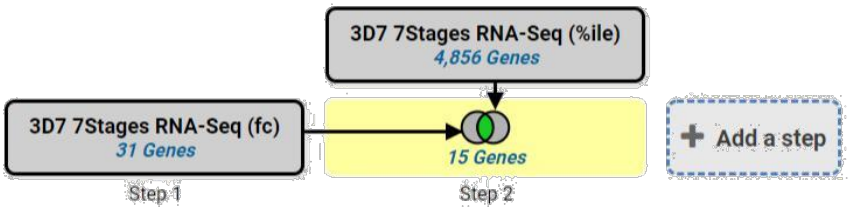


b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the average expression level of other stages. Despite the high fold change, some genes in the list may be highly expressed in the other stages. How can you remove genes from the list that are highly expressed in the other stages?

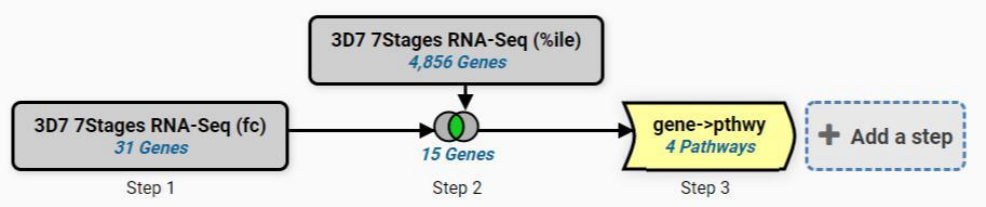
- *Hint: Add a search for genes based on RNA Seq evidence from the same experiment, but this time select the percentile search: P.f. seven stages - RNA Seq (percentile). What minimal percentile values should you choose? 40 – 100%? How does setting the any / all samples impact the result …. Which would be better in this case?*



- *Hint II: Try changing the operator from average to maximum for the set of non-ookinete stages in your initial fold change search. What does this do? How do the resulting genes compare with the two step strategy you generated in the first hint? Which hint do you think works better?*

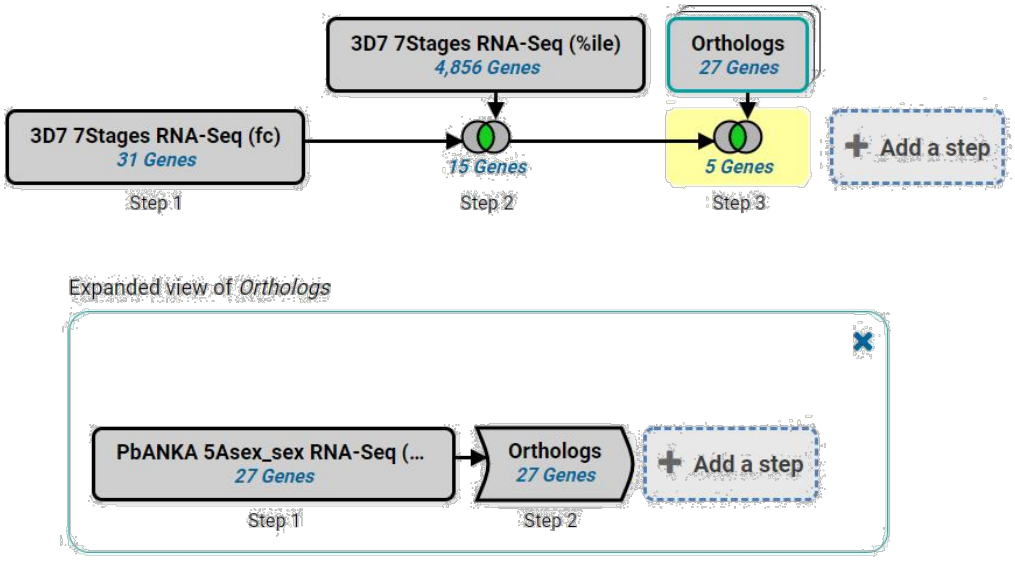


c. Which metabolic pathways are represented in this gene list? *Hint: add a step and transform results to pathways*. How does this result compare to running a pathways enrichment on step 2?



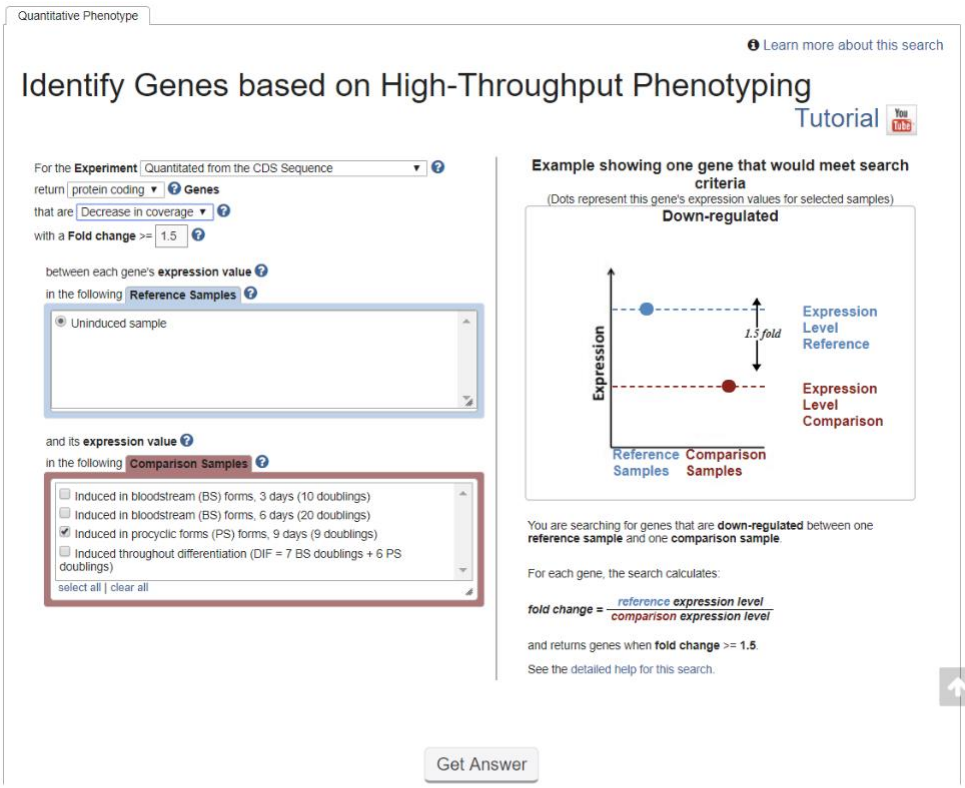
d. What happens if you revise the first step and modify the fold difference to a lower value - 10 for example? Compare results when you also modify the "between each genes" parameter. What happens if you set this to maximum? Which value do you think is most stringent for ensuring at 10 fold up regulation compared to the other samples?

e. PlasmoDB also has an experiment examining gene expression during sexual development in *Plasmodium berghei* (rodent malaria). Can you determine if there are genes that are up-regulated in both human and rodent ookinetes (compared to all other stages)? *Hint:* start by deleting the last step you added in this exercise (transform to pathways). To do this click on edit then delete in the popup. Next, add steps for the *P. berghei* experiments "P berghei ANKA 5 asexual and sexual stage transcriptomes RNASeq". Note that you will have to use a nested strategy or by running a separate strategy then combining both strategies.



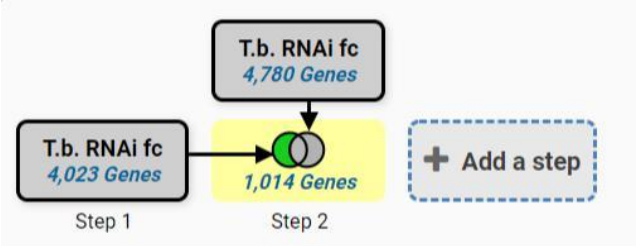
4. Find genes that are essential in procyclics but not in blood form *T. brucei*.
   Note: for this exercise use http://TriTrypDB.org.

   - Find the query for High Throughput Phenotyping. Think about how to set up this query (*Hint*: you will have to set up a two-step strategy). Remember you can play around with the parameters but there is no one correct way of setting them up –

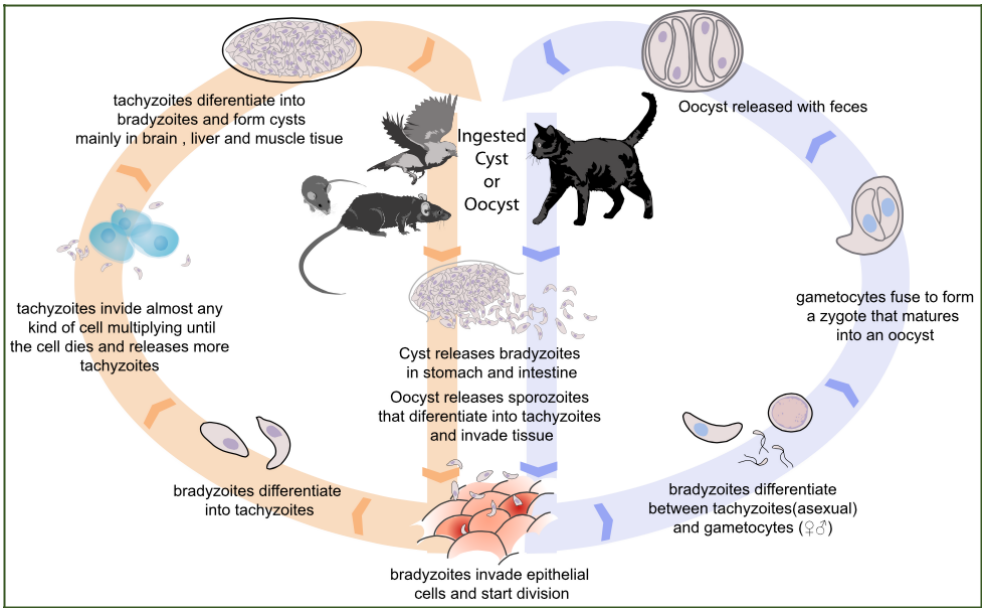


- Next add a step and run the same search except this time select the “induced bloodstream form” samples.

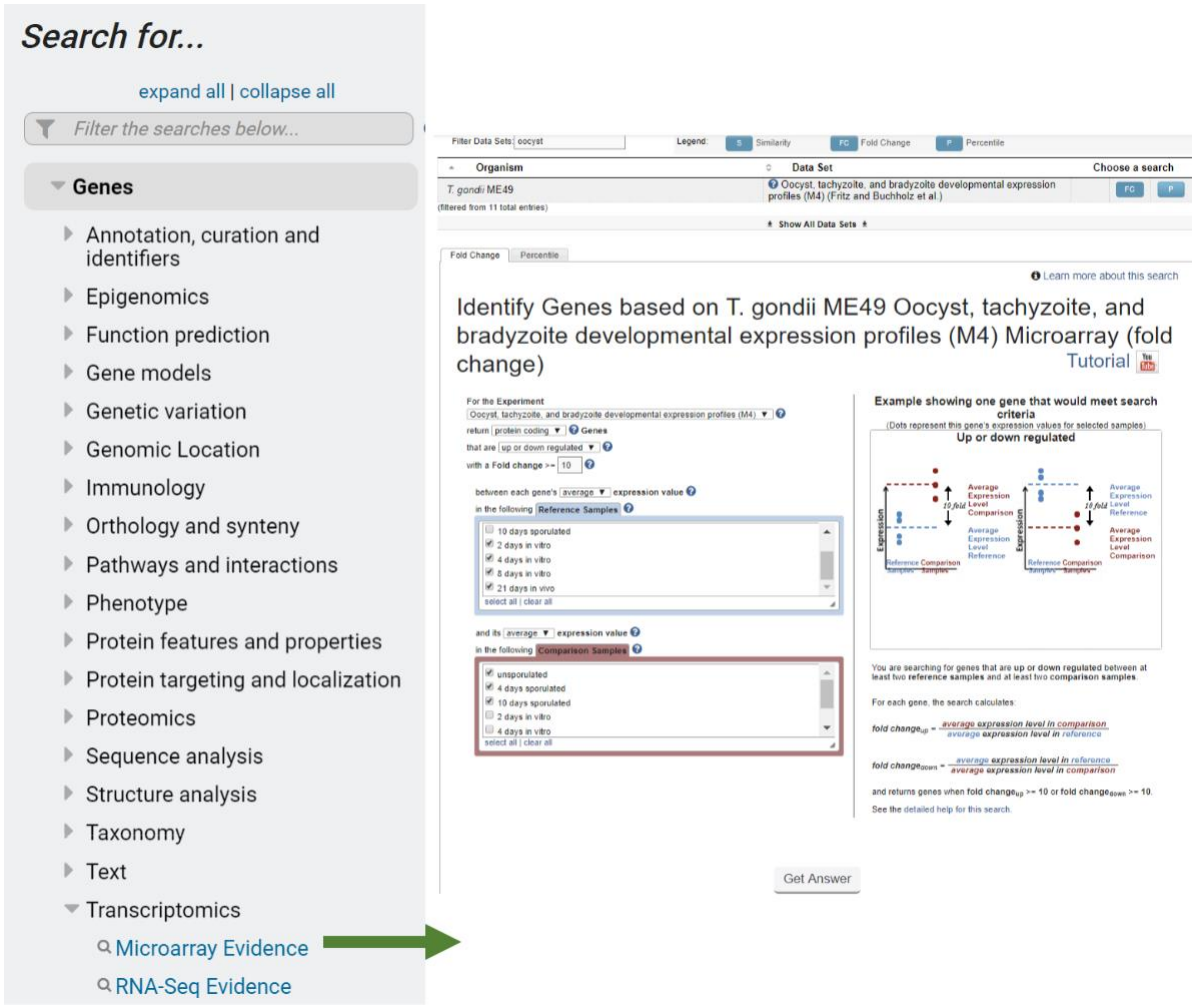
- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.



5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
   Note: For this exercise use http://toxodb.org



   a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the "**Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (Fritz and Buchholz)**" microarray experiment.

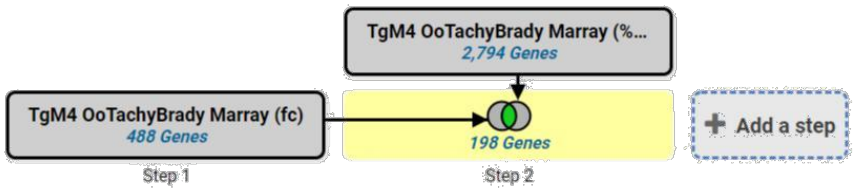


b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile … i.e., likely not expressed at those stages. *(Hint*: after you click on add step find the same experiment under microarray expression and chose the percentile search).

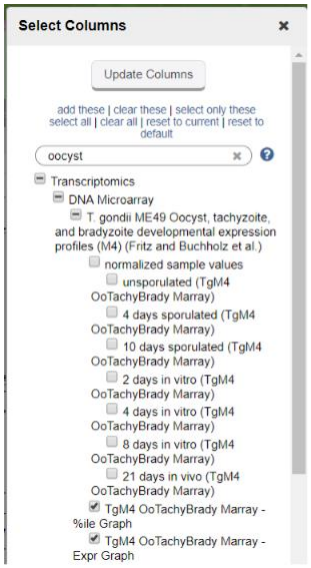
   - Select the 4 **non-oocyst** samples.

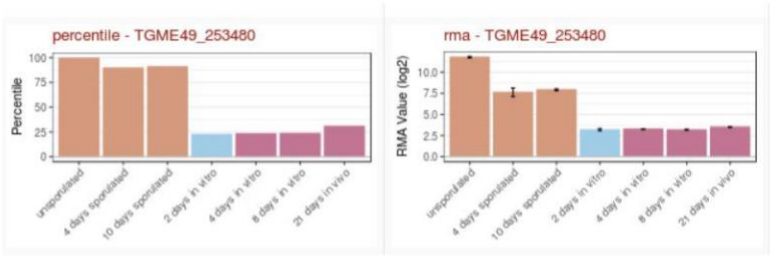
   - We want all to have less than 50th percentile so set ***minimum percentile*** to 0 and ***maximum percentile*** to 50.

- Since we want all of them to be in this range, choose ALL in the "***Matches Any or All Selected Samples***".



- To view the graphs in the final result table, turn on the columns called "TgM4 OoTachyBrady Marray - Expr Graph" and "TgM4 OoTachyBrady Marray - %ile Graph" (inside the "T. gondii ME49 Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (Fritz and Buchholz et al.)" Microarray).



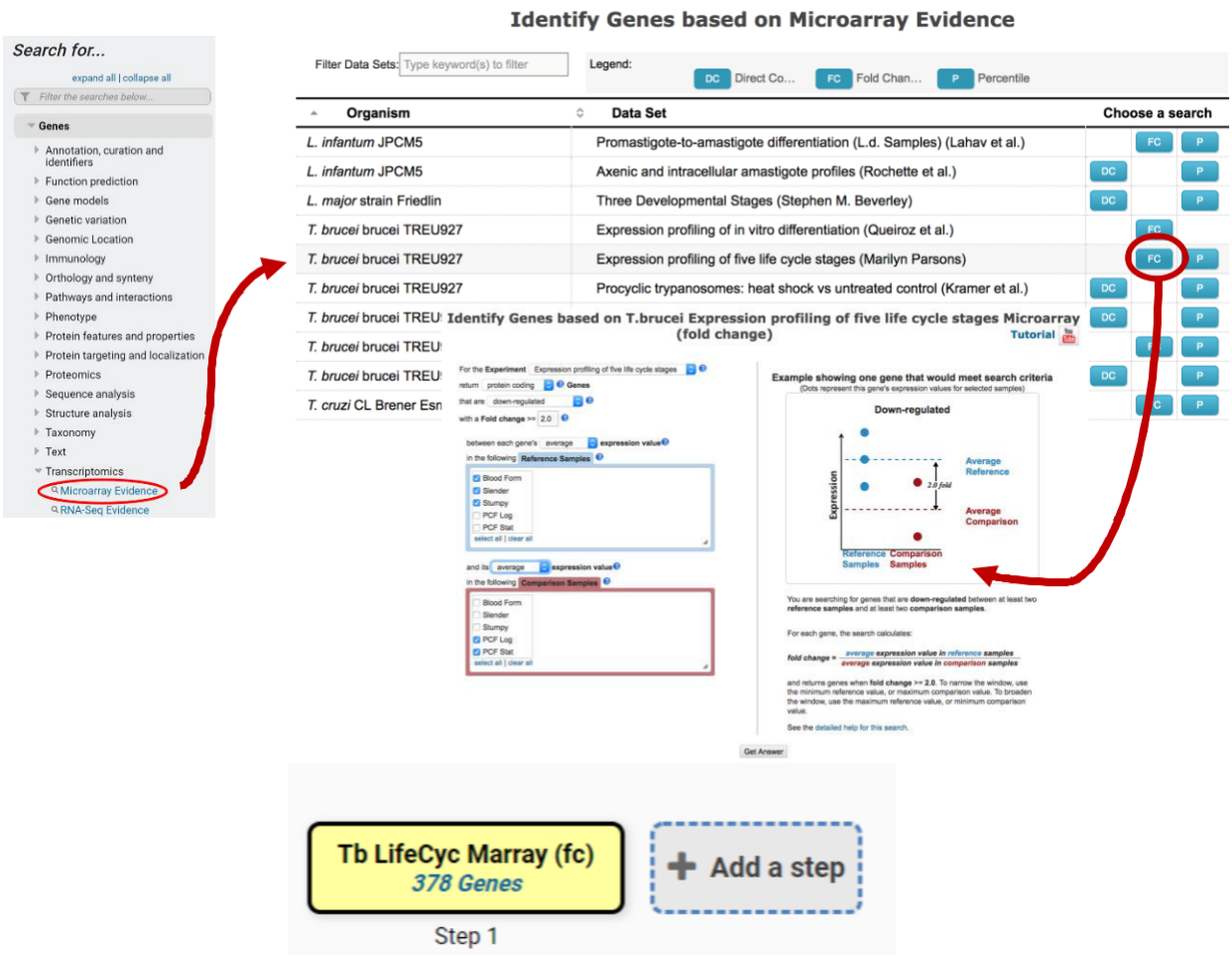


6. Comparing RNA abundance and Protein abundance data.

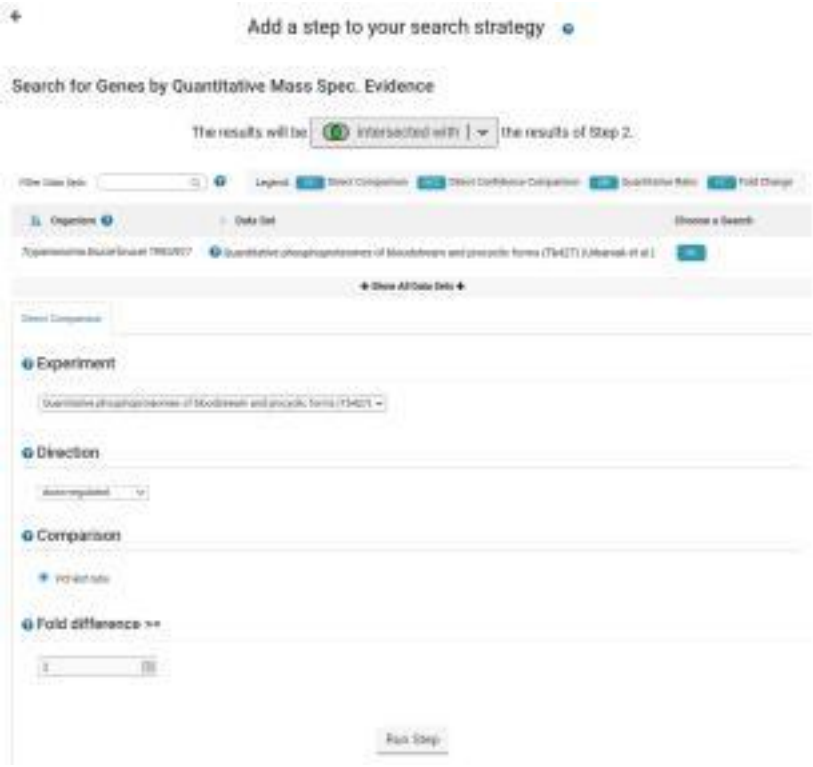
Note: for this exercise use http://TriTrypDB.org.

