

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.

<http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf>

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 interface. At the top, there are tabs for 'My Strategies: New, Opened (1), All (84), Basket, Public Strategies (26), Help'. Below this is a strategy workflow diagram for 'Strategy: P vivax genes that are likely proteases expressed in gametocytes (MIMB2017)'. The workflow consists of four steps: Step 1 (protease, 1631 Genes), Step 2 (GO:proteolysis, 2160 Genes), Step 3 (PfNf54 Gametocy, 1699 Genes), and Step 4 (Pf to P vivax, 74 Genes). A red 'Add Step' button is visible.

Below the workflow is a table titled '74 Genes from Step 4' with the same strategy. The table has columns for 'All Results', 'Ortholog Groups', and various Plasmodium species. The 'All Results' column shows 74 genes. The 'Ortholog Groups' column shows 65 groups. The table is filtered to show 74 genes.

At the bottom, there is a 'Gene Results' section with a 'Download' button circled in red. Below this is a table with columns: Gene ID, Transcript ID, Organism, Genomic Location (Transcript), and Input Ortholog(s).

Gene ID	Transcript ID	Organism	Genomic Location (Transcript)	Input Ortholog(s)
PVX_089425	PVX_089425.1	<i>P. vivax</i> Sal-1	Pv_Sal1_chr05:541285..543357(-)	PF3D7_0818900
PVX_089315	PVX_089315.1	<i>P. vivax</i> Sal-1	Pv_Sal1_chr07:670460..672876(+)	PF3D7_0818900
PVX_117322	PVX_117322.1	<i>P. vivax</i> Sal-1	Pv_Sal1_chr12:1772270..1773851(-)	PF3D7_1462800

- **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.

Download 74 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 [?](#)

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... Q [?](#)

- ▼ Search Specific
 - Input Ortholog(s)
 - Search Weight
- ▶ Gene models
- ▶ Annotation, curation and identifiers
- ▶ Link outs
- ▼ Genomic Location
 - Chromosome
 - Genomic Location (Gene)
 - Genomic Location (Transcript)
 - Genomic Sequence ID
- ▶ Taxonomy
- ▶ Orthology and synteny
- ▶ Genetic variation
- ▶ Sequences
- ▶ Protein features and properties
- ▶ Protein targeting and localization
- ▶ Function prediction

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)

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 list of results as in 1a. Choose Download again but this time choose **FASTA (sequence retrieval, configurable)**. Explore the tool. What kind of sequences can you retrieve? Protein? Genomic? Coding?

Download your gene sequences in fasta format and include the 500bp upstream of the start sites.

Download 74 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

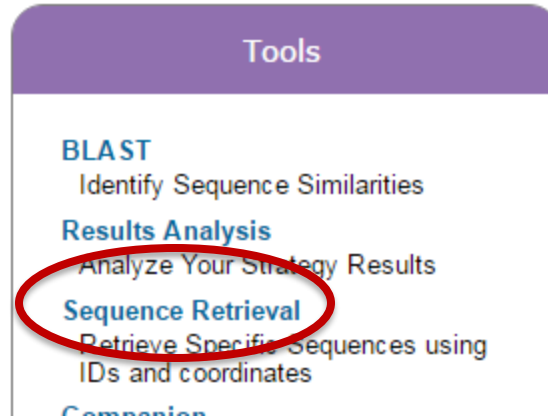
Note:
For "genomic" sequence: If UTRs have not been annotated for a gene, then choosing "transcription start" may have the same effect as choosing "transcription start + 500".
For "protein" sequence: you can only retrieve sequence contained within the ID(s) listed, i.e. from downstream of amino acid sequence to the amino acid end (last amino acid in the protein = 0).

Diagram:
transcriptional start → ATG → stop codon → polyA
← 5' UTR → exon → intron → exon → 3' UTR →
CDS: (coding sequence n)

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.

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

Note that you can download sequence with the sequence retrieval tool (SRT) accessed from the tools menu on the home page:



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.

PiroplasmaDB Genomics Resource

Release 32 20 Apr 17

Gene ID: TA14985

Home New Search My Strategies My Basket (0) Tools Data Summary Downloads Community

Data Summary

News and Tweets

19 April 2017 PiroplasmaDB 32 Released

8 March 2017 PiroplasmaDB 31 Released

21 February 2017 EuPathDB is Hiring

All PiroplasmaDB News >>>

Tweets by @eupathdb

EuPathDB @eupathdb

Post dinner lecture by Jessica Kissinger @cjckuga to kick off the 12th annual #EuPathDB workshop in Athens, GA @universityofga

Search for Genes

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Transcriptomics
- Sequence analysis
- Structure analysis
- Protein features and properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

Downloads

- Understanding Downloads
- Data Files
- Sequence Retrieval
- Upload Community Files
- Download Community Files
- EuPathDB Publications
- Genomic Sequences
- Genomic Segments
- ESTs
- ORFs
- Metabolic Pathways
- Compounds

expand all | collapse all

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.

Name	Last modified	Size	Description
Current_Release/			
release-1.0/			
release-1.1/			
release-2.0/			
release-3.0/			
release-4.0/			
release-5.0/			
TannulataAnkara/	31-Jan-2014 11:14	-	
fasta/	31-Jan-2014 11:14	-	
gff/	31-Jan-2014 11:14	-	
txt/	31-Jan-2014 11:14	-	
PiroplasmaDB-5.0_TannulataAnkara_CodonUsage.txt	31-Jan-2014 11:14	1.1K	Codon usage table
PiroplasmaDB-5.0_TannulataAnkara_GeneAliases.txt	31-Jan-2014 11:14	139K	Gene information table
PiroplasmaDB-5.0_TannulataAnkara_InterproDomains.txt	31-Jan-2014 11:14	910K	Interpro features, table
PiroplasmaDB-5.0_TannulataAnkara_UniProtMapping.txt	31-Jan-2014 11:14	207K	
PiroplasmaDB-5.0_TannulataAnkaraGene.txt	31-Jan-2014 11:14	44M	Gene information table
PiroplasmaDB-5.0_TannulataAnkaraSequence.txt	31-Jan-2014 11:14	845K	Assembly and scaffold com

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?