Data retrieval and download

10.1 Downloading a set of gene results and associated data.

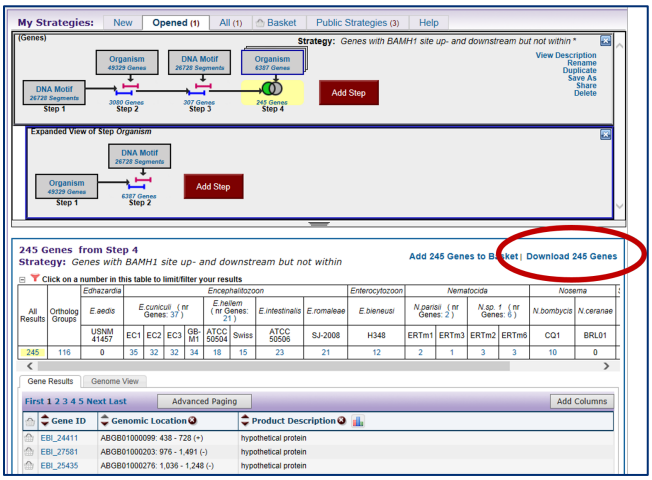
There are several ways to obtain gene results and associated data. You can download this information via:

(1) gene results section by clicking on Download Genes function;
(2) Sequence Retrieval tool in the Tools Section on the right side of the Home page;
(3) Downloads menu at the top and Data files submenu.

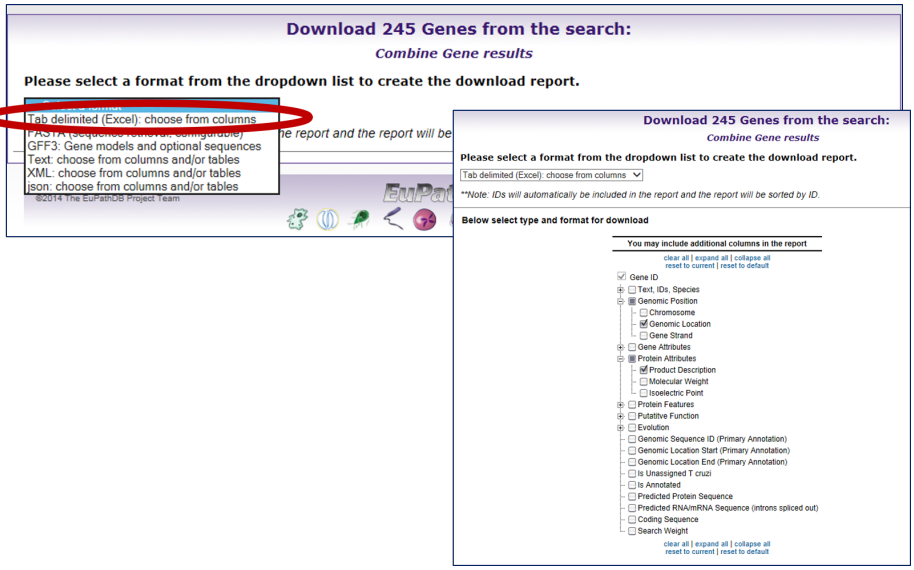
For this exercise you can start with any result list you generated this morning, or pick a strategy from the list below:

http://microsporidiadb.org/micro/im.do?s=0ae5b5bd79a40271

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.
    - **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.

- **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.

- **Text: choose from columns and/or tables** – create a text file of data associated with your sequences. 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.

- **XML: choose from columns and/or tables** – create an xml file of columns or tables of data for your sequences. **XML** (extensive mark-up language) stores data in plain text format and is both human- and machine-readable format. This is also a software- and hardware- independent way of storing, transporting, and sharing data. XML is commonly used for the interchange of data over the Internet.

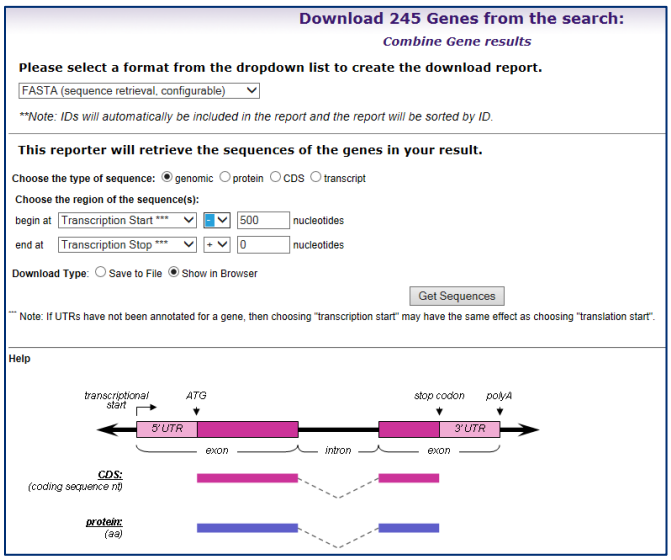
- **json** (JavaScript Object Notation): choose from columns and/or tables – Create a json formatted file that can be used as an alternative to xml. It is a way to store information in an organized, easy-to-access manner. In a nutshell, it gives us a human-readable collection of data that we can access in a really logical manner.