In this exercise we will compare genes that show differential RNA abundance levels between procyclic and blood form stages in *T. brucei* with genes that show differential protein abundance in these same stages.

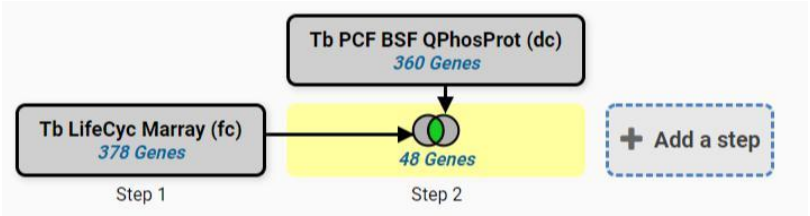
a. Find genes that are down-regulated 2-fold in procyclic form cells. Go to the search page for Genes by Microarray Evidence and select the fold change search for the "Expression profiling of five life cycle stages (Marilyn Parsons)" experiment and configure the search to return protein-coding genes that are down-regulated 2 fold in procyclic form (PCF) relative to the Blood Form reference sample. Since there are two PCF samples, it is reasonable to choose both and average them.



b. Add a step to compare with quantitative protein expression. Select protein expression then “Quantitative Mass Spec Evidence” and the "Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427) (Urbaniak et al.)" experiment. Configure this search to return genes that are down-regulated in procyclic form relative to blood form.



c. How many genes are in the intersection? Does this make sense? Make certain that you set the directions correctly.



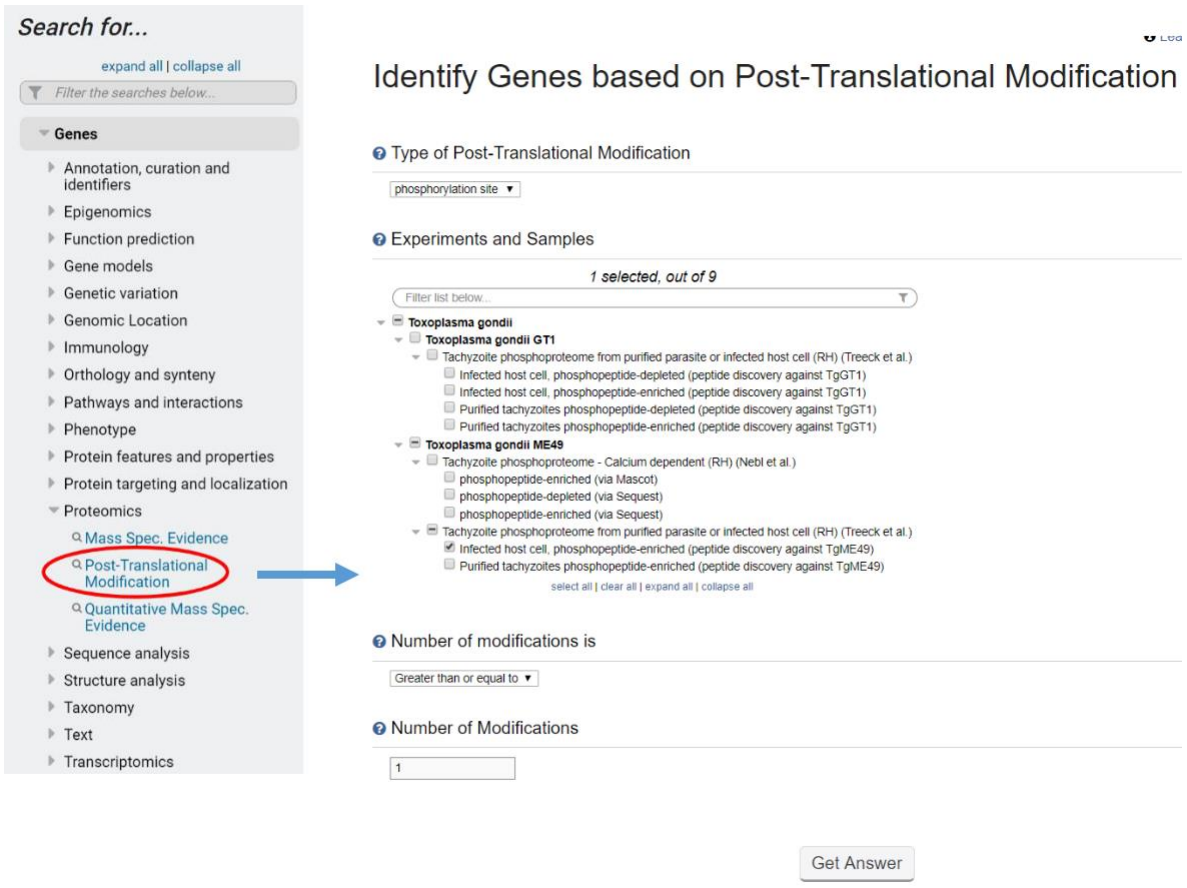
d. Try changing directions and compare up-regulated genes/proteins. (*Hint:* revise the existing strategy … you might want to duplicate it so you can keep both). When you change one of the steps but not the other do you have any genes in the intersection? Why might this be?

e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? (*Hint:* you could insert steps to restrict based on percentile or add a RNA Sequencing step that has the same samples).

7. Find genes with evidence of protein phosphorylation in intracellular *Toxoplasma* tachyzoites.
   For this exercise use http://www.toxodb.org

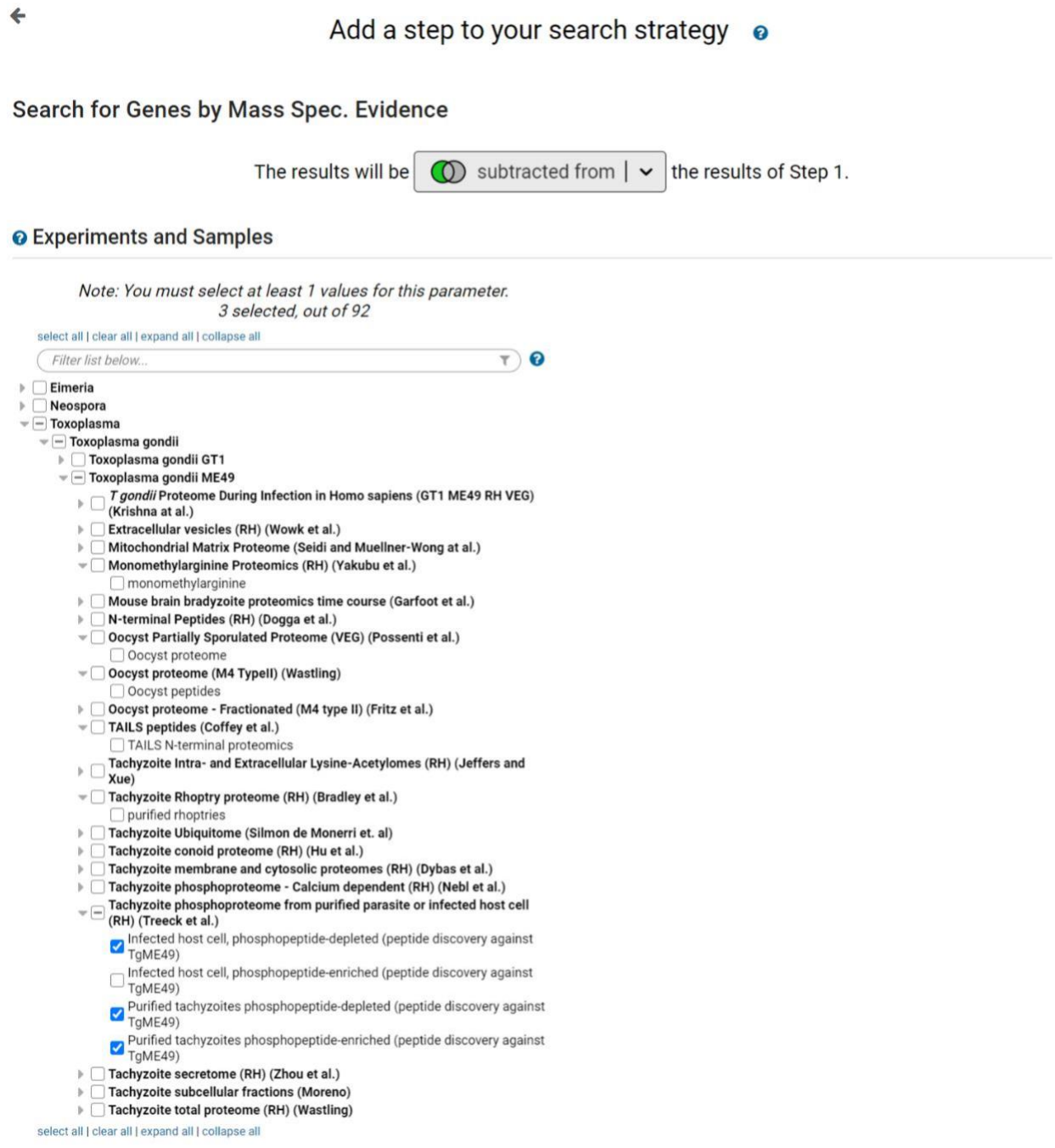
Phosphorylated peptides can be identified by searching the appropriate experiments in the Mass Spec Evidence search page.

7a. Find all genes with evidence of protein phosphorylation in intracellular tachyzoites. Navigate to the Post-Translational Modification search. Select the "**Infected host cell, phosphopeptide-enriched** (peptide discovery against **TgME49**)" sample under the experiment called "**Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)**"



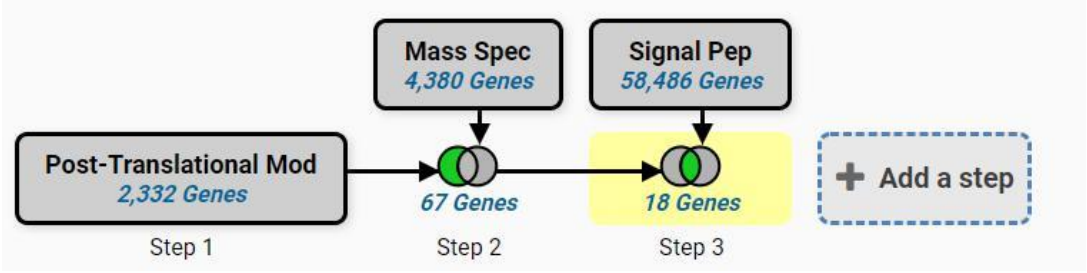
7b. Remove all genes with phosphorylation evidence from purified tachyzoites and the phosphopeptide depleted fractions.

Hint: Use the Mass Spec Evidence search to access the tachyzoite and depleted fractions. Subtract (1 minus 2) these results from your first search.



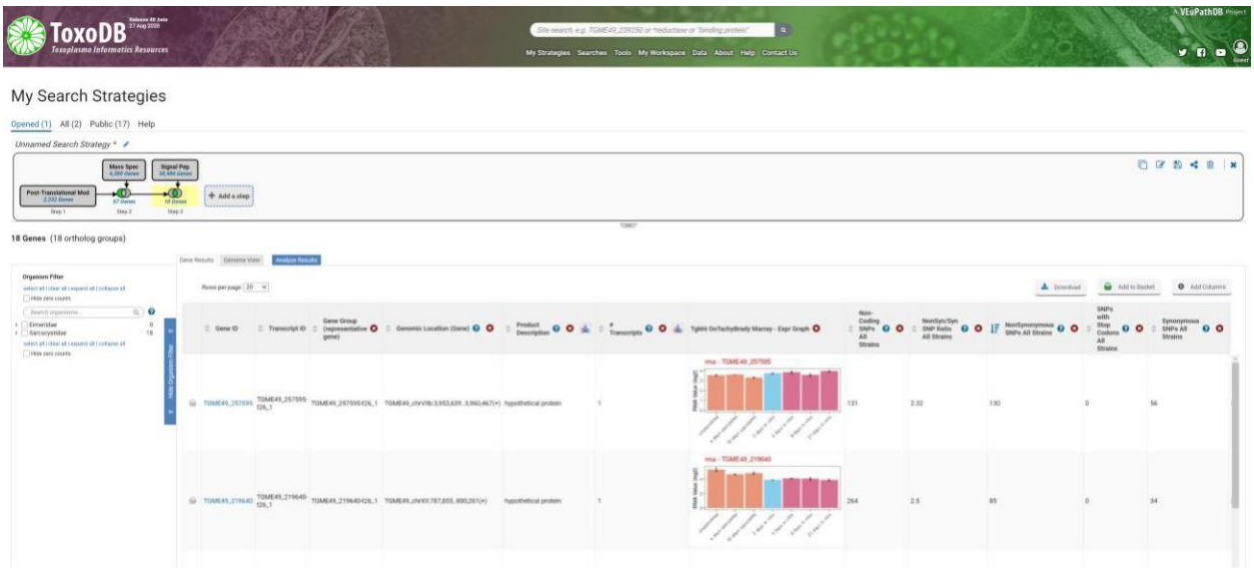
7d. Explore your results. What kinds of genes did you find? *Hint: use the Product description word column or perform a GO enrichment analysis of your results.*

7e. Are any of these genes likely to be secreted? Hint: add a step searching for genes with secretory signal peptides.



7f. Pick one or two of the hypothetical genes in your results and visit their gene pages. Can you infer anything about their function? Hint: explore the protein and expression sections.

7g. What about polymorphism data? Go back to your strategy and add columns for SNP data found under the population biology section. Explore the gene page for the gene that has the most number of non-synonymous SNPs. Hint: you can sort the columns by clicking on the up/down arrows next to the column names.



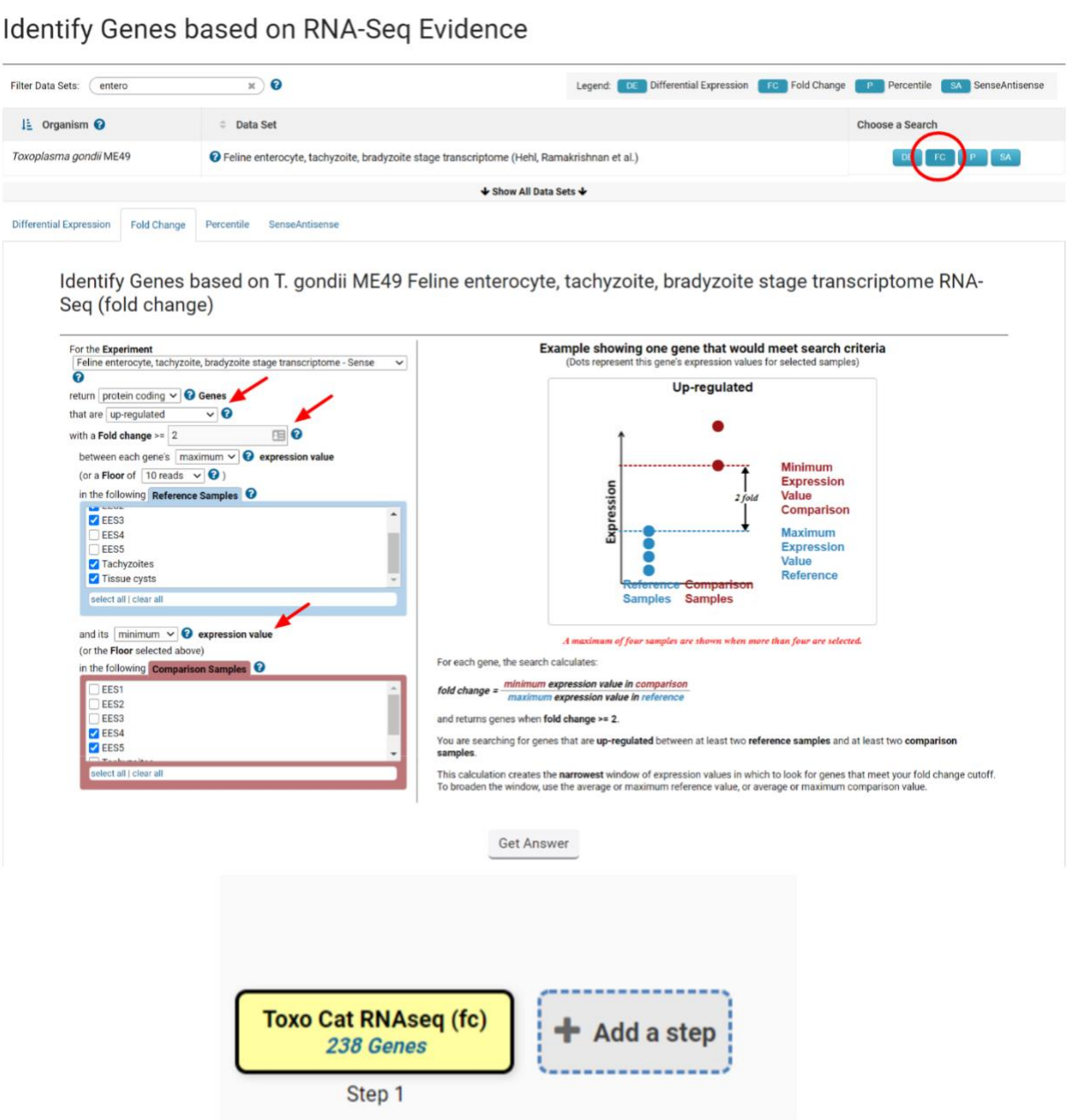
8. Find *T. gondii* genes expressed in late enteroepithelial stages

Toxoplasma gondii is a zoonotic pathogen for which felids serve as definitive hosts. In cats, the parasite undergoes several rounds of asexual replication before entering the sexual cycle which gives rise to oocysts that are shed into the environment. These then sporulate and become infective to humans and livestock. To understand the genes involved in the parasite development in the felid host and identify potential intervention targets, we designed a transcriptomic approach to compare the cat intestinal stages with the well characterized tachyzoites that mediate acute infection and tissue cysts that are responsible for chronic infection. Cats were infected with *T. gondii* CZ clone H3 tissue cysts from mouse brain and the intestinal stages were sampled at day 3, 5 and 7 post infection. As an input sample, we also collected tissue cysts from mouse brain. In vitro cultivated tachyzoites were also harvested. Total RNA was extracted, enriched for mRNA and used for cDNA synthesis. RNA-Seq was then performed to describe the transcriptomic repertoire of each developmental stage. RNA-seq datasets from each time point post inoculation with bradyzoites in kittens were subjected to cluster analysis and assigned to five enteroepithelial developmental stages (EES) according to their profile.

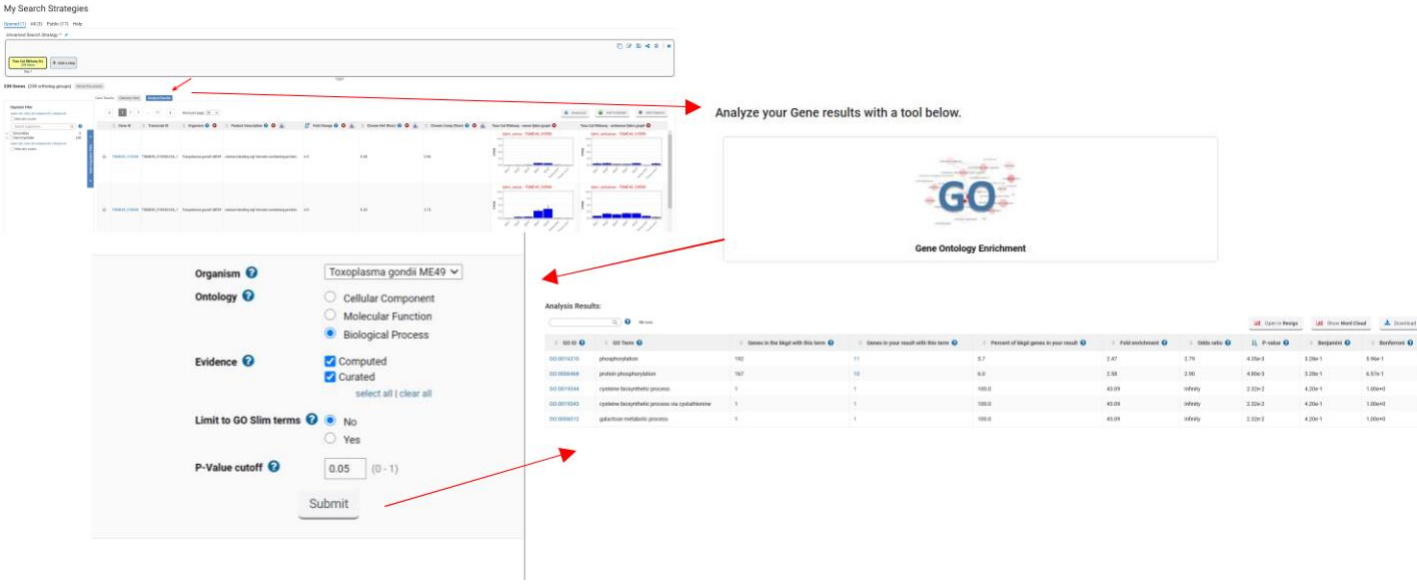
Cat enteroepithelial stages:

- EES1 = very early enteroepithelial stages
- EES2 = early enteroepithelial stages
- EES3 = mixed enteroepithelial stages
- EES4 = late enteroepithelial stages
- EES5 = very late enteroepithelial stages

- Navigate to the RNAseq searches and identify the experiment of cat enterocyte stages. Configure the search to identify call *T. gondii* genes that are upregulated by at least 2-fold in late and very late enteroepithelial stages (EES4 and EES5) compared to all other stages available from this experiment.



