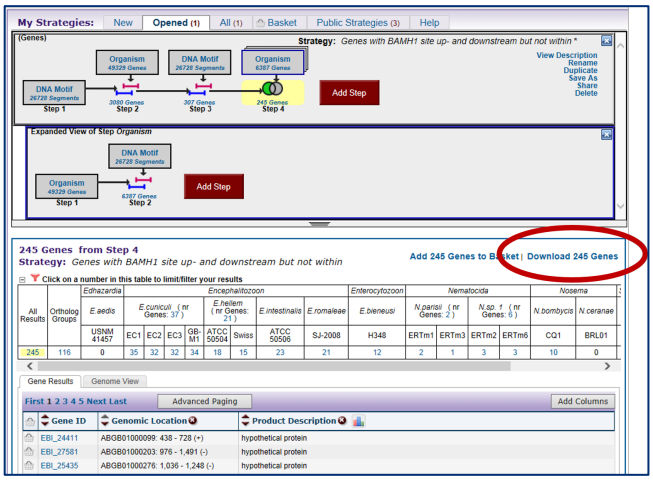
Data retrieval and download

10.1 Downloading a set of gene results and associated data.

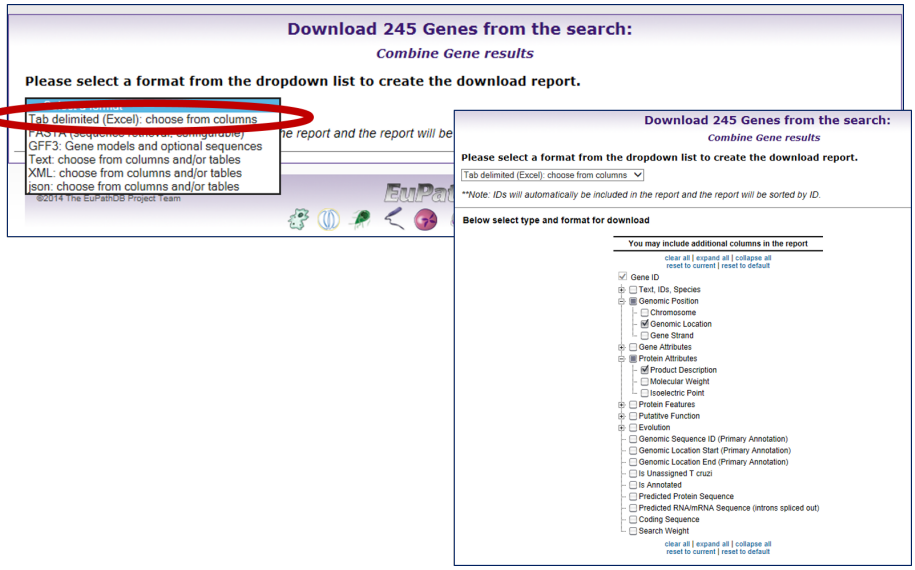
For this exercise you can start with any gene list of results. Start with any result list you generated this morning, such as the DNA Motif search. For example, you could use the strategy below if you don't have one you prefer.

http://microsporidiadb.org/micro/im.do?s=d106dcb74a962249

Download this list of genes with the following associated data: Genomic Location, Product Description, Transcript Length and Predicted GO Function.
Hint: click on the Download ## Genes link.



Hint: select the Tab delimited type of report to download and then click on the boxes to customize your report. The gene ID is automatically downloaded and so is not an option in the popup.

