

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.

The screenshot shows a multi-step search workflow in iPathDB. The main workflow consists of four steps: Step 1 (DNA Motif, 13856 Segments), Step 2 (Organism, 12339 Genes), Step 3 (DNA Motif, 13856 Segments), and Step 4 (Organism, 1859 Genes). An expanded view of Step 2 shows the transition from DNA Motif to Organism. Below the workflow, a table titled 'My Step Result:' shows the number of results for various species. The 'Download 84 Genes' link is circled in red.

Filter results by species	(results removed by the filter will not be combined into the next step.)					
All Results	Ortholog Groups	Encephalitozoon cuniculi	Encephalitozoon hellem	Encephalitozoon intestinalis	Enterocytozoon bienersi	Nosema ceranae
84	70	31	18	23	12	0

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

Genes | Genome View (beta)

First 1 2 3 4 5 Next Last | Advanced Paging | Select Columns

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.

The screenshot shows the 'Download 84 Genes from the search:' dialog box. It has a title bar 'Download 84 Genes from the search:' and a subtitle 'Combine Gene results'. The main text says 'Please select a format from the dropdown list to create the download report. **Note: Gene IDs will automatically be included in the report.' A dropdown menu is open, showing options: '--- 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', and 'JSON: choose from columns and/or tables'. A red circle highlights the dropdown menu, and a red arrow points to the 'Columns' section on the right. The 'Columns' section has a title bar 'Columns' and a subtitle '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.' It contains a list of columns with checkboxes: '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', and 'Molecular Weight'. The 'Genomic Location', 'Transcript Length', and 'Product Description' checkboxes are checked.

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

*** 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 model shows a 5' UTR, an exon, an intron, another exon, a stop codon, and a 3' UTR. The transcriptional start site is marked with an arrow at the beginning of the 5' UTR, and the ATG start codon is marked at the beginning of the first exon. The stop codon is marked at the end of the second exon, and the polyA site is marked at the end of the 3' UTR. Below the gene model are four tracks representing different sequence types: CDS (coding sequence in nucleotides), protein (in amino acids), transcript (including CDS and UTRs if available), and genomic (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:

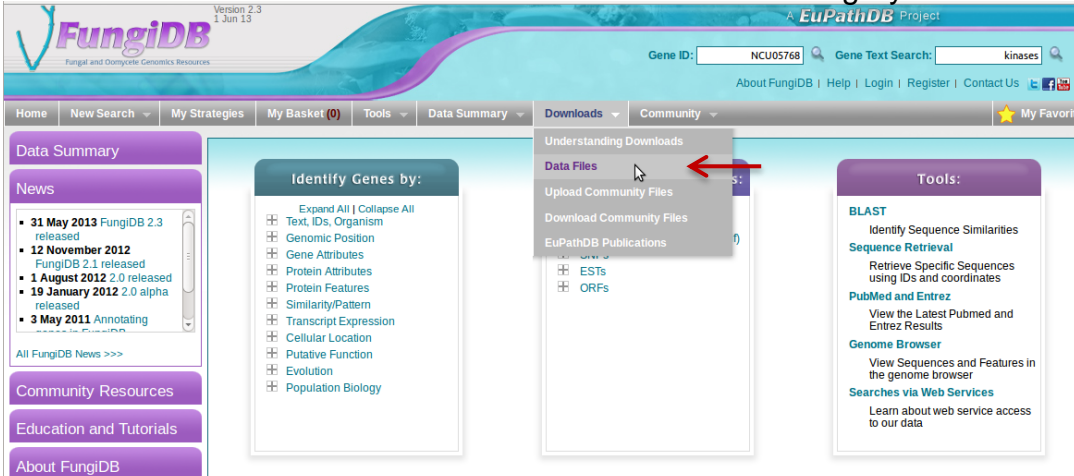
BLAST
Identify Sequence Similarities

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.

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/	16-Apr-2013 17:44	-	
release-2.0/	02-Nov-2012 19:37	-	
release-2.1/	02-Nov-2012 16:46	-	
release-2.3/			
Ndiscreta_FGSC_8579/	14-May-2013 14:06	-	
Nfischeri/	14-Mar-2013 17:18	-	
Nfischeri_NRI			
Ntetrasperma			
Ntetrasperma			
Pcapsici/			
Pcapsici_LTL			
Pchrysospori			
Pchrysospori			
Paraminis/	14-Mar-2013 17:18	-	

Name	Last modified	Size	Description
Parent Directory		-	
fasta/	14-Mar-2013 17:18	-	
gff/	14-Mar-2013 17:19	-	
transcriptExpression/	14-Mar-2013 17:18	-	
txt/	14-Mar-2013 17:19	-	

Name	Last modified	Size	Description
Parent Directory		-	
data/	08-Mar-2013 23:54	-	
<u>FungiDB-CURRENT_Phyca_LT1534_CodonUsage.txt</u>	14-Mar-2013 17:19	1.1K	Codon usage table
FungiDB-CURRENT_Phyca_LT1534_InterproDomains.txt	14-Mar-2013 17:19	3.0M	Interpro features,