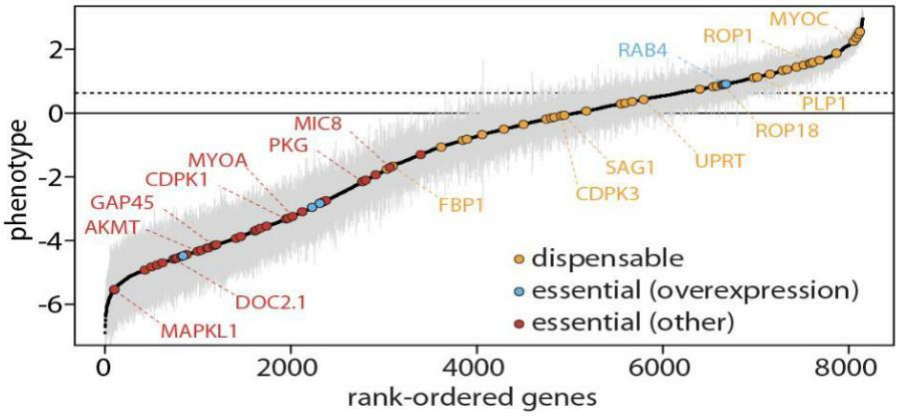
- What kinds of genes did this search identify? How can you determine if your results are enriched for a particular function? Try clicking on Analyze Results and explore the GO enrichment tool.

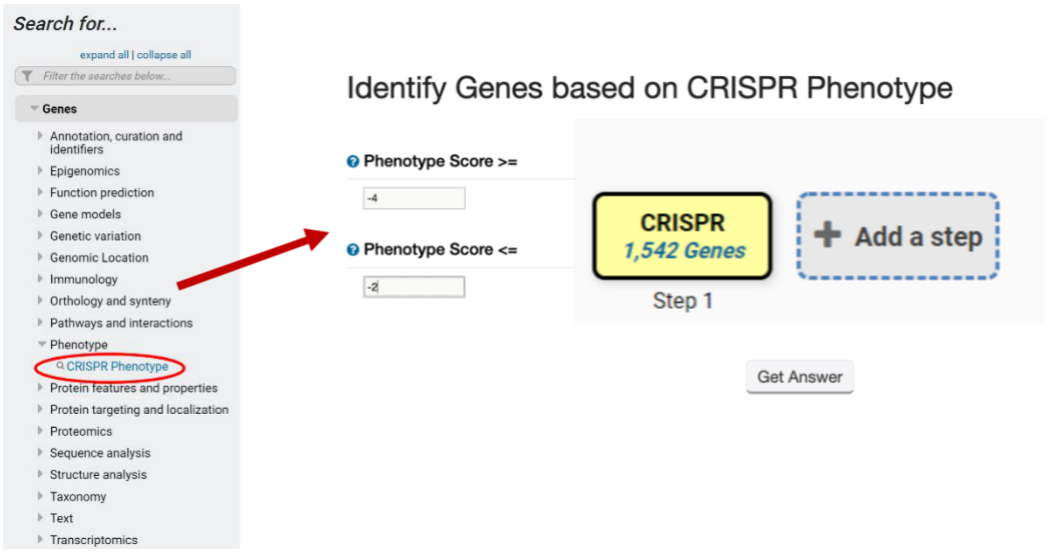

My Search Strategies
Analyze your Gene results with a tool below.
Gene Ontology Enrichment
Organism
Toxoplasma gondii ME49
Ontology
Cellular Component
Molecular Function
Biological Process
Evidence
Computed
Curated
select all | clear all
Limit to GO Slim terms
No
Yes
P-Value cutoff
0.05
(0 - 1)
Submit
Analysis Results:

9. Finding genes based on high throughput mutagenesis and fitness analysis.

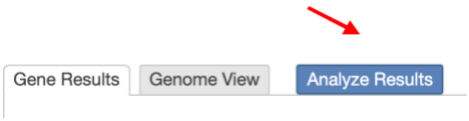
In EuPathDB we have a variety of studies where genome scale phenotypic analyses were carried out. In this exercise we'll use ToxoDB.org and look at fitness following CRISPR mutagenesis. You could also explore phenotyping studies in PlasmoDB or FungiDB if you prefer, the principles are the same.

- Navigate to the CRISPR phenotype search. Note that this search form is quite simple just requiring a range of fitness values. The defaults return all genes not limiting the search at all. This is only useful in as much as it tells you which genes were assayed which is nearly the entire genome. The tricky bit is deciding where to make the cutoffs. Again, the description on the search form is very helpful in this regard (as is the link to the paper … remember these phenotypes were assayed under specific conditions so just because a particular gene doesn't show a phenotype doesn't mean it wouldn't in other conditions (or infecting an actual host). The plot showing the phenotype score (fitness) is particularly useful. Red points along the plot are genes known to be essential under these conditions while yellow are known to be expendable. This will help you determine where to set the values. The last essential gene has a fitness score just >= than -4 so setting the phenotype score <= -2 would provide a pretty stringent search but still return more than 1000 genes. Try it. Do you get the expected results based on the number of genes returned?

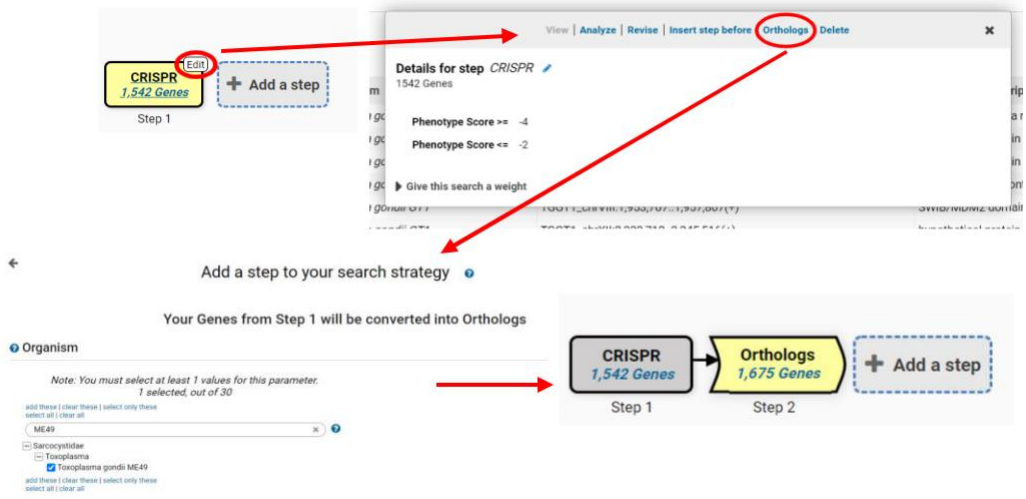




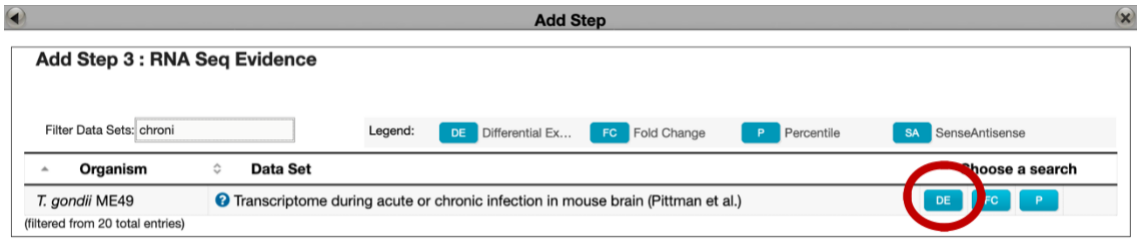
- Can you find additional evidence that these genes are essential? One way is to use the analysis tools to assess biological process and go function. Are the results what you would expect?



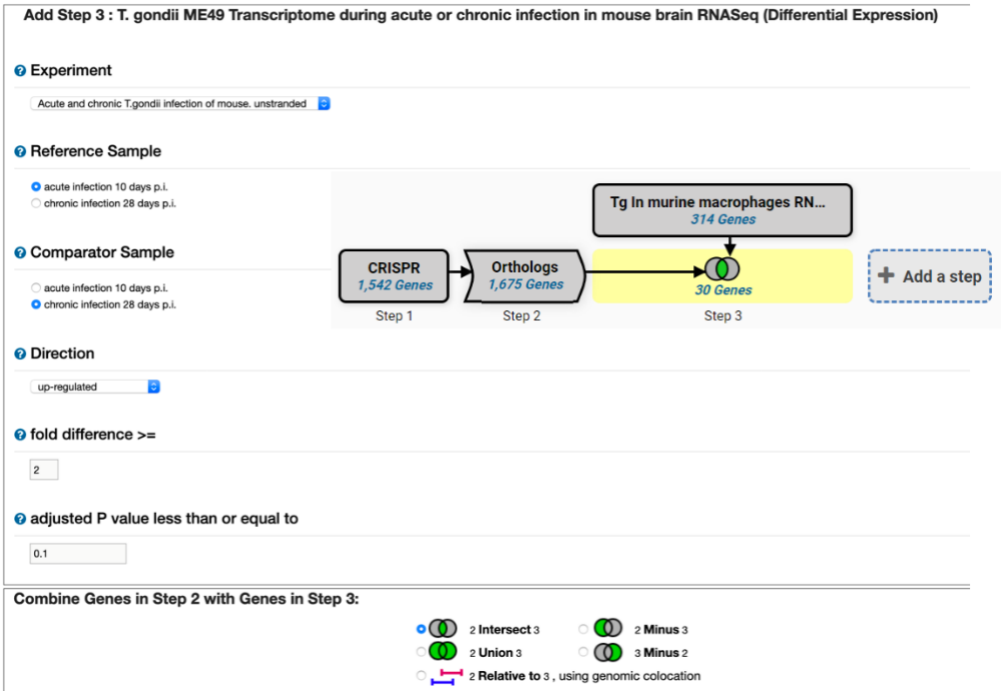
- Try adding columns to show additional data or intersecting these results with other queries, perhaps expression queries, to further assess this list. NOTE: this experiment was done in GT1 while all *T. gondii* functional data in ToxoDB is mapped to ME49 so an ortholog transform to ME49 is required before adding any additional functional studies.
- To do this, click on add step and select the Transform to orthologs option and select *T. gondii* ME49 to transform to.



- How many of these genes are upregulated in *in vivo* chronic stages of *T. gondii*?
  - Click on add step
  - Select the RNAseq searches under the Transcriptomics category
  - Find the experiment with chronic stages and run a search based on differentially expressed genes (DE).



- Intersect genes that are 2-fold upregulated in chronic stages compared to acute stages.
- What do these results look like? Do you find any interesting genes?



Related sites of interest to our communities

- Previous EuPathDB Workshops
- Companion
- OrthoMCL
- GeneDB
- ModBase at UCSF
- Tetrahymena Genome
- The Arabidopsis Information Resource
- NAR Database Summary Paper Categories

VEuPathDB Publications and Citations

This PubMed filter provides a current list of the most recent publications about VEuPathDB resources

Recent publications include:

Omar Harb, Jessica C. Kissinger and David S. Roos on behalf of the EuPathDB group
ToxoDB: the functional genomic resource for *Toxoplasma*
Toxoplasma gondii 3rd edition, Edited by Louis Weiss and Kami Kim (2020)
https://doi.org/10.1016/B978-0-12-815041-2.00023-2

Susanne Warrenfeltz and Jessica Kissinger on behalf of the EuPathDB Consortium
Accessing *Cryptosporidium* omic & isolate data via CryptoDB.org
Methods in Molecular Biology, Editor, Jan Mead (2020)
https://doi.org/10.1007/978-1-4939-9748-0_22

Omar S. Harb and David S. Roos on behalf of the EuPathDB Consortium
ToxoDB: Functional Genomics Resource for *Toxoplasma* and Related Organisms
Methods in Molecular Biology, Editor, Christopher J. Tonkin (2020)
https://doi.org/10.1007/978-1-4939-9857-9_2

Susanne Warrenfeltz, Evelina Y Basenko, Kathryn Crouch, Omar S. Harb, Jessica C. Kissinger, Achchuthan Shanmugasundram and Fatima Silva-Franco
EuPathDB: the eukaryotic pathogen genomics database resource
Methods in Molecular Biology, Editor, Martin Kollmar (2018)
https://doi.org/10.1007/978-1-4939-7737-6_5

Evelina Y. Basenko, Jane A. Pulman, Achchuthan Shanmugasundram, Omar S. Harb, Kathryn Crouch, David Starns, Susanne Warrenfeltz, Cristina Aurrecoechea, Christian J. Stoeckert, Jr., Jessica C. Kissinger, David S. Roos and Christiane Hertz-Fowler
FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. (2018)
J. Fungi 2018, 4(1), 39
https://doi.org/10.3390/jof4010039
Cristina Aurrecoechea; Ana Barreto; Evelina Y. Basenko; John Brestelli; Brian P. Brunk; Shon Cade; Kathryn Crouch; Ryan Doherty; Dave Falke; Steve Fischer; Bindu Gajria; Omar S. Harb; Mark Heiges; Christiane Hertz-Fowler; Sufen Hu; John Iodice; Jessica C. Kissinger; Cris Lawrence; Wei Li; Deborah F. Pinney; Jane A. Pulman; David S. Roos; Achchuthan Shanmugasundram; Fatima Silva-Franco; Sascha Steinbiss; Christian J. Stoeckert Jr; Drew Spruill; Haiming Wang; Susanne Warrenfeltz; Jie Zheng
EuPathDB: the eukaryotic pathogen genomics database resource
Nucleic Acids Research 2017 doi: 10.1093/nar/gkw1105

Warren, A. S., Aurrecoechea, C., Brunk, B., Desai, P., Emrich, S., Giraldo-Calderón, G. I., Harb, O., Hix, D., Lawson, D., Machi, D., Mao, C., McClelland, M., Nordberg, E., Shukla, M., Vosshall, L. B., Wattam, A. R., Will, R., Yoo, H. S., & Sobral, B.
RNA-Rocket: an RNA-Seq analysis resource for infectious disease research

Bioinformatics, 1 May 2015; 31(9), 1496–1498
https://doi.org/10.1093/bioinformatics/btv002

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S; VectorBase Consortium, Madey G, Collins FH, Lawson D.
VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases
Nucleic Acids Res. 2015 Jan;43(Database issue):D707-13
https://doi.org/10.1093/nar/gku1117

Publications that use our resource

Our project resources are cited in >20,000 publications. You can view the **publications that cite us** in Google Scholar.

Release Notes

Release notes are generated for each project site within VEuPathDB with each new release. They are located in the “New” tab on the right-hand expandable panel of each project site. An example for VectorBase is shown below.

Example release notes - VectorBase 48 Released

(*27 Aug 2020*)

We are pleased to announce the release of VectorBase 48.

Beginning with VectorBase 48, the URL https://VectorBase.org directs you to the newest version of our interface. The site contains previous VectorBase.org data plus some new tools and functions in an updated and streamlined interface that empasizes easy access to help information. Navigate to https://legacy.VectorBase.org for the previous version of the interface. The legacy interface will remain active for at least one release.

Release 48 Webinar

Join us on September 3, 2020 at 10am EDT for our Release 48 Highlights webinar where we will demonstrate and discuss new data and features in VEuPathDB 48. Register here

New features in VectorBase 48

- Gene pages: The sequence section now contains links to copy the sequence to your clipboard. Sequences are copied in FASTA format with the gene ID in the def line.
- Gene pages: It is now possible to scroll and zoom within the JBrowse views that are present in the 'Gene models' and 'Protein features and properties' sections.
- Contact Us form: A new option on the Contact Us form makes it possible to paste a screenshot into the form when sending us questions or suggestions.
- Search result page: The Download and Add to Basket options from a search result tab now appear as buttons for better visibility.
- Error messages: The text in our error messages was reviewed and edited to improve clarity. And we've made it easier to report errors by adding links in error messages to the Contact Us form.

Omics data sets

- Fifteen additional microarray datasets are available in the merged VectorBase site. These are marked as 'New' in the Identify Genes based on Microarray Evidence search page.
- The results of nine comparative genomic study analyses (VCF files) are availible in JBrowse. The data are categorized under Genetic Variation in the Select Tracks tool.

Population Biology

This release represents our second largest abundance data release ever – over 219,000 new non-zero abundance records. It is also our largest increase in pathogen status assays, doubling our records to over 174,000 assays.

45 projects have been added. You can use any of the Project IDs from the list below, in the Search box of the MapVEu (formerly PopBio map) tool: https://vectorbase.org/popbio-map/web/

- Indoor Human landing catches from Western Cameroon (VBP0000626: map, 52 collections, 166 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Uganda by PRISM ICEMR group, 2018 (VBP0000627: map, 1113 collections, 7791 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Marion County, Indiana, USA. 2019. (VBP0000628: map, 1758 collections, 7539 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Manatee County, Florida, USA. 2017 (VBP0000629: map, 1845 collections, 10390 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Manatee County, Florida, USA. 2018 (VBP0000630: map, 1217 collections, 8462 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Manatee County, Florida, USA. 2019 (VBP0000631: map, 1859 collections, 11223 samples, 0 phenotypes, 0 genotypes)
- Maryland Dept. of Agriculture mosquito surveillance, Maryland, USA, 2016 (VBP0000632: map, 558 collections, 4373 samples, 0 phenotypes, 0 genotypes)
- Maryland Dept. of Agriculture mosquito surveillance, Maryland, USA, 2019 (VBP0000634: map, 757 collections, 7016 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Lucas County, Ohio, USA. 2018 (VBP0000635: map, 1438 collections, 6626 samples, 0 phenotypes, 0 genotypes)
- Supplemental Maryland Dept. of Agriculture mosquito surveillance, Maryland, USA, 2016 (VBP0000636: map, 45 collections, 259 samples, 0 phenotypes, 0 genotypes)

- Mosquito surveillance in Lucas County, Ohio, USA. 2019 (VBP0000637: map, 1554 collections, 7260 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Hernando County, Florida, USA. 2018 (VBP0000638: map, 361 collections, 2167 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Hernando County, Florida, USA. 2019 (VBP0000639: map, 513 collections, 2656 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2016 (VBP0000664: map, 1750 collections, 4745 samples, 0 phenotypes, 0 genotypes)
- Dynamic of resistance alleles of two major insecticide targets in Anopheles gambiae (s.l.) populations from Benin, West Africa (VBP0000640: map, 28 collections, 28 samples, 54 phenotypes, 0 genotypes)
- Contrasting resistance patterns to type I and II pyrethroids in two major arbovirus vectors Aedes aegypti and Aedes albopictus in the Republic of the Congo, Central Africa (VBP0000641: map, 7 collections, 17 samples, 45 phenotypes, 12 genotypes)
- Maryland Dept. of Agriculture mosquito surveillance, Maryland, USA, 2017 (VBP0000633: map, 614 collections, 5500 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Norfolk County Mosquito Control, in Massachusetts, USA, 2019 (VBP0000644: map, 934 collections, 4297 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Desplaines Valley Mosquito Abatement, Illinois, USA, 2019 (VBP0000645: map, 3006 collections, 25974 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in Salt Lake County, Utah, USA. 2018 (VBP0000646: map, 1585 collections, 6134 samples, 0 phenotypes, 0 genotypes)
- Mosquito infection assays for Plasmodium falciparum in Uganda by PRISM ICEMR group, 2015-2017 (VBP0000647: map, 7621 collections, 92524 samples, 92524 phenotypes, 0 genotypes)
- Mosquito infection assays for Plasmodium falciparum in Uganda by PRISM ICEMR group, 2018 (VBP0000648: map, 379 collections, 1497 samples, 1497 phenotypes, 0 genotypes)
- Mosquito surveillance in Uganda by PRISM ICEMR group, 2017 - 2019 (VBP0000649: map, 2706 collections, 6791 samples, 0 phenotypes, 0 genotypes)
- Susceptibility of An. gambiae s.l. from Côte d'Ivoire to Insecticides used on Insecticide-Treated Nets: Evaluating the Additional Entomological Impact of Piperonyl Butoxide and Chlorfenapyr (VBP0000650: map, 15 collections, 165 samples, 165 phenotypes, 0 genotypes)
- Mosquito abundance in Central Mozambique (VBP0000651: map, 28 collections, 55 samples, 0 phenotypes, 0 genotypes)
- Spatial and temporal distribution of Anopheles arabiensis larvae (Sudan) (VBP0000652: map, 3293 collections, 3296 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2017 (VBP0000653: map, 1623 collections, 5221 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2019 (VBP0000654: map, 1376 collections, 4778 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2007 (VBP0000655: map, 602 collections, 1408 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2008 (VBP0000656: map, 692 collections, 1627 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2009 (VBP0000657: map, 831 collections, 2389 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2011 (VBP0000659: map, 1064 collections, 3050 samples, 0 phenotypes, 0 genotypes)

- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2012 (VBP0000660: map, 1482 collections, 3972 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2013 (VBP0000661: map, 1819 collections, 5103 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2014 (VBP0000662: map, 1488 collections, 5151 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2015 (VBP0000663: map, 1625 collections, 4856 samples, 0 phenotypes, 0 genotypes)
- Montana State University mosquito data for 2015 (VBP0000666: map, 223 collections, 1458 samples, 0 phenotypes, 0 genotypes)
- Montana State University mosquito data for 2018 (VBP0000667: map, 239 collections, 1640 samples, 0 phenotypes, 0 genotypes)
- Montana State University mosquito data for 2016 (VBP0000668: map, 283 collections, 1790 samples, 0 phenotypes, 0 genotypes)
- Montana State University mosquito data for 2019 (VBP0000669: map, 80 collections, 639 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance in South Walton County, Florida, USA. 2019 (VBP0000670: map, 856 collections, 5264 samples, 0 phenotypes, 0 genotypes)
- Mosquito surveillance, Ada County Weed, Pest, and Mosquito, in Idaho, USA, 2010 (VBP0000658: map, 798 collections, 1976 samples, 0 phenotypes, 0 genotypes)
- Larval habitats seasonality and species distribution (VBP0000671: map, 4278 collections, 8866 samples, 0 phenotypes, 0 genotypes)
- Comparative toxicity of larvicides and growth inhibitors on Aedes aegypti from select areas in Jamaica (VBP0000642: map, 6 collections, 238 samples, 238 phenotypes, 0 genotypes)
- Variation in Malaria Transmission Dynamics in Three Different Sites in Western Kenya (VBP0000665: map, 133 collections, 1701 samples, 0 phenotypes, 0 genotypes)

Note: Our bimonthly database releases incorporate new data and correct errors in old data when necessary. Changes in annotation and new experimental data may slightly alter your search results by increasing or decreasing the number of hits. When search parameters change with a new release, we invalidate (Ø) the search and ask you to rerun it. When IDs are updated or removed, we map the old IDs to the new ones, remove the old IDs from your Basket, and leave your Favorites page alone.

