

Data retrieval and download

1a Downloading a set of gene results and associated data from a search result

For this exercise, you can start with any result list you generated this morning, or use this shared strategy that returns a list of *P. vivax* genes that are likely proteases expressed in gametocytes.

<https://plasmodb.org/plasmo/im.do?s=24d4a19a9d15ca92>

Use the Download tool to create a table with one row per gene and columns for the associated data: **Genomic Location, Product Description, Transcript Length and all Curated GO Function.** Which report type would you choose to create your table?

The screenshot shows the Plasmodb.org interface. At the top, a search strategy is defined: "P vivax genes that are likely proteases expressed in gametocytes". The strategy is visualized as a flowchart with four steps: Step 1 (protease, 3315 Genes), Step 2 (GO:proteolysis, 3617 Genes), Step 3 (PF154 Gametocy, 1699 Genes), and Step 4 (Pf to Pvivax, 76 Genes). Below the flowchart, a table shows the results for 76 genes from Step 4. The table has columns for various ortholog groups and a 'Download' button circled in red.

All Results	Ortholog Groups	P.falciparum (0)																					
		Padleri	P.berghei	P.billcollinsi	P.blacklocki	P.chabaudi	P.coatneyi	P.cynomolgi (0)		3D7	7G8	CD01	Dd2	GA01	GB4	GN01	HB3	IT	KE01	KH01	KH02	ML01	
76	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene ID	Transcript ID	Input Ortholog(s)	Organism	Genomic Location (Gene)	Product Description	Ortholog Group
PVX_089425	PVX_089425.1	PF3D7_0818900	P. vivax Sal-1	Pv_Sal1_chr05:541,285..543,933(-)	heat shock 70 kDa protein, putative	OG5_1265
PVX_099315	PVX_099315.1	PF3D7_0818900	P. vivax Sal-1	Pv_Sal1_chr07:670,249..672,876(+)	heat shock protein 70, putative	OG5_1265

- **Tab delimited (Excel) - choose columns to make a custom table**– create a file with one row per gene and unlimited (almost) columns per gene. Any data that is available as a column on the result page can be downloaded with this option.
 - **Tab-delimited** text, also known as **tab-separated** values (TSV), is a format that can be created or viewed by most spreadsheet programs and text editors. The TSV

format follows these rules: Each entry in the **file** takes up a single line. The first line in the **file** is the header line, which labels each field.

- **Tab delimited (Excel) - choose a pre-configured table** – This option allows you to download data that has multiple associations per gene, such as multiple GO terms assigned to one gene. The file structure is NOT one row per gene. Only one table can be downloaded at a time.
- **FASTA (sequence retrieval, configurable)** – create a multi-fasta file of your sequences. Each sequence begins with a single-line description, which contains greater-than (“>”) symbol, followed by lines of sequence data. You have the option to configure the start and end points of the sequence
- **GFF3: Gene models and optional sequences** – a simple **tab delimited** format for describing genomic features in a 9-column text file. GFF stands for *Generic Feature Format*. GFF3 allows multi-level grouping and multi-level descriptive attributes.
Hint: choose the option for a ‘Tab delimited (Excel) - choose columns to make a custom table’ to open the tool. Under Choose Columns you can either expand every category and browse to find the data you want, or you can use the search function.

Results are from search: Transform by Orthology

Choose a Report: Tab delimited (Excel) - choose columns to make a custom table [?](#)
 Tab delimited (Excel) - choose a pre-configured table [?](#)
 FASTA (sequence retrieval, configurable) [?](#)
 GFF3: Gene models and optional sequences [?](#)

Note: IDs will automatically be included in the report and the report will be sorted by ID.

Choose Columns select all | clear all | expand all | collapse all

Search Columns... [?](#)

- Search Specific
 - Input Ortholog(s)
 - Search Weight
- Gene models
 - Annotation, curation and identifiers
 - Gene has Unmatched Transcripts
 - Gene Name or Symbol
 - gene_source_id
 - Previous ID(s)
 - Product Description
 - Transcript Product Description
 - Link outs
 - Genomic Location
 - Chromosome
 - Genomic Location (Gene)
 - Genomic Location (Transcript)
 - Genomic Sequence ID
 - Taxonomy
 - Orthology and synteny
 - Ortholog count
 - Ortholog Group
 - Paralog count
 - Phenotype
 - Genetic variation
 - Transcriptomics
 - Sequences
 - Protein features and properties
 - Protein targeting and localization
 - Function prediction
 - Computed GO Component IDs
 - Computed GO Components
 - Computed GO Function IDs
 - Computed GO Functions
 - Computed GO Process IDs
 - Computed GO Processes
 - Curated GO Component IDs
 - Curated GO Components
 - Curated GO Function IDs
 - Curated GO Functions
 - Curated GO Process IDs
 - Curated GO Processes
 - EC numbers
 - EC numbers from OrthoMCL
 - Proteomics

select all | clear all | expand all | collapse all

Choose Rows Include only one transcript per gene (the longest)

Download Type [?](#)

- Text File
- Excel File*
- Show in Browser

Additional Options [?](#)

- Include header row (column names)

[Get Genes](#)

1b Download the genomic sequences of genes in a list of results. This is a good way to get sequences for further analysis.