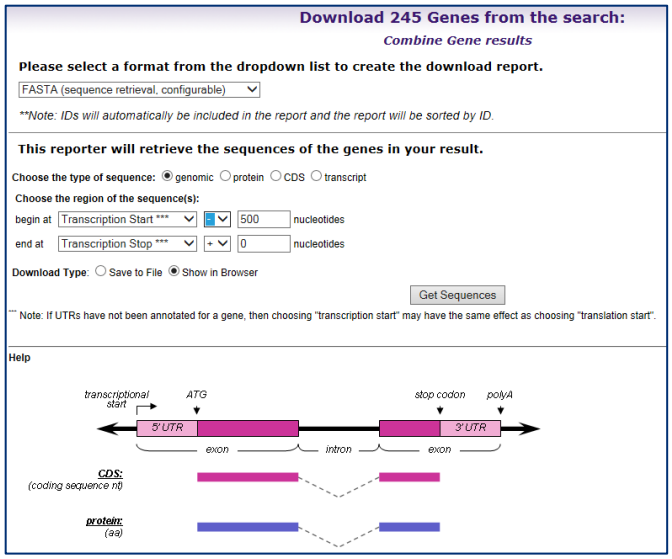


- **Tab delimited (Excel): choose from columns** – create a file with one row per gene and unlimited columns per gene. Any data that is available as a column on the result page can be downloaded with this option.
- **FASTA (sequence retrieval, configurable)** – create a multi-fasta file of your sequences. 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. GFF3 allows multi-level grouping and multi-level descriptive attributes.
- **Text: choose from columns and/or tables** – create a text file of data associated with your sequences. This options 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.
- **XML: choose from columns and/or tables** – create an xml file of columns or tables of data for your sequences. XML is commonly use for the interchange of data over the Internet.
- **json**: choose from columns and/or tables – Create a json formatted file that can be used as an alternative to xml

10.2 Download the sequences of genes in a list of results.

What if you are interested in examining the 5' flanking sequences of these genes? How can you easily get these sequences for subsequent analysis? What kind of sequences can you retrieve? Protein? Genomic? Coding?

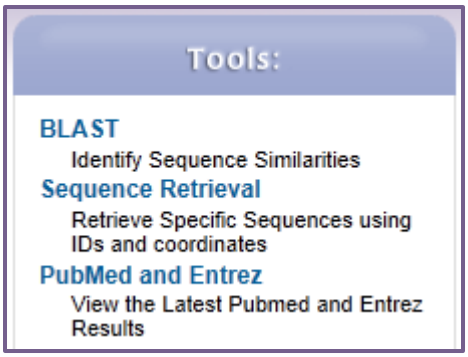
Hint: use same list of results as in 10.1. Choose Download ### Genes again but this time choose **FASTA (sequence retrieval, configurable).** Now, retrieve the 500 nucleotides upstream of the start site of your genes.



10.3 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:

- 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.



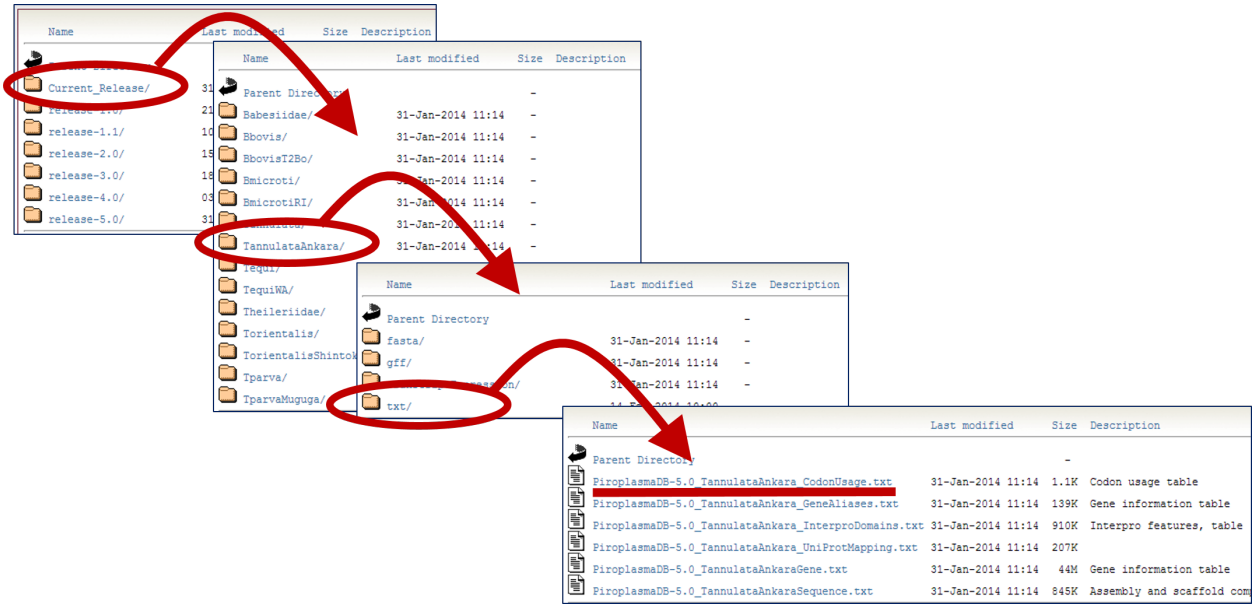
Hint: copy the list of IDs from your gene result into the Retrieve Sequences by Gene ID option of the Sequence Retrieval Tool.

10.4 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, transcriptExpression, txt? Is there any data in the transcriptExpression folder for *T. annulata*? Look at the Transcript Expression searches to determine which of the organisms have this data type.