Analyses methods

VEuPathDB draws data from many sources. To facilitate comparisons across data sets, we analyze all data with standardized, data type-specific analyses. All data of one type are analyzed with the same workflow. Although our results may show some differences from an author's publication, our re-analysis of the data makes it feasible to compare data sets from very different sources and to update the data analysis with contemporary methods. For transparency, the methods we use to analyze data are presented below. They are also located online at: *Link: https://veupathdb.org/veupathdb/app/static-content/methods.html*

Genome analyses

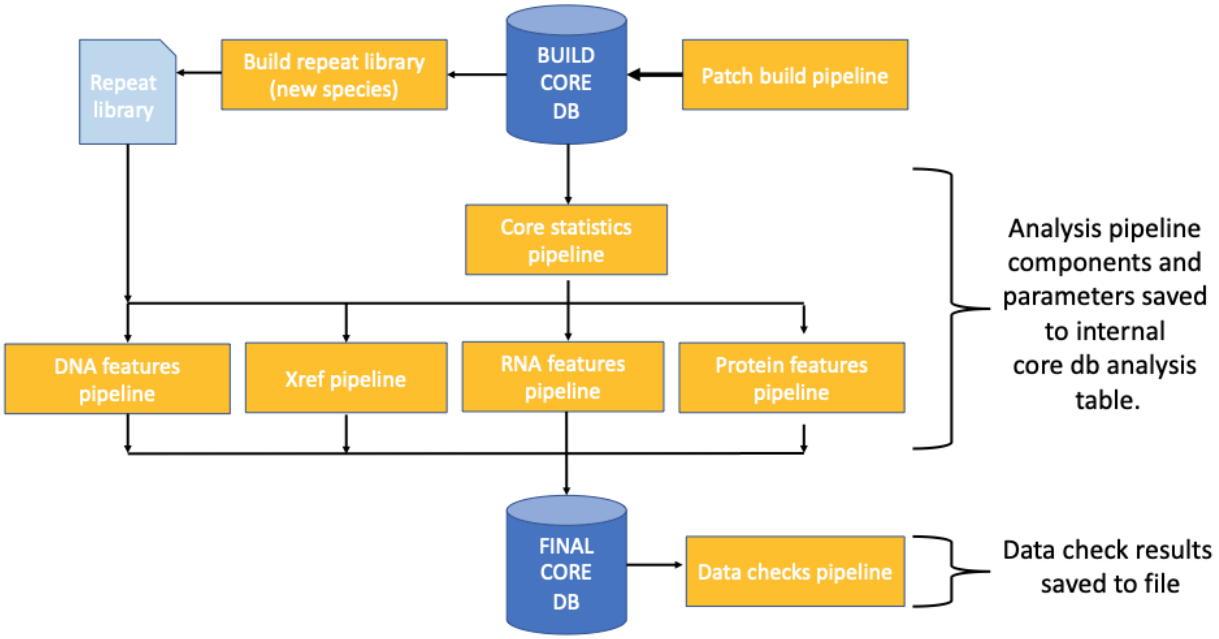
VEuPathDB employs the Ensembl genome analysis pipelines for analyzing genomic sequence to enhance annotations. While most of the genomic sequence (FASTA) are integrated into VEuPathDB from an INSDC repository, genome annotation (GFF3) may come from either the INSDC repository or a community submission.

Core database pipelines (next figure)- Primary genomic sequence and structural annotation data are loaded into a core database and run through 6 pipelines: core statistics, DNA feature annotation, external cross reference annotation, RNA gene annotation, repeat feature annotation, and protein feature annotation. The main pipelines applied to the core database and their components are listed in table 1.

Table 1 - Core database analysis pipelines and hive components.

Pipeline	Hive pipeline component	Git repository link
Core db build	Bio::EnsEMBL::EGPipeline::LoadGFF3::LoadGFF3	Under code review, details coming soon
Protein features	InterProScan	https://github.com/Ensembl/ensembl-production
DNA features	Bio::EnsEMBL::EGPipeline::DNAFeatures::*	Under code review, details coming soon
RNA features	Bio::EnsEMBL::EGPipeline::RNAFeatures::*	Under code review, details coming soon
Xrefs	Bio::EnsEMBL::EGPipeline::Xref::*	Under code review, details coming soon
Core Statistics	shortnoncodingdensity	https://github.com/Ensembl/ensembl-production
Core Statistics	snpdensity	https://github.com/Ensembl/ensembl-production
Core statistics	codingdensity	https://github.com/Ensembl/ensembl-production
Core statistics	percentgc	https://github.com/Ensembl/ensembl-production
Core statistics	percentagerepeat	https://github.com/Ensembl/ensembl-production

Core statistics	pseudogenedensity	https://github.com/Ensembl/ensembl-production
Core statistics	longnoncodingdensity	https://github.com/Ensembl/ensembl-production



Core database analysis pipelines and hive components

Example ehive pipelines, modules, programs and parameter data from coredb analysis table

Table 2 - Example ehive pipelines, modules, programs and parameter data from coredb analysis table.

Pipeline	Program	Program version	Parameters	Ehive module (Bio::Ens EMBL::)	Database	Database version
DNA features	dustmasker	NULL	NULL	Analysis::Runnable::D		

				ustMasker		
DNA features	trf	4	2 5 7 80 10 40 500 -d -h	Analysis::Runnable::TRF		
DNA features	RepeatMasker	4.0.5	-nolow -gccalc -species "Aedes aegypti" - engine crossmatch -q	Analysis::Runnable::RepeatMasker		
DNA features	RepeatMasker	4.0.5	-nolow -gccalc -lib “location1” -engine crossmatch -q	Analysis::Runnable::RepeatMasker		
DNA features	RepeatMasker	4.0.5	-nolow -gccalc -lib “location2” -engine crossmatch -q	Analysis::Runnable::RepeatMasker		
	NULL	NULL	NULL	EGPipeline::LoadGFF3::LoadGFF3		
RNA features	Infernal	1.1		Analysis::Runnable::CMScan		
RNA features	tRNAscan-SE	1.23		Analysis::Runnable::tRNAscan		
RNA features	Infernal	1.1		Analysis::Runnable::CMScan		
RNA features	rfam_12.2_gene	NULL	NULL	EGPipeline::RNAFeatures::CreateCmscanGenes		
RNA features	mirbase_gene	NULL	NULL	EGPipeline::RNAFeatures::CreateMirbaseGenes		
RNA features	trnascan_gene	NULL	NULL	EGPipeline::RNAFeatures::CreateTrnascanGenes		
Xref	xrefchecksum	NULL	NULL	EGPipeline::Xref::LoadUniParc		
Xref	xrefuniparc	NULL	NULL	EGPipeline::Xref::LoadUniProt		
Xref	gouniprot	NULL	NULL	EGPipeline::Xref::LoadUniProtGO		
Xref	xrefuniprot	NULL	NULL	EGPipeline::Xref::LoadUniProtXrefs		

DNA features	blastp	NULL	-word_size 3 -num_alignments 100000 -num_descriptions 100000 -lcase_masking -seg yes -num_threads 3	Analysis::Runnable::BlastEG		
DNA features	blastp	NULL	-word_size 3 -num_alignments 100000 -num_descriptions 100000 -lcase_masking -seg yes -num_threads 3	Analysis::Runnable::BlastEG		
Protein Features	InterProScan	5.37-76.0	NULL		Prosite patterns	2019_01
Protein Features	InterProScan	5.37-76.0	NULL		SFLD	4
Protein Features	InterProScan	5.37-76.0	NULL		CDD	3.17
Protein Features	InterProScan	5.37-76.0	NULL		Gene3D	4.2.0
Protein Features	InterProScan	5.37-76.0	NULL		HAMAP	2019_01
Protein Features	InterProScan	5.37-76.0	NULL		PANTHER	14.1
Protein Features	InterProScan	5.37-76.0	NULL		ncoils	2.2.1
Protein Features	InterProScan	5.37-76.0	NULL		Prosite profiles	2019_01
Protein Features	InterProScan	5.37-76.0	NULL		Pfam	32
Protein Features	InterProScan	5.37-76.0	NULL		PRINTS	42
Protein Features	InterProScan	5.37-76.0	NULL		Smart	7.1
Protein Features	InterProScan	5.37-76.0	NULL		SuperFamily	1.75

Protein Features	InterProScan	5.37-76.0	NULL		TIGRfam	15
Protein Features	InterProScan	5.37-76.0	NULL		InterPro- 2GO	NULL
Protein Features	InterProScan	5.37-76.0	NULL		PIRSF	3.02
Protein Features	InterProScan	5.37-76.0	NULL		SignalP	4.1
Protein Features	InterProScan	5.37-76.0	NULL		TMHMM	2.0c
Protein Features	InterProScan	5.37-76.0	NULL		Seg	NULL
Protein Features	InterProScan	5.37-76.0	NULL		MobiDBLite	2
Protein Features	InterProScan	5.37-76.0	NULL		InterPro-2Pathway	NULL

*Location 1 “/homes/jallen/scratch/vb/recent_assemblies/aedes_aegypti/RepeatModeler/aedes_aegypti.rm.lib"

*Location 2 "/nfs/panda/ensemblgenomes/vectorbase/data/tefam/aegypti_tefam.lib"

Supplements to the EBI Pipelines

VEuPathDB supplements the EBI pipeline with workflows that produce data for EST alignments, Open reading frames, and synteny (Table).

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Open reading frame generation: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

Synteny: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Details for the supplements to the EBI pipelines

Table 3: Details for supplements to the EBI Pipeline for genomes

Data	Program	Parameters or VEuPathDB GitHub repository	Version
EST alignments	BLAT	-nohead -maxIntron=1000 -t=dna -q=dna -dots=10	35
Open Reading Frames	orfFinder	--minPepLength 50 https://github.com/VEuPathDB/.../orfFinder	
Synteny	Mercator and MAVID	-p <PATH TO MERCATOR DIRECTORY> -t <tree string> -m <MAVID_EXE> -c <CNDSRC_DIR> -d draftGenome1... -d draftGemoneN -n nonDraftGeome1... -n nonDraftGenomeN -r referenceGenome https://github.com/VEuPathDB/.../runMercator	Mercator mapmaker 0.4 (2016-01-21) Mavid Version 2.0.4

In-house genome analyses in lieu of the EBI Pipelines

On rare occasions the EBI pipeline cannot be applied to a genome. For example, genomes that are not housed at an INSDC repository cannot be analyzed by the EBI pipeline. VEuPathDB uses the following in-house analyses in lieu of the EBI pipeline.

BLAT against NRDB: For every genome, VEuPathDB runs BLAT alignments of the annotated proteins against the GenBank Non-Redundant Protein Sequence Database (NRDB) to identify possible relationships and alignments outside the scope of VEuPathDB-supported organisms.

Compute open reading frames: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

DNA repeats: The Tandem Repeats Finder program locates and displays tandem repeats in genomic sequences.

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Protein domain annotations: InterProScan scans protein sequences against the protein signatures of the InterPro member databases and generates a file containing the domain matched, description of the InterPro entry, GO descriptions and E-values.

Signal peptide prediction: Signal P is used to identify signal peptides and their likely cleavage sites. A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined towards the secretory pathway.

Syntenic sequences: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Transmembrane domain prediction: TMHMM is used to predict transmembrane domain presence and topology from protein sequences.

tRNA gene prediction: tRNAScan identifies transfer RNA genes in transcript or genome sequences.

Details for the VEuPathDB in-house pipelines

Table 4: Details for genome analyses in lieu of EBI Pipeline

Data	Program	Parameters or VEuPathDB GitHub repository	Version
BLAT against NRDB	BLAT	-nohead -maxIntron=1000 -t=dnax -q=prot -dots=10 -minScore=25 -minIdentity=20	35
Computed Open Reading Frames	orfFinder	--minPepLength 50 https://github.com/VEuPathDB/.../orfFinder	
DNA Repeat regions	Tandem Repeats Finder	2 7 7 80 20 50 500	4.04
EST alignments	BLAT	-nohead -maxIntron=1000 -t=dna -q=dna -dots=10	35
Signal peptide predictions	SignalP	-t euk -f short -m nn+hmm -q -trunc 70	3.0

Synteny	Mercator and MAVID	-p <PATH TO MERCATOR DIRECTORY> -t <tree string> -m <MAVID_EXE> -c <CNDSRC_DIR> -d draftGenome1... -d draftGemoneN -n nonDraftGeome1... -n nonDraftGenomeN -r referenceGenome https://github.com/VEuPathDB/.../runMercator	Mercator mapmaker 0.4 (2016-01-21) Mavid Version 2.0.4
Transmembrane domain prediction	TMHMM	Nice -short	2.0c

Proteomics

VEuPathDB integrates the results of proteomics experiments as peptides aligned to a reference genome or as abundance data assigned to a gene. We do not reanalyze the raw mass spec data but instead use an in-house plugin that loads found peptides or abundance data from tab delimited input files of a specific format.

Details for the VEuPathDB in-house proteomics pipeline

RNA-Sequence

VEuPathDB integrates RNA-Seq data from many different experiments and analyzes all data with the same EBI RNA-Seq analysis pipeline. The RNA sequence data that we integrate is processed at EBI.

The following is a general outline of the analysis process.

- Trim poor quality data (Trimmomatic)
- HiSAT2 alignment to a reference genome
- HT-Seq-count to tally aligned reads per gene
- Convert to transcripts per kilobase million (TPM)
- DESeq2 to determine differential expression

EBI RNA-Seq pipeline details

ChIP-Sequence

VEuPathDB integrates ChIP-Seq data from many different experiments and sources. DNA seq data are aligned to the reference genome using Bowtie2. Alignment results are converted to bigwig and displayed in JBrowse.

Copy Number Variation

VEuPathDB uses coverage from whole genome sequencing data to estimate gene and chromosome copy numbers in sequenced strains. The bowtie2 alignments generated during SNP analysis are used as a starting point. HTseq-count is used to count the number of reads that align to each gene and the values are converted to transcripts per million (TPM). Assuming that the median TPM value represents a single copy gene on a chromosome of constitutive ploidy, we can infer gene or chromosome duplications by comparing the TPM values for individual genes or the median TPM for individual chromosomes to the whole genome median using custom scripts based on the method described in PMID: 22038252. Additionally, coverage is calculated in 1 kb bins across the genome, normalised to the constitutive ploidy and converted to bigwig format for visualisation in JBrowse.

Haploid number and gene dose are metrics used to define copy number in VEuPathDB. Haploid number is the number of genes on an individual chromosome. Gene dose is the total number of genes in an organism, accounting for copy number of the chromosome. For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2.

Searches for genes based on Copy Number Variation

There are two searches that query copy number data in VEuPathDB. The first, Identify Genes based on Copy Number (CNV) returns genes that are present at copy numbers within a range that you specify. The search can be configured to return genes based on the haploid number or gene dose. The second search, Identify Genes based on Copy Number Comparison (CNV), returns genes for which the copy number varies between the reference and your chosen isolates. This search compares the estimated copy number of a gene in the resequenced strain(s) with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are both on the same chromosome and in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor. In this search, the metric for copy number is the haploid number, which is the number of copies of a gene on a single chromosome.

Genetic Variation and SNP calling

VEuPathDB analyzes whole genome resequencing data to call single nucleotide polymorphisms of isolates. The method employed by VEuPathDB to call SNPs from short read sequencing like Illumina reads, follows these steps:

- Reads are aligned to the reference genome using bowtie2
- The resulting BAM file from bowtie2 is sorted and a pileup file using samtools is generated
- Reads around indels are realigned using GATK
- SNPs and indels are called consensus sequence using VarScan is generated:
  - P value <= 0.01
  - minimum aligned reads >= 5
  - minimum read frequency >= 0.8
- SNP calls where coverage is >2.5x the median coverage are removed to limit erroneous calls in repeat regions
- At each SNP position “like reference” calls are generated for each strain that is identical to the reference to give the full picture of each SNP

Microarray data

VEuPathDB integrates microarray data from high density oligonucleotide as well as spotted arrays. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. We process the data we receive according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

Protein Array data

VEuPathDB integrates protein array data from serm antibody microarray experiments. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. Although each experiment and data set can have special considerations, we process the data according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

Metabolic Pathways

VEuPathDB integrates metabolic pathways from KEGG and MetaCyc. For TriTrypDB, pathways are also integrated from LeishCyc and TrypanoCyc. Metabolic pathways are associated with genes via Enzyme Commission annotations.

Dataset descriptions

This document provides a high-level overview of the software infrastructure utilized by the VEuPathDB BRC to load, integrate and provide data to users. Please check a list of all the data sets loaded in our VEuPathDB sites utilizing this infrastructure.

Link: https://veupathdb.org/veupathdb/app/search/dataset/AllDatasets/result

All loaded datasets are described appropriately with links to publications and repositories as appropriate.

- Data Set (Name)

- Organism(s) (source or reference)
- VEuPathDB Project
- Release # /Date
- Category
- Publications
- Summary
- Contact

Organisms - Genome info and stats

This document provides a comprehensive overview of the organisms and genome sequence versions and annotation versions within the VEuPathDB BRC.

Link: https://veupathdb.org/veupathdb/app/search/organism/GenomeDataTypes/result

All genome sequences and annotations are searchable and described appropriately with links to VEuPathDB organism pages.

- VEuPathDB project
- Organism
- NCBI taxon ID
- Species NCBI taxon ID
- Genome source
- Genome version
- Genome size (Mbp)
- Annotation version
- Total gene count
- Is reference strain indicator?
- Contig count
- Supercontig count
- Chromosome count
- Protein coding gene count
- Non protein coding gene count
- Pseudogene count
- User comment count
- Ortholog count
- Go Terms count

Technical infrastructure and software documentation:

Link: *https://beta.veupathdb.org/veupathdb.beta/app/static-content/infrastructure.html*

The above document describes the overall infrastructure of our data loading system and websites with links to appropriate GitHub repositories for users who want to explore using VEuPathDB infrastructure to build their own genomics database and website.

Browser Compatibility Statement

We recognize that our users access VEuPathDB using various Internet Browsers and Operating Systems. Our goal is to ensure that you have the best possible experience on VEuPathDB, but it is impossible to develop applications that work identically, efficiently and effectively on all web browsers.

Based on our site usage statistics we support the following browsers used by greater than 95% of our visitors:

- Firefox
- Safari
- Chrome

Feel free to contact us about any browsing issues you might come across.