Use same results as in 1a. Go to the result page (My Strategies and use the strategy panel to activate the result you want) and choose Download again but this time choose **FASTA (sequence retrieval, configurable)**. Explore the tool. What kind of sequences can you retrieve? Protein? Genomic? Coding?

i. Download your gene sequences in fasta format and include the 500bp upstream of the start sites. Notice that there are options to configure the fasta file header (define). This makes it easier to integrate the file with downstream programs that you may be using since some programs (outside EuPathDB) require specific formats for the define.

Download 76 Genes

Results are from search: Transform by Orthology

Choose a Report: Tab delimited (Excel) - choose columns to make a custom table [?](#)
 Tab delimited (Excel) - choose a pre-configured table [?](#)
 FASTA (sequence retrieval, configurable) [?](#)
 GFF3: Gene models and optional sequences [?](#)

Choose the type of sequence:
 Genomic
 Protein
 CDS
 Transcript

Choose the region of the sequence(s):
Begin at Transcription Start*** | + | - | 0 | nucleotides
End at Transcription Stop*** | + | - | 0 | nucleotides

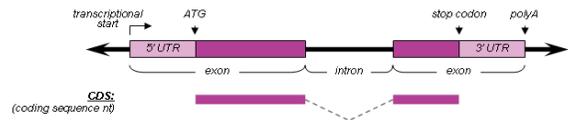
Download Type:
 Text File
 Show in Browser

Fasta define:
 Only Gene ID
 Full Fasta Header

Sequence format:
 Single line
 Default (60 chars on a line)

Note:
For "genomic" sequence: If UTRs have not been annotated for a gene, then choosing "transcription start" may have the same effect as for "protein" sequence: you can only retrieve sequence contained within the ID(s) listed. i.e. from downstream of amino acid sequence start amino acid end (last amino acid in the protein = 0).

Use this section to configure the tool to return the 500bp upstream of the gene.



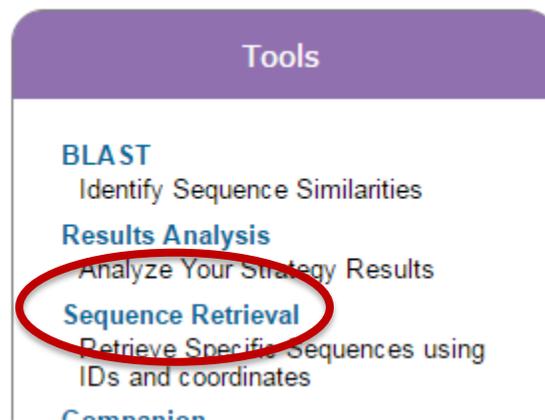
ii. Now retrieve 5' UTR sequences for the list of genes. Begin with setting a transcription start parameter at 0 and end at a translation start (ATG) and parameter (-1). Setting a

translation start (ATG) site parameter to (-1) eliminates incorporating “A” of the start codon into the 5’ UTR sequence.

Did you retrieve 5’UTRs for every gene? Why?

1c Use the Sequence Retrieval Tool to download the genomic sequence for your genes.

In addition to the download FASTA function from a result page, EuPathDB sites also have a Sequence Retrieval Tool (SRT). This tool is not linked to a specific strategy or result. The SRT is accessed from the tools menu on the home page or the Tools dropdown menu:



The tool contains several options for downloading sequences.

- Retrieve Sequences By Gene IDs.
- Retrieve Sequences By Genomic Sequence IDs.
- Retrieve Multiple Sequence Alignments by Contig / Genomic Sequence IDs.
- Retrieve Sequences By Open Reading Frame IDs.