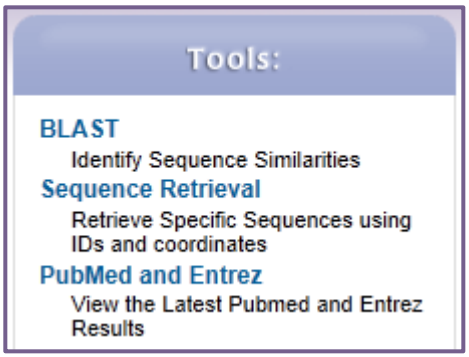
10.2 Download the sequences of genes in a list of results.

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.





Now retrieve 5' UTR sequences for the list of genes. Begin with setting a transcription start parameter at 0 and end at a translation site (ATG) and parameter (-1). Setting a translation start (ATG) site parameter to (-1) would eliminate incorporating "A" of the start codon into the 5' UTR sequence.

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 (Gene IDs can be accessed from gene result page and the uploaded to the retrieve sequence tool that can be accessed form the home page).

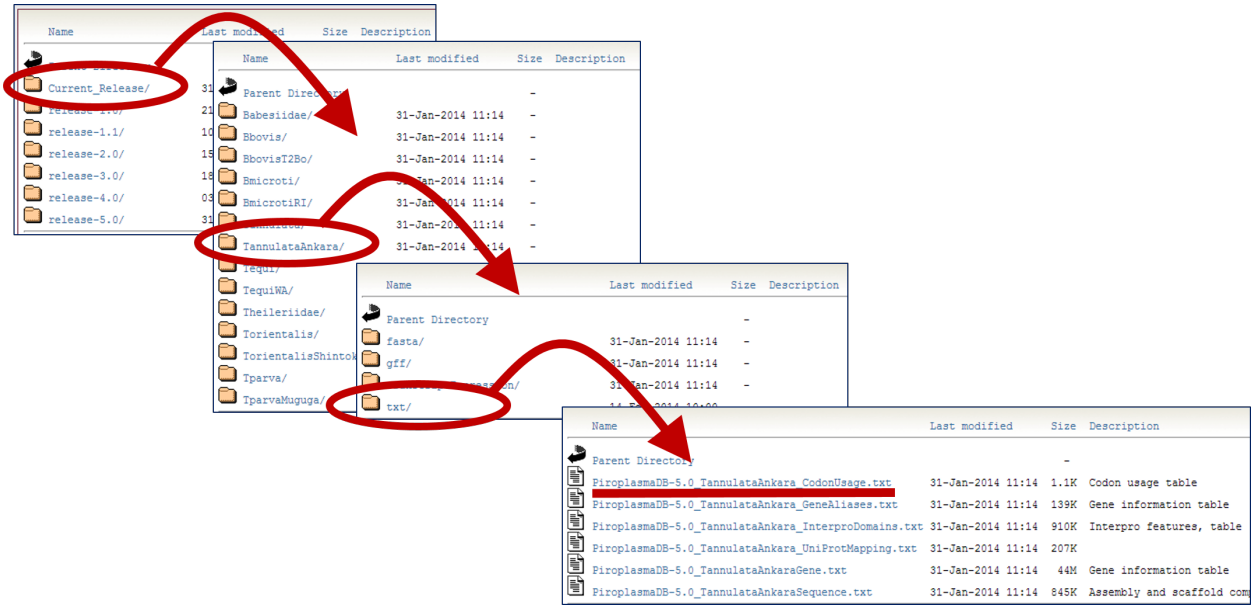
- Retrieve Sequences By Genomic Sequence IDs (Retrieve Genomic Sequence IDs by accessing Home > Genomic Sequence > Organism > Nosema > Get answer).
- Retrieve Multiple Sequence Alignments by Contig / Genomic Sequence IDs (follow instructions above, then select a contig of interest and follow prompts to Retrieve contig sequence or perform multiple sequence alignment in the Sequence section).
- Retrieve Sequences By Open Reading Frame IDs (hint: you can find ORFs using gene IDs).

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