Data Loading and Database Schema

We use the **G**enomics **U**nified **S**chema (GUS) database schema and data loading infrastructure and its framework available at GusAppFramework. This includes not only a comprehensive database schema for integrating and representing genomic and functional (or post) genomic data but also tools for loading said data into that system. We have made some extensions to the schema and tools for VEuPathDB specific purposes primarily to generate de-normalized views of the data for query optimization purposes.

Our data are all stored in Oracle12c databases. Our software infrastructure also supports PostgreSQL but we have some Oracle specific SQL constructs in our model that would need to be changed in order to run successfully in PostgreSQL.

We load all data using an in house engineered workflow system called ReFlow. Briefly, ReFlow is engineered to be an efficient graph-based workflow system. In it each step (node in the graph) has the ability to be undone and subsequently rerun with updated data. This was a significant requirement as it enables us to undo entire genomes when the annotation or underlying sequence changes. This results in automated removal of all data dependent on that

genome. When the step is re-run with the new annotation, all dependent data are recomputed and reloaded automatically, thus greatly improving our ability to keep these complex databases up-to-date.

The ReFlow workflow system utilizes another piece of software developed at the University of Pennsylvania to schedule, manage and monitor running tasks called DistribJob. DistribJob distributes tasks generated from a large input dataset such as a set of sequences to compute nodes in a cluster for analysis and retrieves and collates the results in an efficient manner. We automate the running of large compute tasks on compute clusters located at the University of Pennsylvania and the University of Georgia.

Code Availability

To facilitate greater transparency and tool reuse, our codebase has been migrated to the GIT repository (https://github.com/VEuPathDB).

Web Presentation System and User Interfaces

Our websites are based on code that we developed and have released to the community called the Strategies-WDK (Strategies Web Development Kit) which enables the graphical strategies search system. You can download the software and see documentation for this toolkit at Strategies-WDK. This toolkit enables us to represent our data as an XML model which is then turned into the web interfaces that are presented to users using these tools.

Software Code Repository

To facilitate greater transparency and tool reuse, our codebase has been migrated to the GIT repository (https://github.com/VEuPathDB).

System Hardware and Third-Party Software

VEuPathDB maintains redundant database and content web servers at the University of Pennsylvania and the University of Georgia to minimize interruptions for our users during maintenance periods. Additionally VEuPathDB compute and data loading servers are located at the University of Pennsylvania.

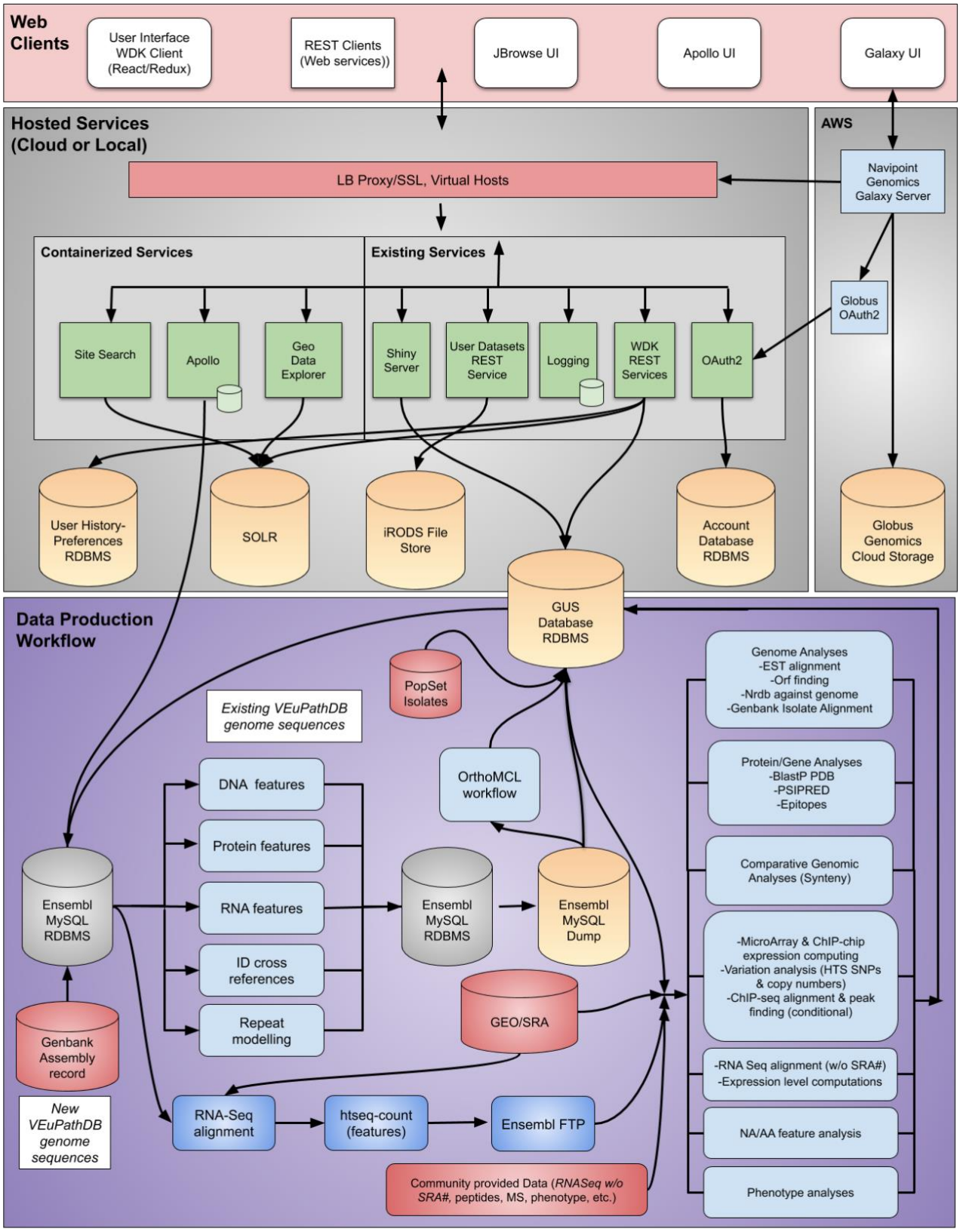
Server configurations are coordinated and deployed through Puppet automation software (http://puppetlabs.com/). Custom infrastructure software is versioned and deployed through standard RPM/YUM mechanisms. When appropriate, software builds are automated using Jenkins Continuous Integration Server (http://jenkins-ci.org/)

System infrastructure statistics (CPU load, I/O, etc) are gathered with collectd (http://collectd.org/) and in-house applications and feed to Graphite (http://graphite.wikidot.com/) for human review. Nagios (http://www.nagios.org/) provides notifications of system degradations.

Both Universities also maintain large compute clusters that are heavily utilized by VEuPathDB in order to analyze and load incoming data in a timely fashion. The linked document below describes our actual hardware and includes a list of third-party software required in order to analyze, load and present data via our websites.

Overview of the VEuPathDB Data Production Workflow and Architecture

The complete pathway from data acquisition to web presentation and utilization by users is detailed in the next figure. The data production workflow (bottom, purple rectangle) begins with data acquired from numerous sources (indicated in red) including direct deposition by the community. These data are processed through two main pipelines (indicated in blue), one targeting genome features and RNA-Seq mapping using Ensembl pipelines and the second focused on comparative genomics and analysis of other types of omics-scale data including proteomics and high-throughput phenotypic screens among others. Quality-controlled analysed data are loaded into the GUS database (indicated in gold spanning the purple data production and grey services rectangles). The GUS database links the data to the servers and services (green) and resources that allow the user community to securely access, interrogate, visualize, download and further analyse data, including their own data via the provided web clients (indicated in the pink rectangle).


Web Clients
User Interface WDK Client (React/Redux)
REST Clients (Web services))
JBrowse UI
Apollo UI
Galaxy UI
Hosted Services (Cloud or Local)
AWS
LB Proxy/SSL, Virtual Hosts
Navipoint Genomics Galaxy Server
Containerized Services
Existing Services
Globus OAuth2
Site Search
Apollo
Geo Data Explorer
Shiny Server
User Datasets REST Service
Logging
WDK REST Services
OAuth2
User History-Preferences RDBMS
SOLR
iRODS File Store
Account Database RDBMS
Globus Genomics Cloud Storage
GUS Database RDBMS
Data Production Workflow
PopSet Isolates
Existing VEuPathDB genome sequences
Genome Analyses
-EST alignment
-Orf finding
-Nrdb against genome
-Genbank Isolate Alignment
OrthoMCL workflow
DNA features
Protein features
RNA features
ID cross references
Repeat modelling
Protein/Gene Analyses
-BlastP PDB
-PSIPRED
-Epitopes
Comparative Genomic Analyses (Synteny)
Ensembl MySQL RDBMS
Ensembl MySQL RDBMS
Ensembl MySQL Dump
-MicroArray & ChIP-chip expression computing
-Variation analysis (HTS SNPs & copy numbers)
-ChIP-seq alignment & peak finding (conditional)
GEO/SRA
Genbank Assembly record
-RNA Seq alignment (w/o SRA#)
-Expression level computations
New VEuPathDB genome sequences
RNA-Seq alignment
htseq-count (features)
Ensembl FTP
NA/AA feature analysis
Community provided Data (RNASeq w/o SRA#, peptides, MS, phenotype, etc.)
Phenotype analyses

VEuPathDB Website Privacy Policy

We do not use or share any of your personal information for any purpose unrelated to the functionality of the websites; however, we do collect some information to help us understand how our sites are being used and to improve community support.

UPDATED: April 12, 2020

Introduction

VEuPathDB (also referred to as 'we' throughout this document) is committed to protecting its users' data and privacy. The purpose of this page is to provide you with information about how the data we collect from users of VEuPathDB websites is used or shared. We may update this Privacy Notice from time to time. We encourage you to visit this page frequently and take note of the date updated field above.

We do not use or share any of your personal information for any purpose unrelated to the functionality of the websites; however, we do collect some information to help us understand how our sites are being used in order to improve community support and to enhance the VEuPathDB community's experience when visiting our sites.

Information Automatically Collected

When you browse VEuPathDB sites, certain information about your visit will be collected. We automatically collect and store the following type of information about your visit:

- The IP address of the client making the request. Often the IP address is that of your personal computer or smart phone; however, it might be that of a firewall or proxy your internet provider manages.
- The operating system and information about the browser used when visiting the site.
- The date and time of each visit.
- Pages visited.
- The address of a referring page. If you click a link on a website that directs you to a VEuPathDB page, the address of that originating web page will be collected. This "referrer" information is transmitted as part of the browser and server communications; it is not based on any marketing or partnering agreements with the referring site.

This automatically collected information does not identify you personally unless you include personally identifying information in a support form request; see the "Contact Us" policy below for details. We use this information to measure the number of visitors

to our site. The aggregate data may be included in prospectuses and reports to funding agencies.

Information You Directly Provide

The Basket, Favorites, Public Strategies, Gene/Sequence Comment and GBrowse Track features of the VEuPathDB websites require that you register for an account. A valid email address is required so we can send you your temporary account password. An anonymous email service can be used if you do not want to provide personally identifying information.

Your email address will be used to send you infrequent alerts if you subscribe to receive them. We do not sell or distribute email addresses to third parties.

We also ask for your name and institution during account registration. If you add a comment to a Gene or a Sequence, your name and institution will be displayed with the comment. If you make one of your strategies public, your name and institution will be displayed with it. We do not routinely verify the validity of names and institutions associated with comments or public strategies; however, we will delete accounts or comments if we believe them to be fraudulent based on inappropriate activity or posted content. We will not sell or distribute your name or institution to third parties.

When you log in, the client IP address is recorded. This IP address can be correlated with the address automatically collected as noted above. If your user profile personally identifies you, then it may be possible to associate you with your detailed activity on VEuPathDB web sites.

“Contact Us” Form

The header on each web page includes a “Contact Us” link to a form where users can submit questions, error reports, feature requests, and dataset proposals. Submissions through this form are emailed to VEuPathDB staff and recorded in a project management application accessible only by VEuPathDB staff.

The form includes a field for an email address. If the email address identifies you personally, say if you use your institutional email, then your correspondence with us will likewise be linked to you. A valid email is not strictly required, although we cannot reply to you without one.

When you submit the form, your IP address and browser version will be recorded for internal use. In the case of reported bugs or other site errors, this information may be used by technical staff to help locate your session in the server logs to aid in troubleshooting the issue. This does have the side effect of making it possible to associate an IP address with an email address which may, in turn, personally identify you. However, VEuPathDB does not publicly release this information.

How VEuPathDB Uses Cookies

VEuPathDB uses cookies to associate multiple requests by your web browser into a stateful session. Cookies are essential to track the state of query strategies, gene baskets and authentication.

Some cookies persist only for a single session. The information is recorded temporarily and is erased when the user quits the session or closes the browser. Others may be persistently stored on the hard drive of your computer until you manually delete them from a browser folder or until they expire, which can be months after they were last used.

Cookies can be disabled in your browser (refer to your browser's documentation for instructions); however, the majority of the website functionality will be unavailable if cookies are disabled.

Google Analytics

Google Analytics provides aggregate measurements of website traffic including counts of page hits and unique users along with statistics on countries of origin.

The raw measurements and statistics are only available to approved VEuPathDB staff. Aggregated data may be included in prospectuses and reports to funding agencies.

Third-Party Websites and Applications

Third-party websites and applications are not exclusively operated or controlled by VEuPathDB. By using these third-party websites, individuals may be providing nongovernmental third-parties with access to personally identifying information.

Twitter

VEuPathDB maintains a presence on Twitter in the form of a VEuPathDB branded page. This page allows for a direct connection with end users to promote information related VEuPathDB services and to disseminate educational information on research publications, news and events related to the biology of eukaryotic pathogens. Postings may also include information about planned service maintenance and outages.

Twitter collects profile information such as name and email address about users who register to use this third-party website. Depending on the user's privacy settings, this information may be displayed on the user's profile page or in the user's tweets which may be retweeted on VEuPathDB's page. The VEuPathDB Twitter account may post the authors and institutions of publicly published scientific papers and news articles. VEuPathDB does not actively collect or maintain personally identifying information through its use of Twitter. VEuPathDB will redact or refrain from retweeting a posting that contains obviously identifiable personal information. A Twitter account is not

required to read VEuPathDB postings on Twitter. VEuPathDB does not collect or use personal information outside of Twitter's site.

Twitter is hosted and maintained by a third party which may use browser tracking and related technologies to collect information about visitors to twitter.com and its affiliates. Refer to Twitter's privacy statement, https://twitter.com/en/privacy, for more information.

Facebook

VEuPathDB maintains a presence on Facebook in the form of a VEuPathDB branded page. This page allows for a direct connection with end users to promote information related VEuPathDB services and to disseminate educational information on research publications, news and events related to the biology of eukaryotic pathogens. Postings may also include information about planned service maintenance and outages.

Like Twitter, Facebook collects profile information, including name and email address, from its users. Depending on the user's privacy settings, this information may be displayed on the user's profile page along with any activity such as comments or "likes" on the VEuPathDB Facebook page or in posts that VEuPathDB shares on Facebook. VEuPathDB does not collect or use any personally identifying information outside of our Facebook page. To understand how Facebook collects and uses personal information, refer to their data policy page, https://www.facebook.com/policy.php.

YouTube

VEuPathDB maintains a presence on YouTube in the form of a VEuPathDB branded page. This page provides tutorials on the use of our websites.

YouTube also requires some information when users create an account, including an email address, and users may choose to provide a name and other identifying information in their public profile. Depending on their individual privacy settings, some personally identifiable information may be available to other users, including VEuPathDB. However, VEuPathDB does not collect or use any of that information outside of its YouTube interactions. You can view videos without signing in to an account, but you must be a registered user in order to comment. As YouTube is a Google service, the VEuPathDB youtube channel is subject to Google's privacy policy.

Globus Genomics

The VEuPathDB Galaxy Data Analysis Service is a workspace for large-scale data analyses. Developed in partnership with Globus Genomics, workspaces offer a private analysis platform with published workflows and pre-loaded annotated genomes. The workspace is accessed through the "Analyze My Experiment" tab on the home page of any VEuPathDB resource and can be used to upload your own data, compose and run custom workflows, retrieve results and share workflows and data analyses with colleagues.

The VEuPathDB Galaxy Data Analysis Service is hosted by Globus Genomics, an affiliate of Globus. The first time you visit VEuPathDB Galaxy you will be asked to sign up with Globus in order to set up your private Galaxy workspace. Linking your Globus account with your VEuPathDB account is necessary so that input data and analysis results can be transferred between the two systems. We encrypt data transfers and storage, but ultimately, we cannot guarantee the security of data transmissions among VEuPathDB, Globus and affiliates, Amazon Cloud Services, and the user. It is your responsibility to back up your data and obtain any required permissions from your study and/or institution prior to uploading data for analyses on the VEuPathDB Galaxy platform. Do not use, transmit, upload or share any human identifiable information in the files you analyze. VEuPathDB, Globus and affiliates, the University of Georgia, the University of Pennsylvania, the University of Liverpool, and Amazon Cloud Services do not take any responsibility and are not liable for the loss and/or release of any data you analyze via the VEuPathDB Galaxy platform. We encourage you to review the Globus' privacy policy.

Your Rights based on the General Data Protection Regulation (GDPR)

To read more about GDPR please check the GDPR website.

1. The right of transparency and modalities. The privacy policy should be clear and easy to follow in explaining what data we collect and how we use it.
2. The right to be informed about when data is gathered. This is described in the privacy policy, during the registration process (if you choose to register), site banner and an email sent out to all registered users on May 25, 2018.
3. The right of access. You can ask for what sepecific data we have about you and how we use it.
4. The right to rectification. We will correct any errors in your personal data that you point out to us.
5. The right to be forgotten. We are happy to delete your account and info when you make such a request.
6. The right to restrict processing. You have the right to request that we restrict the use of your data.
7. The right for notification obligation regarding rectification/erasure/restriction.
8. The right to data portability.
9. The right to object to the processing of your personal data at any time.
10. The right in relation to automated decision making and profiling. Basically, you have the right not to be subject to decisions based solely on automated processing which significantly affect you.

To make any of the above stated requests or if you have any questions please email us at help@VEuPathDB.org

VEuPathDB Accessibility Conformance

We strive to make the VEuPathDB website accessible. Wherever possible, our sites have alternative text for images, the pages can be magnified and read by screen readers. The results of some searches are less accessible as they are dynamic and user-speciifc, thus alternative text describing images cannot be generated. We update our Voluntary Product Accessibility Template, VPAT, 508 accessibility report annually. The more recent version is located at: https://veupathdb.org/documents/VEuPathDB_Section_508_BRC4.pdf

VEuPathDB personnel

VEuPathDB Management

Beatrice Amos, Annotation Manager
Cristina Aurrecoechea, User Interface and Portal Manager
Bob Belnap, Systems and Databases Manager
John Brestelli, Data Development Manager
Brian Brunk, VEuPathDB Senior Manager
Mark Caddick, Wellcome Trust PI; Co-I NIAID BRC Contract
George Christophides, Co-I, NIAID BRC Contract
Kathryn Crouch, Co-I, Wellcome Trust
Jeremy DeBarry, Project Coordinator
Steve Fischer, Software and Infrastructure Manager
Paul Flicek, Co-I, NIAID BRC Contract
Omar Harb, Director of Scientific Outreach & Education
Jessica C Kissinger, Joint-PI, NIAID BRC Contract; WT Co-PI
Dan Lawson, Project Coordinator
Wei Li, Data Loading Manager
Mary Ann McDowell, Joint-PI, NIAID BRC Contract
David S Roos, Joint-PI, NIAID BRC Contract; WT Co-PI
Chris J Stoeckert, Co-I, NIAID BRC Contract

To contact any one of us please use the contact us form.

Current VEuPathDB Team members

Beatrice Amos[4], Rachel Ankirskiy[1], Cristina Aurrecoechea[1], Matthieu Barba[9], Ana Barreto[3], Evelina Basenko[4], Wojtek Bazant[2], Dan Beiting[2], Bob Belnap[1], Ulrike Böhme[5], John Brestelli[3], Brian Brunk[2], Mark Caddick[4], Danielle Callan[2], Mikkel Christensen[9],

George Christophides[8], Kathryn Crouch[6], Katie Cybulski[7], Elaine Daugan[4], Jeremy DeBarry[1], Ryan Doherty[3], Yikun Duan[2], Dave Falke[1], Steve Fischer[3], Paul Flicek[9], Bindu Gajria[2], Gloria I. Giraldo-Calderón[7], Omar S. Harb[2], Elizabeth Harper[2], Danica Helb[2], Mark Hickman[2], Connor Howington[7], Sufen Hu[2], Jay Humphrey[1], John Iodice[3], John Judkins[2], Sarah Kelly[8], Jessica C. Kissinger[1], Dae Kun Kwon[7], Kris Lamoureux[1], Daniel Lawson[8], Wei Li[2], Brianna Lindsay[2], Jamie Long[2], Bob MacCallum[8], Gareth Maslen[9], Mary Ann McDowell[7], Greg Milewski[2], Jarek Nabrzyski[7], David S. Roos[2], Samuel Rund[7], Steph Wever Schulman[2], Achchuthan Shanmugasundram[4], Vasili Sitnik[9], Drew Spruill[1], David Starns[4], Christian J. Stoeckert Jr.[3], Sheena Shah Tomko[2], Haiming Wang[1], Susanne Warrenfeltz[1], Robert Wieck[7], Mariann Winkelman[2], Lin Xu[2], Jie Zheng[3].

