

## Data retrieval and download Exercise 9

### 9.1 Downloading a set of 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.

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

**My Step Result:**

All Results	Ortholog Groups	Encephalitozoon cuniculi	Encephalitozoon hellem	Encephalitozoon intestinalis	Enterocytozoon bieneusi	Nosema ceranae
84	70	31	18	23	12	0

**DNA Motif - step 4 - 84 Genes** [Add 84 Genes to Basket](#) [Download 84 Genes](#)

Gene ID	Genomic Location	Product Description
EBI_24411	ABGB01000099: 438 - 728 (+)	hypothetical protein
EBI_27581	ABGB01000203: 976 - 1,491 (-)	hypothetical protein

Hint: select the 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.

**Download 84 Genes from the search:**  
*Combine Gene results*

Please select a format from the dropdown list to create the download report.  
\*\*Note: Gene IDs will automatically be included in the report.

--- Select a format ---  
 Tab delimited (Excel): choose from columns  
 Text: choose from columns and/or tables  
 Configurable FASTA  
 GFF3: Gene models and optional sequences  
 XML: choose from columns and/or tables  
 json: choose from columns and/or tables

**Generate a tab delimited report of your search result. Select columns to include in the report. Optionally (see below) include a first line with column names.**

**Columns**

clear all | expand all | collapse all  
reset to current | reset to default

- Text, IDs, Species
- Genomic Sequence ID
- Organism
- Genomic Position
  - Chromosome
  - Genomic Location
  - Gene Strand
- Gene Attributes
  - Gene Type
  - # Exons
  - Transcript Length
  - CDS Length
  - Is Pseudo
- Protein Attributes
  - Product Description
  - Molecular Weight

## 9.2 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 this sequence for subsequent analysis?

Hint: use same list of results as in 9.1. Go to the download section and select "Configurable FASTA". Now, retrieve the 500 nucleotides upstream of the start site of your genes.

**Combine Gene results**

Please select a format from the dropdown list to create the download report.  
**\*\*Note: Genes IDs will automatically be included in the report.**

Configurable FASTA

---

**This reporter will retrieve the sequences of the genes in your result.**

Choose the type of sequence:  genomic  protein  CDS  transcript

Choose the region of the sequence(s):

begin at Transcription Start\*\*\* + 0 nucleotides

end at Translation Start (ATG) + 0 nucleotides

Download Type:  Save to File  Show in Browser

[Get Sequences](#)

\*\*\* Note: If UTRs have not been annotated for a gene, then choosing "transcription start" may have the same effect as choosing "translation start".

---

**Help**

The diagram illustrates the structure of a gene and its corresponding sequences. At the top, a gene structure is shown with a 5' UTR, an exon, an intron, another exon, and a 3' UTR ending with a polyA tail. Key features include the transcriptional start site, the ATG start codon, and a stop codon. Below this, four tracks are shown: CDS (coding sequence nt), protein (aa), transcript (CDS + UTRs, if avail), and genomic (includes introns). The CDS track shows the coding sequence with a dashed line indicating the intron. The protein track shows the amino acid sequence. The transcript track shows the full transcript including UTRs. The genomic track shows the full genomic sequence including introns.

Note, that you can access and 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.

**Tools:**

**BLAST**  
Identify Sequence Similarities

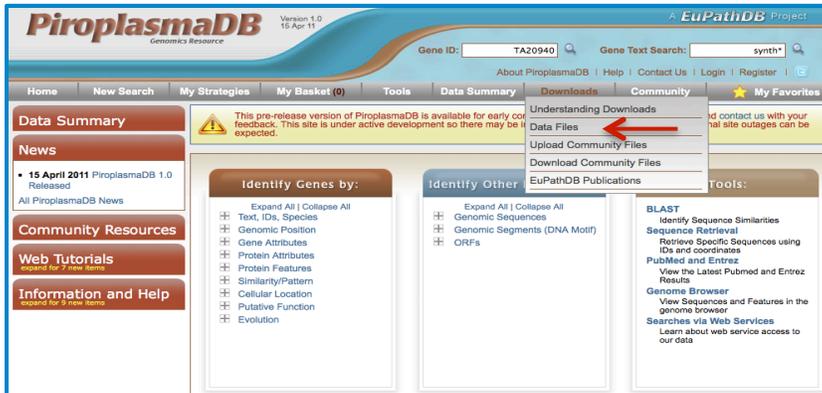
**Sequence Retrieval**  
Retrieve Specific Sequences using IDs and coordinates

**PubMed and Entrez**  
View the Latest Pubmed and Entrez Results

### 9.3 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>

Download files are available in the file 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 files containing codon usage information for *T. annulata* Ankara. Folders without a strain designation contain species level data.

Name	Last modified	Size	Description
Parent Directory		-	
Current_Release/	18-Jan-2013 16:52	-	
release-1.0/	21-Feb-2011 12:54	-	
release-1.1/			
release-2.0/			
release-3.0/			

Name	Last modified	Size	Description
Parent Directory		-	
Babesiidae/	25-Feb-2013 20:21	-	
Bbovis/	18-Jan-2013 16:52	-	
BbovisT2Bo/	18-Jan-2013 16:52	-	
Bmicroti/	18-Jan-2013 16:52	-	
BmicrotiRI/			
Tannulata/			
TannulataAnkara/			
Theileriidae/			
Tparva/			
TparvaMuguga/			

Name	Last modified	Size	Description
Parent Directory		-	
fasta/	18-Jan-2013 16:52	-	
gff/	18-Jan-2013 16:52	-	
transcriptExpression/	18-Jan-2013 16:52	-	
txt/	25-Feb-2013 20:16	-	

Name	Last modified	Size	Description
Parent Directory		-	
data/	25-Feb-2013 20:18	-	
PiroplasmaDB-3.0_TannulataAnkara_CodonUsage.txt	18-Jan-2013 16:52	1.1K	Codon usage table
PiroplasmaDB-3.0_TannulataAnkara_GeneAliases.txt	18-Jan-2013 16:52	139K	Gene information table
PiroplasmaDB-3.0_TannulataAnkara_InterproDomains.txt	18-Jan-2013 16:52	608K	Interpro features, table
PiroplasmaDB-3.0_TannulataAnkara_UniProtMapping.txt	18-Jan-2013 16:52	64K	