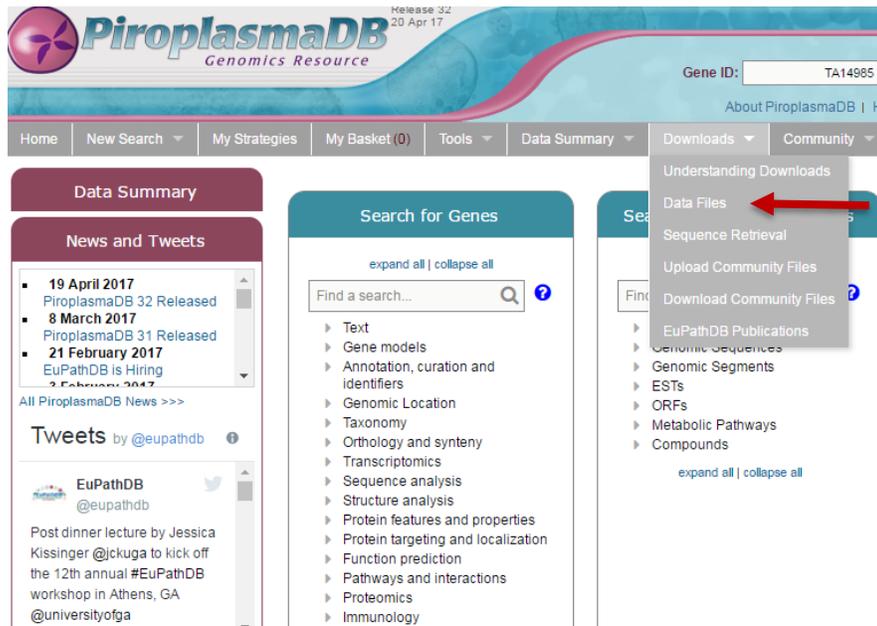
Hint: copy the list of IDs from your gene result into the Retrieve Sequences by Gene ID option of the Sequence Retrieval Tool. How will you retrieve just the gene IDs for your genes? Maybe you can use the download tool described in 1a to retrieve only the IDs.

1d Downloading large data files such as all coding sequences or all protein sequences for an entire genome.

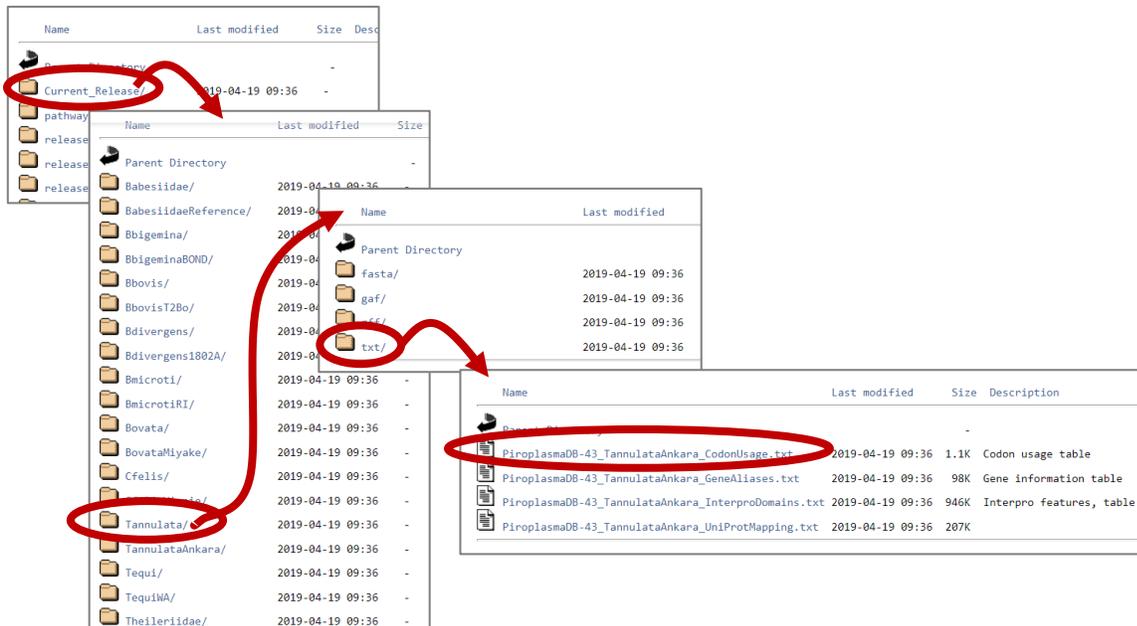
For this exercise use any EuPathDB site. The example below illustrates a use case in PiroplasmaDB: <http://piroplasmadb.org>

Files are available from the Download section of all EuPathDB sites

Hint: select “Data Files” under the “Download” menu in the grey tool bar.



Hint: navigate through the subfolders and find the txt files containing codon usage information for *T. annulata* Ankara. Folders without a strain designation contain species level data.



What other data are available for download? Do the directories make sense ... fasta, gff, txt? How would you download the complete genome sequence and annotation for *T annulata* Ankara?