Previous VEuPathDB Team members, 2004-2020

Antelmo Aguilar[7], James Allen[9], Alexis Allot[9], Nora Besansky[7], Austin Billings[2], Sanjay Boddu[9], Steve Bogol[7], Ewan Birney[9], Andrew Brockman[8], Robert Bruggner[7], Ja'Shon Cade[3], David Campbell[7], Cristian Cocos[2], Frank Collins (VectorBase Principal Investigator, 2004-2018)[7], Kathy Couch[1], Greg Davies[7], Ale Diaz Miranda[2], Emmanuel Dialynas, Jennifer Dommer[3], Vicky Dritsou, Scott Emrich[10], Xin Gao[2], William Gelbart[12], Sandra Gesing[7], Alan Gingle[1], Greg Grant[3], Matt Guidry[1], Martin Hammond[9], Mark Heiges[1], Christiane Hertz-Fowler (Principal Investigator, WT 2008-2019)[4], Nicholas Ho[8], Daniel Hughes[9], Frank Innamorato[3], San James[14], Amie Jaye[8], Fotis Kafatos[8], Paul Kersey[9], Ioannis Kimitzoglou[8], Nathan Konopinski[7], Carolyn Knoll[2], Eileen T. Kraemer[1], Nick Langridge[9], Cris Lawrence[2], Neil Lobo[7], Christos (Kitsos) Louis[11], Ross Madden[6], Greg Madey[7], Elisabetta Manduchi[3], Karine Megy[9], John A. Miller[6], Elvira Mitraka[11], Vishal Nayak[3], Cary Pennington[1], Deborah F. Pinney[3], Brian Pitts[1], Jane A.Pulman[4], Caleb Reinking[7], Seth Redmon, Chris Ross[1], Andrew Sheehan[7], Fatima Silva[4], Ganesh Srinivasamoorthy[1], Scott Szakonyi[7], Pantelis Toplais[11], Ryan Thibodeau[1], Charles Treatman[2], Betsy Wenthe[1], Matt Vander Werf[7], Maggie Werner-Washburne[13], Patricia L. Whetzel[3], Derek Wilson[9], Andrew Yates[9]

[1]University of Georgia, Athens, GA 30602, USA
[2]University of Pennsylvania, Philadelphia, PA 19104, USA
[3]University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
[4]University of Liverpool, UK
[5]Wellcome Sanger Institute, Hinxton, CB10 1RQ, UK
[6]Wellcome Centre for Integrative Parasitology, University of Glasgow, UK
[7]University of Notre Dame, Notre Dame, IN 46556, USA
[8]Imperial College London, South Kensington, London SW7 2BU, UK
[9]European Bioinformatics Institute, Hinxton, CB10 1SD, UK
[10]University of Tennessee, Knoxville, TN 37996, USA
[11]Institute of Molecular Biology and Biotechnology-FORTH, Heraklion, Crete, Greece

[12]Harvard University, Cambridge, MA 02138, USA
[13]University of New Mexico, Albuquerque, NM 87131, USA
[14]Makerere University and Infectious Diseases Research Collaboration (IDRC), Kampala, Uganda

VEuPathDB acknowledgements

- All the community members who have contributed data (often pre-publication), entered user comments or sent us their suggestions.
- Scientists who provided images to be used as a template for the logos in our websites and for images used in the header section of our sites:

AmoebaDB:	**MicrosporidiaDB:**	**ToxoDB:**
William Petri	Gira Bhabha	David Roos
Serge Ankri	Pattana Jaroenlak (Michael)	**TrichDB:**
Craig Roberts	Damian Ekiert	Antonio Pereira-Neves
Fiona Henriquez	Michael Cammer	Marlene Benchimol
Hugo Aguilar-Díaz	**PiroplasmaDB:**	**TriTrypDB:**
Julio C. Carrero	Ellen Yeh	Rick Tarleton
CryptoDB:	Lowell Kappemeyer	Richard Wheeler
Boris Striepen	Audrey Lau	Leandro Lemgruber Soares
FungiDB:	Dirk Dobbelaere	Margaret Mullin
ZyGoLife	Manoj Duraisingh	Camila Silva Gonçalves
Jason Stajich	Brendan Elsworth	Wanderley de Souza
Zachary Lewis	Caroline Keroack	Maria Cristina Machado Motta
GiardiaDB:	Isabelle Coppens	
Fran Gillin	**PlasmoDB:**	
Tineke Lauwaet	Lawrence Bannister	
Barbara Davids	Lewis Tilney	
Scott Dawson	Pedro Moura	

VEuPathDB Community Representatives

VEuPathDB encourages community members to provide feedback about our resources. We get feedback from many community members including those listed below who have been active in our open community meetings. We encourage you to get involved. Feel free to contact us any time.

Amoeba: Open community call and Carol Gilchrist, Upi Singh
Cryptosporidium: Gregory Buck, Guy Robinson, Karin Troell, Sumiti Vinayak, Jonathon Wastling, Giovanni Widmer, Lihua Xiao, Guan Zhu
Fungi: Bridget Barker, Elaine Bignell, Katherine Borkovich, Michael Bromley, Christina Cuomo, Tamara Doering, Jay Dunlap, Michael Freitag, Louise Glass, Kim Hammond-Kosack, Guilhem Janbon, Seogchan Kang, Theo Kirkland, Corby Kistler, Jennifer Lodge, Robin May, Jessie Uehling, Sinem Beyhan, Douglas Lake, Natalie Mitchell, Maureen Donlin, Vera Meyer, Marc Orbach, Nadia Ponts, Antonis Rokas, Jason Stajich, Matt Sachs, George R. Thompson, Martin Urban, Nathan P. Wiederhold
Giardia: Scott Dawson, Fran Gillin, Adrian Hehl, Aaron Jex, Hilary Morrison, John Samuelson, Cornelia Spycher, Staffan Svard
Microsporidia: James Becnel, Nicolas Corradi, Elizabeth Didier, Patrick Keeling, Emily Troemel, Louis Weiss
Piroplasma: Open community call and Choukri Mamoun
Plasmodium: John Adams, Chris Janse, Rays H.Y. Jiang, Shahid Khan, Stuart Ralph, Akhil Vaidya, Andy Waters
Toxoplasma: John Boothroyd, Jon P. Boyle, Vern B. Carruthers, Marc-Jan Gubbels, Kami Kim, Markus Meissner, Jeroen Saeij, Lilach Sheiner, Ross Waller, Michael White
Trichomonas: Jane Carlton, Patricia Johnson, Steven Sullivan, Jan Tachazy
Trypanosoma/Kinetoplastids: Fernan Aguero, Vivian Bellofatto, Richard Burchmore, George Cross, Angela Cruz, Antonio Estevez, Mark Field, Catarina Gadelha, Eva Gluenz,Keith Gull, John Kelly, Annette MacLeod, Jeremy Mottram, Torsten Ochsenreiter, Marc Ouellette, Barbara Papadopoulou, Laurie Read, Sergio Schenkman, Rick Tarleton, Brent Weatherly, Bill Wickstead, Michael (Mick) Urbaniak
Vectors: Gregory Dasch, Jeff Grabowski, María de Lourdes Muñoz, Monika Gulia-Nuss, Sukanya Narasimhan, Kristin Michel, Michael Povelones, Igor Sharakhov, Ronald van Rij, Rob Waterhouse

Previous Scientific Working Group

VEuPathDB wishes to acknowledge previous scientific working group members. They provided regular feedback oversight and guidance.

Lyric Bartholomay	Michael Gottlieb	Malcolm McConville	John Taylor
Matt Berriman	Keith Gull	Nicola Mulder	Jake Tu

Bill Black	Matthew Hahn	Ull Munderloh	Brett Tyler
John Boothroyd	Adrian Hehl	Daniel Neafsey	Kenneth Vernick
Greg Buck	Steve Higgs	Kenneth Olson	Sarah Volkman
Geraldine Butler	Catherine Hill	Bill Petri	Jonathan Wastling
Angela Cruz	Marcelo Jacobs-Lorena	Barry Pittendrigh	Scott Weaver
George Dimopoulos	Anthony (Tony) James	Jeffrey Powell	Louis Weiss
Martin Donnelly	Pedro Lagerblad de Oliveira	Hillary Ranson	Dyann Wirth
Patrick Duffy	Greg Lanzaro	Alexander Raikhel	Jennifer Wortman
Pascale Gaudet	Daniel Masiga	Lincoln Stein	Guiyun Yan

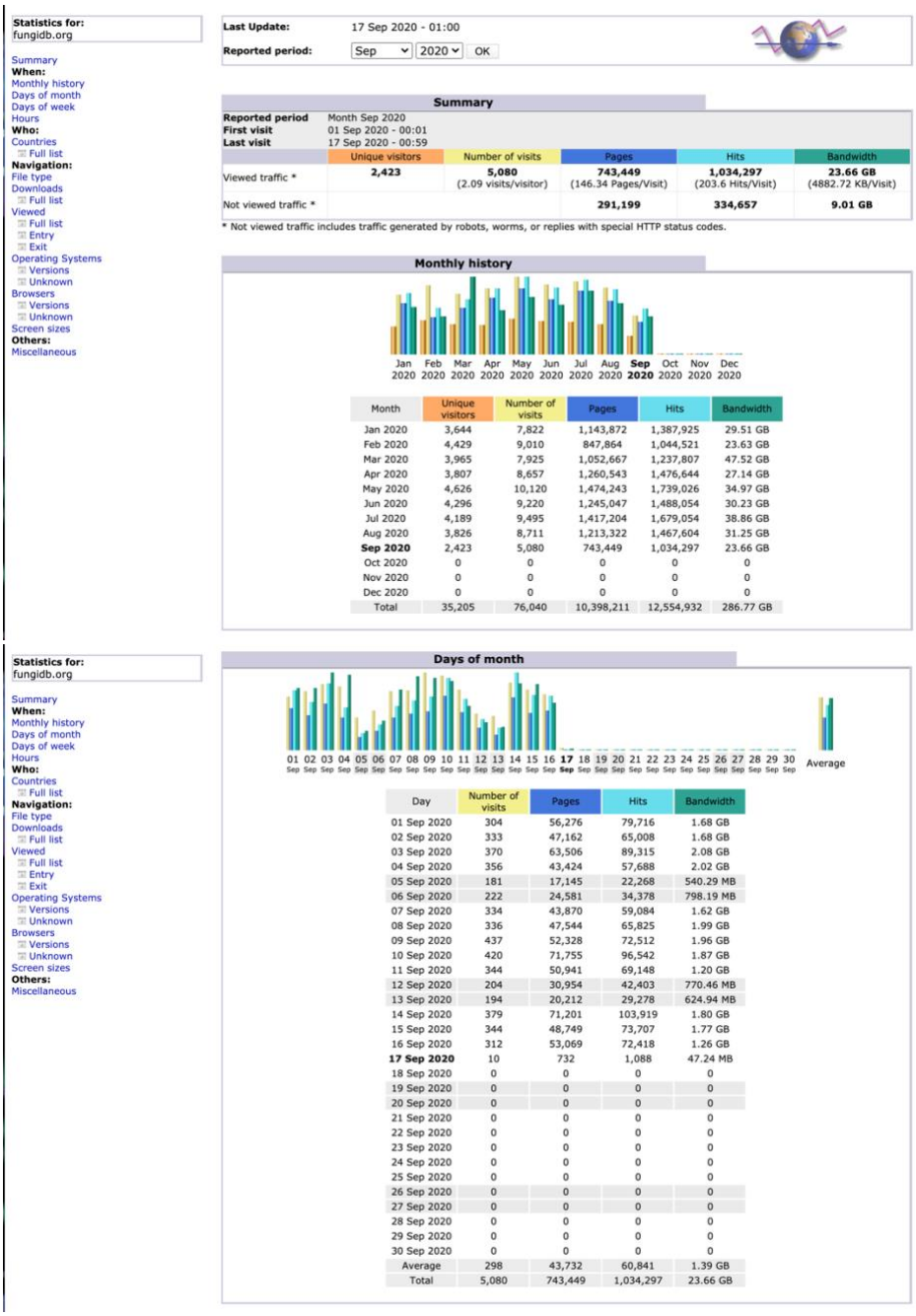
Website usage statistics

We collect project-specific website usage statistics using the program awstats. The links to all statistics (they are continually updated) are listed below. A sample website usage statistics report, as prepared by awstats is shown for an example of the types of data that are available. Users can visit the site and select custom reporting periods.

Website usage links:

- https://amoebadb.org/awstats/awstats.pl
- https://cryptodb.org/awstats/awstats.pl
- https://fungidb.org/awstats/awstats.pl
- https://giardiadb.org/awstats/awstats.pl
- https://hostdb.org/awstats/awstats.pl
- https://microsporidiadb.org/awstats/awstats.pl
- https://piroplasmadb.org/awstats/awstats.pl
- https://plasmodb.org/awstats/awstats.pl
- https://toxodb.org/awstats/awstats.pl
- https://trichdb.org/awstats/awstats.pl
- https://tritrypdb.org/awstats/awstats.pl
- https://orthomcl.org/proxystats/awstats.pl?config=orthomcl.org (this link will be updated when OrthoMCLDB moves into the new UI)
- https://veupathdb.org/awstats/awstats.pl
- https://vectorbase.org/awstats/awstats.pl

Sample awstats report from FungiDB.org



Statistics for: fungidb.org

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Navigation:
File type
Downloads
Full list
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Screen sizes
Others:
Miscellaneous

Last Update: 17 Sep 2020 - 01:00
Reported period: Sep 2020 OK

Summary

Reported period	Month Sep 2020				
First visit	01 Sep 2020 - 00:01				
Last visit	17 Sep 2020 - 00:59				
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	**2,423**	**5,080** (2.09 visits/visitor)	**743,449** (146.34 Pages/Visit)	**1,034,297** (203.6 Hits/Visit)	**23.66 GB** (4882.72 KB/Visit)
Not viewed traffic *			**291,199**	**334,657**	**9.01 GB**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Monthly history

Jan 2020 Feb 2020 Mar 2020 Apr 2020 May 2020 Jun 2020 Jul 2020 Aug 2020 **Sep 2020** Oct 2020 Nov 2020 Dec 2020

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2020	3,644	7,822	1,143,872	1,387,925	29.51 GB
Feb 2020	4,429	9,010	847,864	1,044,521	23.63 GB
Mar 2020	3,965	7,925	1,052,667	1,237,807	47.52 GB
Apr 2020	3,807	8,657	1,260,543	1,476,644	27.14 GB
May 2020	4,626	10,120	1,474,243	1,739,026	34.97 GB
Jun 2020	4,296	9,220	1,245,047	1,488,054	30.23 GB
Jul 2020	4,189	9,495	1,417,204	1,679,054	38.86 GB
Aug 2020	3,826	8,711	1,213,322	1,467,604	31.25 GB
Sep 2020	2,423	5,080	743,449	1,034,297	23.66 GB
Oct 2020	0	0	0	0	0
Nov 2020	0	0	0	0	0
Dec 2020	0	0	0	0	0
Total	35,205	76,040	10,398,211	12,554,932	286.77 GB

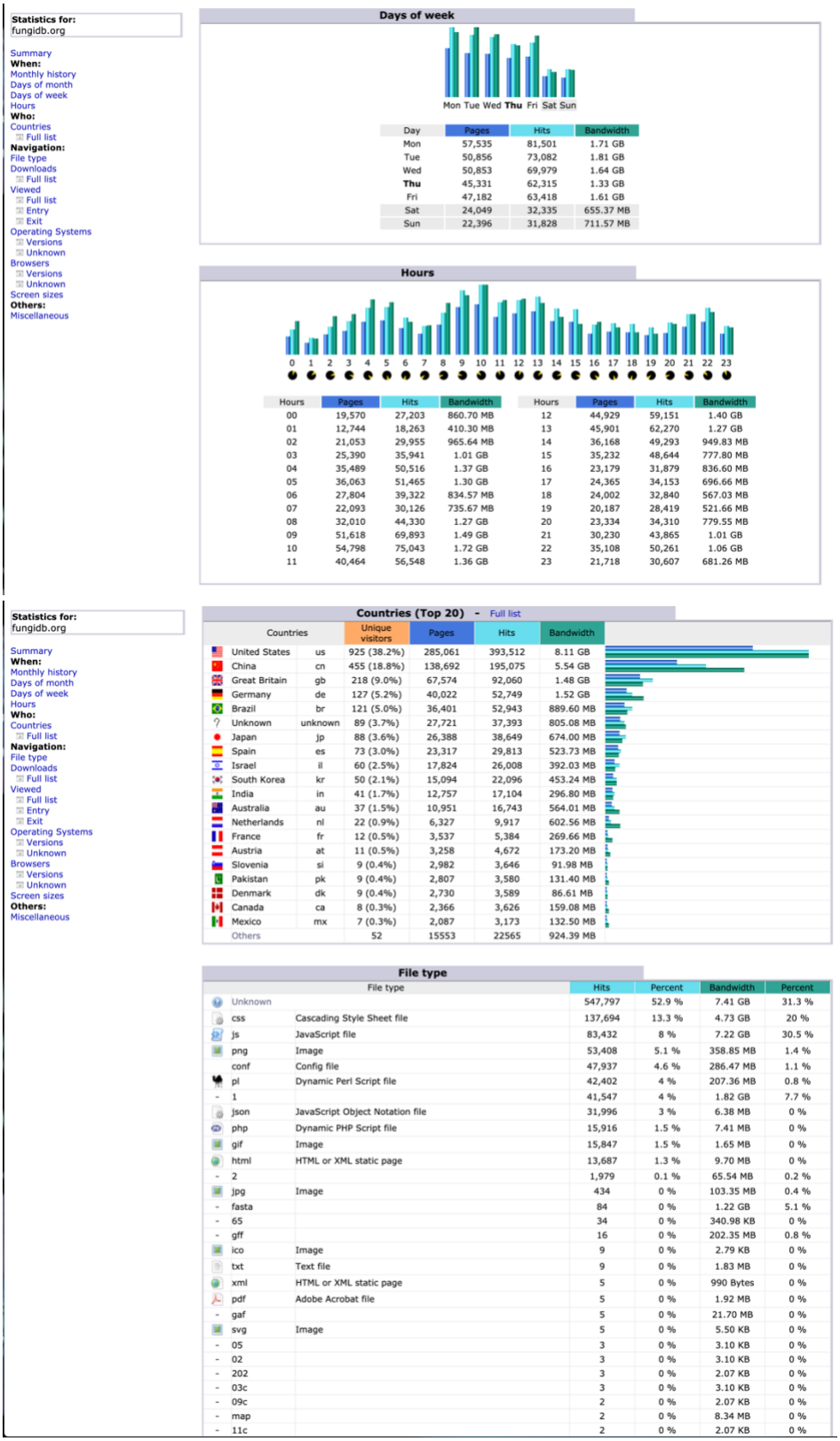
Statistics for: fungidb.org

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Navigation:
File type
Downloads
Full list
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Screen sizes
Others:
Miscellaneous

Days of month

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 **17** 18 19 20 21 22 23 24 25 26 27 28 29 30 Sep Average

Day	Number of visits	Pages	Hits	Bandwidth
01 Sep 2020	304	56,276	79,716	1.68 GB
02 Sep 2020	333	47,162	65,008	1.68 GB
03 Sep 2020	370	63,506	89,315	2.08 GB
04 Sep 2020	356	43,424	57,688	2.02 GB
05 Sep 2020	181	17,145	22,268	540.29 MB
06 Sep 2020	222	24,581	34,378	798.19 MB
07 Sep 2020	334	43,870	59,084	1.62 GB
08 Sep 2020	336	47,544	65,825	1.99 GB
09 Sep 2020	437	52,328	72,512	1.96 GB
10 Sep 2020	420	71,755	96,542	1.87 GB
11 Sep 2020	344	50,941	69,148	1.20 GB
12 Sep 2020	204	30,954	42,403	770.46 MB
13 Sep 2020	194	20,212	29,278	624.94 MB
14 Sep 2020	379	71,201	103,919	1.80 GB
15 Sep 2020	344	48,749	73,707	1.77 GB
16 Sep 2020	312	53,069	72,418	1.26 GB
17 Sep 2020	10	732	1,088	47.24 MB
18 Sep 2020	0	0	0	0
19 Sep 2020	0	0	0	0
20 Sep 2020	0	0	0	0
21 Sep 2020	0	0	0	0
22 Sep 2020	0	0	0	0
23 Sep 2020	0	0	0	0
24 Sep 2020	0	0	0	0
25 Sep 2020	0	0	0	0
26 Sep 2020	0	0	0	0
27 Sep 2020	0	0	0	0
28 Sep 2020	0	0	0	0
29 Sep 2020	0	0	0	0
30 Sep 2020	0	0	0	0
Average	298	43,732	60,841	1.39 GB
Total	5,080	743,449	1,034,297	23.66 GB



Statistics for:
fungidb.org

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Navigation:
File type
Downloads
Full list
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Screen sizes
Others:
Miscellaneous

Days of week

Mon Tue Wed **Thu** Fri Sat Sun

Day	Pages	Hits	Bandwidth
Mon	57,535	81,501	1.71 GB
Tue	50,856	73,082	1.81 GB
Wed	50,853	69,979	1.64 GB
Thu	45,331	62,315	1.33 GB
Fri	47,182	63,418	1.61 GB
Sat	24,049	32,335	655.37 MB
Sun	22,396	31,828	711.57 MB

Hours

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Hours	Pages	Hits	Bandwidth	Hours	Pages	Hits	Bandwidth
00	19,570	27,203	860.70 MB	12	44,929	59,151	1.40 GB
01	12,744	18,263	410.30 MB	13	45,901	62,270	1.27 GB
02	21,053	29,955	965.64 MB	14	36,168	49,293	949.83 MB
03	25,390	35,941	1.01 GB	15	35,232	48,644	777.80 MB
04	35,489	50,516	1.37 GB	16	23,179	31,879	836.60 MB
05	36,063	51,465	1.30 GB	17	24,365	34,153	696.66 MB
06	27,804	39,322	834.57 MB	18	24,002	32,840	567.03 MB
07	22,093	30,126	735.67 MB	19	20,187	28,419	521.66 MB
08	32,010	44,330	1.27 GB	20	23,334	34,310	779.55 MB
09	51,618	69,893	1.49 GB	21	30,230	43,865	1.01 GB
10	54,798	75,043	1.72 GB	22	35,108	50,261	1.06 GB
11	40,464	56,548	1.36 GB	23	21,718	30,607	681.26 MB

Statistics for:
fungidb.org

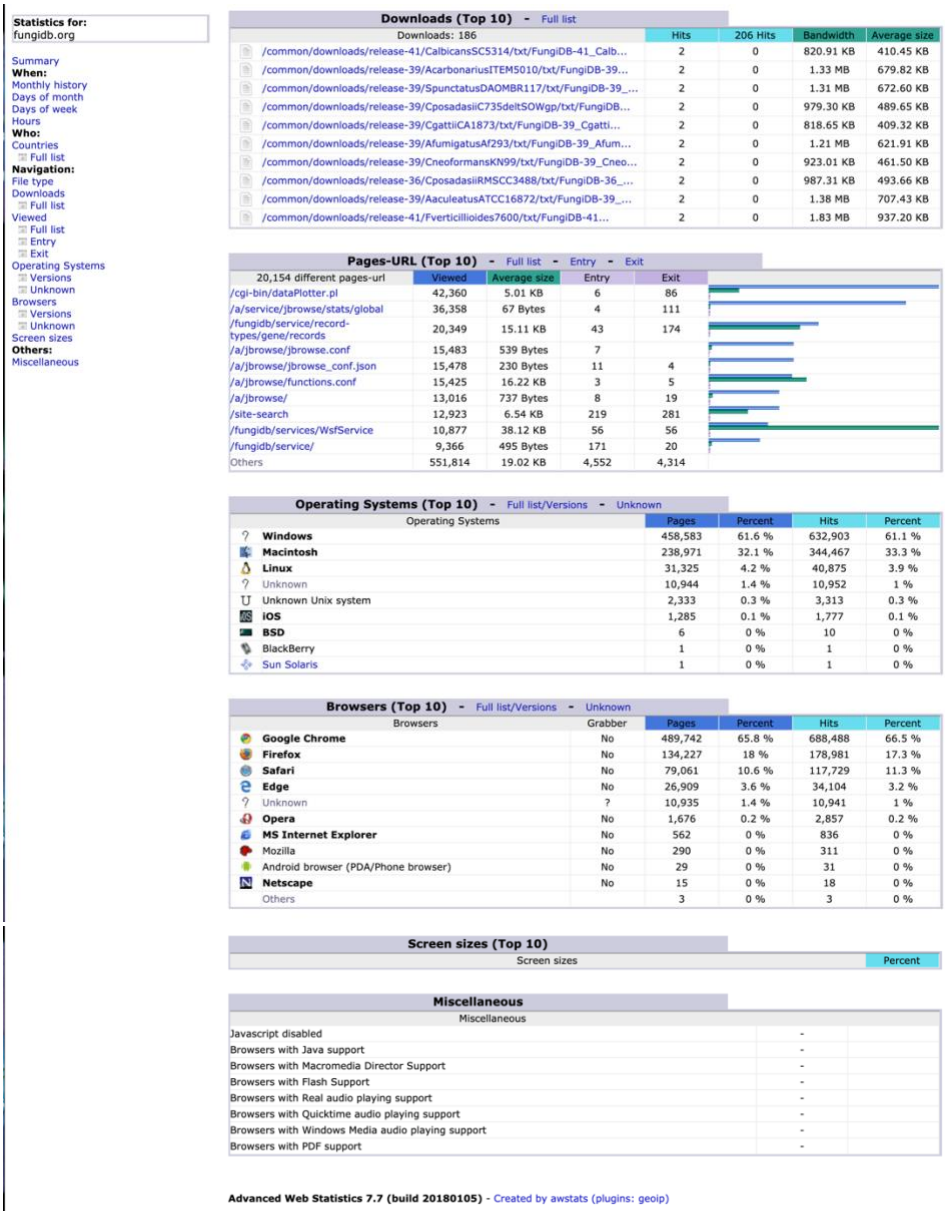
Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Navigation:
File type
Downloads
Full list
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Screen sizes
Others:
Miscellaneous

Countries (Top 20) - Full list

Countries		Unique visitors	Pages	Hits	Bandwidth
United States	us	925 (38.2%)	285,061	393,512	8.11 GB
China	cn	455 (18.8%)	138,692	195,075	5.54 GB
Great Britain	gb	218 (9.0%)	67,574	92,060	1.48 GB
Germany	de	127 (5.2%)	40,022	52,749	1.52 GB
Brazil	br	121 (5.0%)	36,401	52,943	889.60 MB
Unknown	unknown	89 (3.7%)	27,721	37,393	805.08 MB
Japan	jp	88 (3.6%)	26,388	38,649	674.00 MB
Spain	es	73 (3.0%)	23,317	29,813	523.73 MB
Israel	il	60 (2.5%)	17,824	26,008	392.03 MB
South Korea	kr	50 (2.1%)	15,094	22,096	453.24 MB
India	in	41 (1.7%)	12,757	17,104	296.80 MB
Australia	au	37 (1.5%)	10,951	16,743	564.01 MB
Netherlands	nl	22 (0.9%)	6,327	9,917	602.56 MB
France	fr	12 (0.5%)	3,537	5,384	269.66 MB
Austria	at	11 (0.5%)	3,258	4,672	173.20 MB
Slovenia	si	9 (0.4%)	2,982	3,646	91.98 MB
Pakistan	pk	9 (0.4%)	2,807	3,580	131.40 MB
Denmark	dk	9 (0.4%)	2,730	3,589	86.61 MB
Canada	ca	8 (0.3%)	2,366	3,626	159.08 MB
Mexico	mx	7 (0.3%)	2,087	3,173	132.50 MB
Others		52	15553	22565	924.39 MB

File type

File type		Hits	Percent	Bandwidth	Percent
Unknown		547,797	52.9 %	7.41 GB	31.3 %
css	Cascading Style Sheet file	137,694	13.3 %	4.73 GB	20 %
js	JavaScript file	83,432	8 %	7.22 GB	30.5 %
png	Image	53,408	5.1 %	358.85 MB	1.4 %
conf	Config file	47,937	4.6 %	286.47 MB	1.1 %
pl	Dynamic Perl Script file	42,402	4 %	207.36 MB	0.8 %
1		41,547	4 %	1.82 GB	7.7 %
json	JavaScript Object Notation file	31,996	3 %	6.38 MB	0 %
php	Dynamic PHP Script file	15,916	1.5 %	7.41 MB	0 %
gif	Image	15,847	1.5 %	1.65 MB	0 %
html	HTML or XML static page	13,687	1.3 %	9.70 MB	0 %
2		1,979	0.1 %	65.54 MB	0.2 %
jpg	Image	434	0 %	103.35 MB	0.4 %
fasta		84	0 %	1.22 GB	5.1 %
65		34	0 %	340.98 KB	0 %
gff		16	0 %	202.35 MB	0.8 %
ico	Image	9	0 %	2.79 KB	0 %
txt	Text file	9	0 %	1.83 MB	0 %
xml	HTML or XML static page	5	0 %	990 Bytes	0 %
pdf	Adobe Acrobat file	5	0 %	1.92 MB	0 %
gaf		5	0 %	21.70 MB	0 %
svg	Image	5	0 %	5.50 KB	0 %
05		3	0 %	3.10 KB	0 %
02		3	0 %	3.10 KB	0 %
202		3	0 %	2.07 KB	0 %
03c		3	0 %	3.10 KB	0 %
09c		2	0 %	2.07 KB	0 %
map		2	0 %	8.34 MB	0 %
11c		2	0 %	2.07 KB	0 %



Statistics for: fungidb.org

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Navigation:
File type
Downloads
Full list
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Screen sizes
Others:
Miscellaneous

Downloads (Top 10) - Full list

Downloads: 186	Hits	206 Hits	Bandwidth	Average size
/common/downloads/release-41/CalbicansSC5314/txt/FungiDB-41_Calb...	2	0	820.91 KB	410.45 KB
/common/downloads/release-39/AcarbonariusITEM5010/txt/FungiDB-39...	2	0	1.33 MB	679.82 KB
/common/downloads/release-39/SpunctatusDAOMBR117/txt/FungiDB-39_...	2	0	1.31 MB	672.60 KB
/common/downloads/release-39/CposadasiiC735deltSOWgp/txt/FungiDB...	2	0	979.30 KB	489.65 KB
/common/downloads/release-39/CgattiiCA1873/txt/FungiDB-39_Cgatti...	2	0	818.65 KB	409.32 KB
/common/downloads/release-39/AfumigatusAf293/txt/FungiDB-39_Afum...	2	0	1.21 MB	621.91 KB
/common/downloads/release-39/CneoformansKN99/txt/FungiDB-39_Cneo...	2	0	923.01 KB	461.50 KB
/common/downloads/release-36/CposadasiiRMSCC3488/txt/FungiDB-36_...	2	0	987.31 KB	493.66 KB
/common/downloads/release-39/AaculeatusATCC16872/txt/FungiDB-39_...	2	0	1.38 MB	707.43 KB
/common/downloads/release-41/Fverticillioides7600/txt/FungiDB-41...	2	0	1.83 MB	937.20 KB

Pages-URL (Top 10) - Full list - Entry - Exit

20,154 different pages-url	Viewed	Average size	Entry	Exit
/cgi-bin/dataPlotter.pl	42,360	5.01 KB	6	86
/a/service/jbrowse/stats/global	36,358	67 Bytes	4	111
/fungidb/service/record-types/gene/records	20,349	15.11 KB	43	174
/a/jbrowse/jbrowse.conf	15,483	539 Bytes	7	
/a/jbrowse/jbrowse_conf.json	15,478	230 Bytes	11	4
/a/jbrowse/functions.conf	15,425	16.22 KB	3	5
/a/jbrowse/	13,016	737 Bytes	8	19
/site-search	12,923	6.54 KB	219	281
/fungidb/services/WsfService	10,877	38.12 KB	56	56
/fungidb/service/	9,366	495 Bytes	171	20
Others	551,814	19.02 KB	4,552	4,314

Operating Systems (Top 10) - Full list/Versions - Unknown

Operating Systems	Pages	Percent	Hits	Percent
Windows	458,583	61.6 %	632,903	61.1 %
Macintosh	238,971	32.1 %	344,467	33.3 %
Linux	31,325	4.2 %	40,875	3.9 %
Unknown	10,944	1.4 %	10,952	1 %
Unknown Unix system	2,333	0.3 %	3,313	0.3 %
iOS	1,285	0.1 %	1,777	0.1 %
BSD	6	0 %	10	0 %
BlackBerry	1	0 %	1	0 %
Sun Solaris	1	0 %	1	0 %

Browsers (Top 10) - Full list/Versions - Unknown

Browsers	Grabber	Pages	Percent	Hits	Percent
Google Chrome	No	489,742	65.8 %	688,488	66.5 %
Firefox	No	134,227	18 %	178,981	17.3 %
Safari	No	79,061	10.6 %	117,729	11.3 %
Edge	No	26,909	3.6 %	34,104	3.2 %
Unknown	?	10,935	1.4 %	10,941	1 %
Opera	No	1,676	0.2 %	2,857	0.2 %
MS Internet Explorer	No	562	0 %	836	0 %
Mozilla	No	290	0 %	311	0 %
Android browser (PDA/Phone browser)	No	29	0 %	31	0 %
Netscape	No	15	0 %	18	0 %
Others		3	0 %	3	0 %

Screen sizes (Top 10)

Screen sizes	Percent

Miscellaneous

Miscellaneous	
Javascript disabled	-
Browsers with Java support	-
Browsers with Macromedia Director Support	-
Browsers with Flash Support	-
Browsers with Real audio playing support	-
Browsers with Quicktime audio playing support	-
Browsers with Windows Media audio playing support	-
Browsers with PDF support	-

Advanced Web Statistics 7.7 (build 20180105) - Created by awstats (plugins: geoip)

VEuPathDB Glossary

Please also check the NCBI glossary

- **3-Frame translation (forward)**
  Translation of a nucleotide sequence in all three possible reading frames in one direction, usually "on the top strand" of DNA.
  - **3-Frame translation (reverse)**
    Translation of a nucleotide sequence in all three possible reading frames in the reverse direction, usually "on the bottom strand" of DNA.
  - **AA sequence**

Amino acid sequence.

- **Affymetrix genotyped SNP probes**
  Probes on Affymetrix SNP (single nucleotide polymorphism) arrays, which are used for SNP genotyping. See Affymetrix microarray technology and www.affymetrix.com
- **Affymetrix microarray technology**
  Microarray manufacturing technology developed by Affymetrix. Combines semiconductor fabrication techniques, solid phase chemistry, combinatorial chemistry, molecular biology, and robotics to generate a photolithographic manufacturing process in which oligonucleotides are synthesized directly on a chip. See www.affymetrix.com
- **Affymetrix probes**
  Probe on an Affymetrix microarray designed to determine whether or not the complementary sequence of RNA or DNA is present in a sample. Generally 25 nucleotides in length (25-mers), their short length provides higher specificity than longer probes. See Affymetrix microarray technology and www.affymetrix.com
- **Amitochondriate**
  Eukaryotic organism that lacks a mitochondrion. Examples include Giardia and other parasites such as Trachipleistophora and Entamoeba. However, most of these organisms contain what appear to be mitochondrial remnants as well as mitochondrial genes in their nuclear genomes.
- **Annotation**
  Identified feature within a sequence, such as a known or predicted gene, domain, motif, post-translational modification, etc.
- **Annotation density**
  Level to which a nucleotide or protein sequence has been annotated. See Annotation.
- **ApiCyc**
  Database/utility on EuPathDB used for searching and visualizing metabolic pathway information for organisms in EuPathDB; derivative database generated by analyzing various genomes (for example from Plasmodium, Cryptosporidium, and Toxoplasma) with SRI International's pathway tools.
- **ApiDots alignments**
  Consensus sequences found in the ApiDots database and generated by clustering and assembling Apicomplexan mRNA and EST sequences. These consensus sequences were subjected to database searches against protein and protein domain sequences.
- **Apicoplast**
  Nonphotosynthetic plastid found in almost all protozoan parasites belonging to the phylum Apicomplexa that have been examined. The apicoplast is surrounded by four membranes, giving rise to the theory that its presence in the Apicomplexa is the result of a secondary endosymbiosis (acquired by the engulfment of an ancestral alga and retention of the algal plastid). Similar to other endosymbiotic organelles (mitochondria, chloroplasts), the apicoplast contains its own genome as well as proteins that are encoded in the nucleus and post-translationally imported. The apicoplast is a vital organelle to the parasite's long-term survival.
- **Attribute**

Inherent characteristic or feature; in a database, a data item related to a database object. For example, attributes of genes can include features such as introns and untranslated regions (UTRs).

- **BLAST**

Basic local alignment search tool, a sequence similarity search tool used to quickly find local alignments between a query sequence and sequences in nucleotide or protein databases. Different versions of this search tool are available to match the types of query sequence and database used. See blastn, blastp, blastx, tblastn, and tblastx.

- **Boolean**

System of logical thought developed by George Boole (1815-1864). In Boolean searching, an "and" operator between two words or values (for example, "apple AND orange") generates a search for items in a database containing both of the words or values. Similarly, an "or" operator between two words or values (for example, "apple OR orange") generates a search for items containing either word.

- **CDS**

Coding sequence. Region of nucleotides that corresponds to the sequence of amino acids in a predicted protein and that includes start and stop codons. Unexpressed sequences (for example, the 5'-UTR, the 3'-UTR, and introns) are not included within a CDS. The CDS usually does not correspond to the actual mRNA sequence.

- **Centromere**

Region of the chromosome or chromosomal structure essential for division and retention of the chromosome within the cell; point of a chromosome where the spindle fibers attach to pull the chromosome apart during cell division.

- **Chromosome**

Macromolecule of DNA constituting the physical organization of DNA in a cell.

- **Coil**

Three-dimensional spiral structure in protein macromolecules.

- **Contig**

Contiguous genomic sequence assembled from overlapping primary sequences representing overlapping regions of a particular chromosome.

- **CryptoCyc**

Database/utility built by analyzing the Cryptosporidium genome with SRI International's pathway tools; used for searching and visualizing Cryptosporidium metabolic pathway information.

- **Curated annotation**

Annotation made under the supervision of a curator as opposed to a purely computational prediction. Curated predictions often contain combinations of different types of evidence to support the annotation.

- **DNA/GC content**

Content of guanine (G) and cytosine (C) in a fragment of DNA or a genome. Because GC pairs are more thermostable compared to the AT pairs, it was commonly believed that GC content played a vital part in adaptation to high temperatures, a hypothesis that has been refuted. In the genome browser, the DNA/GC content track displays a GC content graph of the reference sequence at low magnifications and the DNA sequence itself at higher magnifications.

- **Dalton**

Unit of mass abbreviated Da and used to express atomic and molecular masses.

- **Deprecated gene**

Genes with little or no evidence (similarities / expression) that overlapped (or were subsumed by) larger genes for which there was evidence such as protein similarities, expression evidence from EST alignments, SAGE or proteomics data were marked as deprecated. These genes will likely be removed in a subsequent release (and in GenBank) unless additional evidence is provided indicating they should be moved into the real gene category.

- **EC numbers**

Enzyme Commission numbers. EC numbers constitute the numerical classification scheme for enzymes based on the chemical reactions they catalyze. EC numbers do not refer to the enzymes, but to the reactions they catalyze.

- **EST**

Expressed sequence tag. Short (typically 100-500 base pairs) partial cDNA produced by single-shot sequencing of a cloned mRNA (cDNA) and often used to identify gene transcripts.

- **EST alignments**

Alignments of expressed sequence tags (ESTs) with a corresponding genomic region. For example, in ToxoDB you can visualize EST alignments by clicking on the "View this sequence in the genome browser" link and turning on the EST Alignments track. Useful for identifying intron boundaries.

- **EST clusters**

Groups of homologous, overlapping EST sequences created to reduce redundancy of the EST database.

- **Expression level**

Level at which an mRNA or protein is present in a sample. Value can be absolute or relative to other mRNA or protein species in the sample.

- **Expression profile**

Pattern of expression of one or more genes or proteins over time or over a set of experimental conditions (for example, during development or treatment, or as a result of a genetic mutation such as a knockout).

- **Expression profile correlation**

Method for correlation of gene expression profiles with gene ontology (GO) annotations developed for the purpose of identifying groups of genes, pathways, and processes reacting in concert to experimental perturbations.

- **Expression timing**

Timing of gene expression during a developmental, metabolic, regulatory, or other biological process or response.

- **GBrowse**

Interactive genome browser developed by the Generic Model Organism Database (GMOD) project (www.gmod.org) that can be customized to show selected chromosomal features as well as display user-provided annotations.

- **GBrowse track**

In the GBrowse viewer, a line of data that corresponds to a particular type of genomic information or feature and that is distinguished by a particular shape or color.

- **Gametocyte**

Eukaryotic germ cell that divides by mitosis to generate other gametocytes or by meiosis to generate gametes. Male gametocytes are called spermatocytes, and female gametocytes are called oocytes. Term often used to describe gametocytes of Plasmodium.

- **GenBank protein record**

Protein sequence file in the GenBank database generally derived by translation of a related nucleotide record.

- **GenPept protein**

Protein record from the GenPept database at the NCBI GenBank, which contains inferred translations of protein-coding sequences.

- **Gene**

Fundamental physical and functional unit of heredity. Ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product, such as a protein or RNA molecule. A gene may have a number of parts, including the promoter region, untranslated regions (5' and 3' UTRs), introns, and exons.

- **GeneDB**

Project developed by the Sanger Institute Pathogen Sequencing Unit (PSU) and aimed at developing and maintaining curated database resources for all projects handled by the PSU. The database is accessible at www.genedb.org

- **Genetic markers**

Known DNA sequences that can be identified by a simple assay. Generally genetic variations caused by mutation or alterations in loci that can be observed, examples include restriction length polymorphisms (RFLPs), short tandem repeats (STRs), variable number tandem repeats (VNTRs), short DNA sequences surrounding single base-pair changes (single nucleotide polymorphisms, or SNPs), or longer microsatellite sequences.

- **Genomic context**

Location of a gene in the genome, which can influence the expression of the gene and functional interactions of the gene expression products. In this database, genes are depicted on individual gene pages with their surrounding genomic region and annotations.

- **Genome Sequence**

The inherited nucleic acid component of the genome. In eukaryotes, organized into linear chromosomes that are capped with telomeres and located in the nucleus.

- **Genotyped SNPs**

Single nucleotide polymorphisms (SNPs) identified during genotyping of individual organism strains. See SNP genotyping.

- **Genotyping**

Process of determining the genotype of an individual with a biological assay using PCR, DNA sequencing, or hybridization to DNA microarrays or beads. Provides a measurement of the genetic variation between members of a species.

- **GLEAN gene**
  Predicted gene sequence generated by GLEAN, an algorithm that integrates different sources of gene structure evidence (for example, gene model predictions, EST and protein sequence alignments to the genome, and SAGE or peptide tags) to produce a consensus gene prediction in the absence of known genes.
- **GO**
  Gene Ontology project. Collaborative project that has developed three structured, controlled vocabularies (ontologies) to describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. The use of a consistent vocabulary allows genes from different species to be compared based on their GO annotations.
- **GO component**
  Gene Ontology term used to describe a cellular component, or the location where a gene product may act, rather than physical features of proteins or RNAs. For example, membrane (GO:0016020), extrinsic to membrane (GO:0019898), and integral to membrane (GO:0016021).
- **GO function**
  Gene Ontology term used to describe the molecular function of a gene product, the jobs that it performs, or the "abilities" that it has (for example, transporting compounds, binding to things, holding things together, and changing one thing into another). This is different from the biological processes the gene product is involved in, which involve more than one activity.
- **GO process**
  Gene Ontology term used to describe a biological process, a recognized series of events, or molecular functions associated with a gene product. A biological process is not equivalent to a pathway, though some GO terms do describe pathways.
- **GO term**
  Gene Ontology term. The building blocks of the Gene Ontology, each term is assigned to one of the three ontologies: molecular function, cellular component, or biological process. Each GO term consists of a unique alphanumerical identifier, a common name, synonyms (if applicable), and a definition. When a term has multiple meanings depending on species, the GO uses a "sensu" tag to differentiate among them. For example, the enzyme fumarase has the GO term GO:0004333, fumarate hydratase activity (fumarase activity), catalysis of the reaction: (S)-malate = fumarate + H2O.
- **GPI anchor**
  C-terminal post-translational modification of many eukaryotic proteins. The two fatty acids within the glycophosphatidylinositol (GPI) moiety anchor the protein to the outer leaflet of the plasma membrane. GPI-anchored proteins are believed to be involved in signal transduction and immune responses, as well as the pathobiology of many parasites.

- **HMM**
  Hidden Markov model. Statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to

determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications.

- **Homolog**
Related by evolutionary descent, either between species (ortholog) or within a species (paralog).
- **Hydropathy**
Hydropathicity. Degree to which a peptide or protein is likely to be soluble in water. Protein hydropathy plots can be useful in predicting transmembrane domains, potential antigenic sites, and regions that are likely to be exposed on the protein's surface.
- **Intergenic region**
Stretch of DNA located between genes that may act to regulate gene expression.
- **Intersect**
Similar to the Boolean operator "AND", an action used in the query history to find items that are common to two query result sets. For example, to search for items that appear in both result set X and result Y, type, "X INTERSECT Y".
- **Join**
Similar to the Boolean operator "UNION", an action used in the query history to combine query sets. For example, to combine result sets X and Y, type, "X JOIN Y".
- **JBrowse**
Interactive genome browser (jbrowse.org) developed by the Generic Model Organism Database (GMOD) project (www.gmod.org) that can be customized to show selected chromosomal features as well as display user-provided annotations.
- **KEGG map**
Metabolic or regulatory interaction pathway generated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) or by the use of their tools (www.genome.jp/kegg).
- **Locus**
Position on a chromosome of a gene, feature (such as a telomere), or other chromosomal marker; also, the DNA at that position. Use of this term is sometimes restricted to mean expressed DNA regions.
- **Low complexity**
Pertaining to sequence regions that have an unusually repetitive nature (for example, a protein sequence of low complexity might look like PPTDPPPPKKDGGPPL, and a low-complexity nucleotide sequence might be AAATAAAAAAAAATAAAAAAATTA). Low-complexity regions can create problems in sequence similarity searching by causing artifactual hits. For this reason, filters are often used to remove low-complexity sequences. Low-complexity regions also contribute to antigenic variation in apicomplexan parasites.
- **Mass spec data**
Mass spectrometry data. Mass spectrometry is an analytical technique used to measure the mass-to-charge ratios of small molecules in several applications, including identification of proteins or peptides. In our databases, the "Identify Genes by Mass Spec Evidence" query is used to identify genes that have evidence of protein expression based on mass spec data.
- **Metabolic pathways**

Series of chemical reactions occurring within a cell and often catalyzed by enzymes. In a pathway, a molecule is often changed or modified into another product, which can be stored by the cell, used as a metabolic product, or used to initiate another pathway.

- **Microarray**

Microscopic array of biological molecules (for example, DNA or protein) used to determine the presence and/or amount (referred to as quantitation) of other biomolecules (other proteins, transcripts, etc.) in biological samples.

- **Microsatellite**

Polymorphic locus in nuclear and organellar DNA that consists of repeating units of 1-4 base pairs in length. Mostly neutral and codominant, microsatellites are used as molecular markers and to study gene dosage (looking for duplications or deletions of a particular genetic region). Also known as simple sequence repeats (SSRs).

- **Microsatellite map**

Map of microsatellite locations and linkages on a genome.

- **Mitochondrion**

Organelle responsible for respiration in a eukaryotic cell. Proteins required for mitochondrial function are encoded both in the nucleus and within the smaller mitochondrial genome.

- **Motif search**

Tool used to identify and locate sequence patterns (motifs) in protein and nucleic acid sequences. In our databases, this flexible search can be based on the general characteristics of the pattern and not solely on specific sequences (for example, Cys-[9-11 amino acids]-Cys or Leu-Leu-[basic residue]-Val). This allows the user to query using previously undescribed motifs.

- **NCBI**

National Center for Biotechnology Information. Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

- **Nonredundant protein DB (NRDB)**

Peptide sequence database containing all nonidentical protein sequences from the same species extracted from GenPept, NCBI RefSeq, Swiss-Prot, and PRF databases. Used for BLAST protein database searches because its smaller size results in shorter search times and more meaningful statistics.

- **Nucmer**

NUCleotide MUMmer. Part of the MUMmer alignment package, an alignment tool used for the rapid alignment of very large DNA and amino acid sequences.

- **ORF**

Amino acid sequences computed by translating the six frames of raw genomic sequence using the standard genetic code. We save the translated sequences having at least 50 amino acids. The sequences are not annotated nor human reviewed. ORF's do not necessarily begin with methionine residues, but they all terminate with a stop codons.

- **Oligo**

Oligonucleotide. Short sequence of DNA or RNA, typically of 20-70 nucleotides.

- **Oligonucleotide microarray**

Collection of microscopic oligonucleotide spots arrayed on a solid surface by covalent attachment to chemically suitable matrices. Used for expression profiling, the monitoring of the expression levels of hundreds or thousands of genes simultaneously. Probes for oligonucleotide microarrays are designed to match parts of known or predicted mRNAs.

- **Open Reading Frame**

See ORF

- **Organellar genome sequence**

The genome sequence contained in an organelle *i.e*. the mitochondria and/or apicoplast organelles. These sequences may be circular or linear and exist in addition to the much larger nuclear genome sequence.

- **OrthoMCL**

Genome-scale algorithm for grouping orthologous protein sequences. It provides genes shared by two or more species/genomes and also genes representing species-specific gene expansion families. Therefore, it serves as a utility for automated eukaryotic genome annotation and phylogenetic profiling. Available at orthomcl.org

- **Ortholog**

Same gene in different organisms; having evolved from the same ancestral locus.

- **Ortholog group**

Orthologous genes shared by group of organisms; the group of genes can also contain paralogs.

- **Orthology-based phylogenetic profile**

Tool used to find genes that are present or not present in a desired group of organisms on the tree of life (currently computed for 81 complete genomes). The user has control over the profile and over whether or not genes must be found in any particular group of organisms (for example, in Apicomplexa but not in mammals). Taxa can also be marked as indifferent (for example, it does not matter if the gene is also found in plants). Ortholog and paralog relationships are determined using the OrthoMCL algorithm.

- **PATS**

Neural network analysis tool that identifies amino acid sequences within a query sequence that are potentially targeted to the apicoplast matrix of Plasmodium falciparum.

- **PDB 3D structure**

Three-dimensional macromolecular structure in the Protein Data Bank (PDB) (www.pdb.org) obtained by one of three methods: X-ray crystallography (over 80%), solution nuclear magnetic resonance (NMR) (about 16%), or theoretical modeling (2%). A few structures were determined by other methods.

- **PROSITE motif**

Protein sequence pattern or profile derived from multiple alignments of homologous sequences and stored in the PROSITE database (prosite.expasy.org), an annotated

collection of motif descriptors dedicated to the identification of protein families and domains.

- **Paralog**

Related by gene duplication within a genome; originated by duplication and then diverged from the parent copy by mutation and selection or drift.

- **Pearson correlation**

Pearson Product Moment Correlation, the most common measure of the correlation between two variables. Reflects the degree of linear relationship between two variables and ranges from +1 to -1, with a correlation of +1 indicating a perfect positive linear relationship between variables.

- **Peptide mass fingerprinting**

Analytical technique for protein identification wherein an unknown protein of interest is cleaved into peptides by a protease such as trypsin, and the peptides resulting from this cleavage are analyzed using a mass spectrometric method such as MALDI-TOF or ESI-TOF. The masses derived for the peptides are then compared to a database containing known protein sequences or even to the genome. Computer programs theoretically cut the protein sequences in the database into peptides with the same protease (for example trypsin), and calculate the absolute masses of the peptides from each protein. They then compare the masses of the peptides of the unknown protein to the theoretical peptide masses of each protein encoded in the genome. The results are statistically analyzed to find the best match.

- **Pfam domain**

Conserved protein region in the Pfam database (Pfam.org), a collection of multiple sequence alignments and hidden Markov models covering many common protein families. The alignments may represent evolutionarily conserved structures that may shed light on protein function. Profile hidden Markov models (profile HMMs) built from the Pfam alignments can be useful for associating a new protein to a known protein family, even if the homology is weak. Unlike standard pairwise alignment methods (for example, BLAST and FASTA), Pfam HMMs deal sensibly with multidomain proteins.

- **Phylogeny**

Historical relationships among lineages of organisms or their parts, including their genes.

- **PlasmoAP**

Algorithm/tool that predicts the likelihood that a protein sequence is targeted to the apicoplast. It provides the position of signal peptide cleavage sites in amino acid sequences if targeting is predicted.

- **PlasmoCyc**

Database/utility built by analyzing the genomes of the Plasmodium species in EuPathDB with SRI International's pathway tools; used for searching and visualizing Plasmodium metabolic pathway information.

- **ProDom**

Database of protein domain families generated from the global comparison of all available protein sequences (prodom.prabi.fr).

- **Promoter**

Regulatory region of DNA located upstream (towards the 5' region) of a gene and providing a control point for regulated gene transcription.

- **Protein-coding**

Capable of encoding a protein sequence; generally refers to a sequence of DNA.

- **Proteomics**

The large-scale study of proteins, particularly of the full set of proteins encoded by a genome.

- **Pseudogene**

Defunct relatives of known genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Although they may have some gene-like features (such as promoters, CpG islands, and splice sites), they are nonetheless considered nonfunctional due to their lack of protein-coding ability resulting from various genetic disablements (stop codons, frameshifts, or a lack of transcription) or their inability to function as an RNA (such as with rRNA pseudogenes).

- **PubCrawler**

Free service that scans daily updates to the NCBI Medline (PubMed) and GenBank databases and alerts users to any relevant updates. Available at pubcrawler.gen.tcd.ie

- **Query**

Sequence or term used in a database search. For example, the sequence submitted for a BLAST search is the query sequence.

- **RNA predictions**

Predictions of genes that encode nonprotein-encoding RNA's such as tRNA, snoRNA, rRNA, etc.

- **Reference genome sequence**

The community gold standard genome sequence. Usually the most complete assembly including an annotation. Reference genome sequences do change as superior assemblies and annotation become available.

- **RefSeq mRNA**

Nonredundant mRNA sequence in the RefSeq database. RefSeq mRNA sequences with an NM_XXXXXX accession are curated sequences and are, therefore, considered more reliable than those with XM_XXXXXX accessions (predicted mRNA sequences).

- **RefSeq noncoding RNA**

Nonredundant noncoding RNA (ncRNA) sequence in the RefSeq database. RefSeq ncRNA sequences with an NR_XXXXXX accession are curated sequences and are, therefore, considered more reliable than those with XR_XXXXXX accessions (predicted ncRNA sequences).

- **RefSeq protein**

Nonredundant protein sequence in the RefSeq database. RefSeq protein sequences with NP_XXXXXX accessions are curated sequences and are, therefore, considered more reliable than those with XP_XXXXXX accessions (predicted protein sequences).

- **RefSeq**

NCBI reference sequences. A curated nonredundant collection of sequences representing genomes, transcripts, and proteins as annotated by NCBI (available

at www.ncbi.nlm.nih.gov/refseq). The annotation in these records is often different from the original GenBank submission, which may not be updated every time new information is obtained.

- **Repeat regions**

Sequences present in many identical or highly similar copies in the genome.

- **SAGE tags**

Serial analysis of gene expression (SAGE) tags. Short (14-nucleotide) sequences found within mRNA, the relative abundance of which indicates the level of expression of the mRNA containing that tag.

- **SNP**

Single nucleotide polymorphism. Small genetic changes or variations that can occur within a DNA sequence, for example when a single nucleotide, such as an A replacing one of the other three nucleotide letters C, G, or T. Most SNPs are found outside of coding sequences, but SNPs found within a coding sequence are more likely to alter the biological function of a protein. SNPs may be synonymous (generating a conservative change not altering the amino acid sequence) or they can be nonsynonymous and change the amino acid that is encoded.

- **SNP density**

Amount or number of single nucleotide polymorphisms (SNPs) in a region of the genome.

- **SNP genotyping**

Identifying and mapping single nucleotide polymorphisms (SNPs) in an effort to determine the genotype members of a species. SNPs usually consist of two alleles (where the rare allele frequency is less than 1%), are evolutionarily conserved, and are the most common type of genetic variation. See Genotyping.

- **Scaffolds**

In genomic mapping, a series of contigs that are in the right order and orientation, but not necessarily connected in one continuous stretch of sequence.

- **Sequence similarity**

Degree of similarity between two or more protein or nucleotide sequences.

- **Signal peptide**

Short (3-60 amino acids) peptide sequence that directs the co-translational import of a protein to certain organelles or for secretion.

- **SignalP**

Program that predicts the presence and location of signal peptide cleavage sites in amino acid sequences from Gram-positive and -negative prokaryotes and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/nonsignal peptide prediction based on a combination of several artificial neural networks and hidden Markov models (HMMs).

- **Strand**

One half of the DNA helix; string or stretch of covalently linked nucleotides.

- **Subtract**

Similar to the Boolean operator "NOT", an action used in the query history to remove items from one result set that occur in another result set. For example, to remove items that exist in set Y from those in set X, type, "X SUBTRACT Y".

- **Syntenic**

Loci located on the same chromosome but not necessarily linked. For example, genes that are part of a syntenic group share a common chromosomal location. Also used to refer to conservation of gene order across species.

- **TIGR**
The Institute for Genomic Research, a nonprofit center dedicated to deciphering and analyzing genomes.
- **TM domains**
See transmembrane domain.
- **TWINSCAN gene models**
Gene models generated using TWINSCAN, a tool that integrates traditional probability models (such as those underlying GENSCAN and FGENESH) with information from alignments between two genomes. TWINSCAN is based on the idea that functional sequences show different patterns of evolutionary conservation than sequences under little selective pressure, such as the central regions of introns. TWINSCAN is designed for the analysis of high-throughput genomic sequences containing an unknown number of genes.
- **Telomere**
Nucleoprotein complexes that constitute the physical ends of linear eukaryotic chromosomes and that have important functions, primarily in the protection, replication, and stabilization of the chromosome ends. Telomeres often contain lengthy stretches of tandemly repeated simple DNA sequences composed of a G-rich strand and a C-rich strand (called terminal repeats). These terminal repeats are highly conserved. Sequences adjacent to the telomeric repeats are often highly polymorphic and rich in repetitive elements (termed subtelomeric repeats); in some cases, genes have been found in the proterminal regions of chromosomes.
- **TigrScan gene**
Gene model generated using TigrScan, a gene-finding tool based on the generalized hidden Markov model (HMM) framework, similar to GENSCAN and Genie. It is highly reconfigurable and includes software for retraining.
- **ToxoCyc**
Database/utility built by analyzing the Toxoplasma gondii genome with SRI International's pathway tools; used for searching and visualizing Toxoplasma metabolic pathway information.
- **Translation**
Synthesis of protein from an mRNA template.
- **Transmembrane domain**
Three-dimensional protein structure that is thermodynamically stable in a membrane. This may be a single alpha helix, a stable complex of several transmembrane alpha helices, a transmembrane beta barrel, a beta-helix of gramicidin A, or any other structure. Transmembrane domains average 20 amino acid residues in length, though they may be much smaller or much longer.
- **Transmembrane protein**
Protein that spans an entire biological membrane.
- **UTR**
Untranslated region. Section of messenger RNA (mRNA) that either precedes (5' UTR) or follows (3' UTR) the coding region and is not itself translated. The UTR

contains several regulatory regions, including the polyadenylation (polyA) site in the 3' UTR, sequences involved in the initiation of translation (in the 5' UTR), and binding regions for proteins and other regulatory molecules in both the 3' and 5' UTR.

- **UniGene**

Project and database at NCBI aimed at defining gene-oriented clusters of expressed sequence tags (ESTs). Sets of ESTs are clustered based on strong sequence homology in an attempt to define a specific, nonredundant cluster for each transcript in a tissue or genome. Each UniGene cluster contains sequences that represent a unique gene in addition to information about the tissue types in which the gene has been expressed and map location.

- **Wildcard character**

Character used to substitute for any other character(s) in a string.

- **Xenolog**

Gene found in an unrelated species and that is related by gene transfer rather than common vertical descent.

- **blastn**

Version of the basic local alignment search tool (BLAST) used to compare a nucleotide query sequence against a nucleotide sequence database.

- **blastp**

Version of the basic local alignment search tool (BLAST) used to compare a protein query sequence against a protein sequence database.

- **blastx**

Version of the basic local alignment search tool (BLAST) used to compare a nucleotide query sequence translated in all reading frames against a protein sequence database.

- **cDNA**

Complementary DNA. DNA molecule synthesized by the enzyme reverse transcriptase using an mRNA as template.

- **cDNA microarray**

Collection of microscopic cDNA spots commonly representing single genes and arrayed on a solid surface (commonly glass slides) by covalent attachment to chemically suitable matrices. Used for expression profiling, the monitoring of the expression levels of hundreds or thousands of genes simultaneously.

- **ePCR**

Electronic PCR (polymerase chain reaction). Computational procedure used to check for uniqueness in spacing and number of primer binding sites within DNA sequences. Searches for subsequences that closely match a set of PCR primers and that have the correct order, orientation, and spacing to make a PCR product. Used to check the expected length of a PCR product, which can provide information regarding unexpected repetitive sequences.

- **ncRNA**

Noncoding RNA. Any RNA that is not translated into a protein. Includes transfer RNA (tRNA), ribosomal RNA (rRNA), small RNAs such as snoRNAs, microRNAs, siRNAs and piRNAs, as well as long ncRNAs.

- **rRNA**

Ribosomal RNA. Component of the ribosomes, which function in protein synthesis.

- **snRNA**

Small nuclear RNA. Class of small RNA molecules found within the nucleus, transcribed by RNA polymerase II or III, and involved in a variety of important processes such as RNA splicing (removal of introns from hnRNA), regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining telomeres. They are always associated with specific proteins, and the complexes are referred to as small nuclear ribonucleoproteins (snRNP) or snurps. These elements are rich in uridine. A large group of snRNAs known as small nucleolar RNAs (snoRNAs) are small RNA molecules that play an essential role in RNA biogenesis and chemical modification of ribosomal RNAs (rRNAs) and other RNA genes (tRNA and snRNAs). They are located in the nucleus and the cajal bodies of eukaryotic cells (the major sites of RNA synthesis).

- **tRNA**

Transfer RNA. Small RNA chain (73-93 nucleotides) that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. A three-base region, the anticodon, pairs to the corresponding three-base codon region on the template mRNA. Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.

- **tblastn**

Version of the basic local alignment search tool (BLAST) used to compare a protein query sequence against a translated nucleotide sequence database.

- **tblastx**

Version of the basic local alignment search tool (BLAST) used to compare the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.