Data retrieval and download

10.1 Downloading a set of gene results and associated data.

For this exercise you can start with any <u>gene</u> 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.

Му	~																	
	Stra	tegie	s: N	ew	Op	pened	1)	All (1) 🔿 Basket	Public	Strategies (3)	Hel	р					
(Ger	nes)								s	trategy: G	enes with BAM	IH1 site	up- and	downstr	eam bu	t not within	* 🖂	
	DNA N 6728 Seg Step	Motif gments o 1	Org 4932 3080 51	Generation 2	1 ui	D) 2672 30	A Moti Segme Genes tep 3	tif ents 6	Organism 6397 Genes 245 Genes Step 4	Add	l Step					View Desc R Du S	ription ename plicate ave As Share Delete	
Г	Expand	ed Viev	of Step O	rganis	sm	_	_	_										
	On 493	ganism 129 Gener Step 1	26	DNA M 728 Se 5387 G Step	Motif gments		Add	Step										~
				_	_					_					-			
24	245 Genes from Step 4 Strategy: Genes with BAMH1 site up- and downstream but not within Add 245 Genes to Backet, Download 245 Genes																	
Sti	5 Gen ategy	nes fi y: Ge	rom Ste nes with umber in f	p 4 BA∧	1H1 :	site up limit/filt	- and	d dowi	nstream but n	ot within		Add 2	45 Gen	es to B	sket I	Download	245 Genes	
Sti	5 Gen ategy 7 Click	nes fi y: Ge kon an	rom Ste nes with umber in t	p4 BAM	1H1 : ble to	s <i>ite up</i> limit/filt	- and er your Ei	d dowi r result	nstream but n s	ot within	Enterocytozoon	Add 2	45 Gen	es to B	sket I	Download	245 Genes	T
Sti	5 Gen rategy Click	tholog roups	rom Ste nes with umber in t Edhazardia E.aedis	p 4 BAN his tal	1H1 : ble to Conic Gene	site up limit/filt uli (nr s: 37)	- and er your Er	d down r results incephal E hellen (nr Geni 21)	nstream but n itozoon 1 ss: E.intestinalis	ot within E.romaleae	Enterocytozoon E.bieneusi	Add 2	45 Gen Nema sii (nr ss: 2)	es to Batocida	sket	Nos N.bombycis	245 Genes ema N.ceranae	
E 1 E 1 Res	S Gen rategy Click	tholog	rom Ste nes with umber in t Edhazardia E.aedis USNM 41457	p 4 BAM his tat EC1	1H1 : ble to Gene EC2	site up limit/filt wi (nr s: 37) EC3	- and er your Er (iB- A1 1 50	d down r results incephal E hellen (nr Geni 21) TCC 5504 Si	nstream but n itozoon s: Eintestinalis wiss S0506	ot within Eromaleae SJ-2008	Enterocytozoon E.bieneusi H348	Add 2 N.pari Genu ERTm1	45 Gen Nema sii (nr si: 2) ERTm3	es to B. atocida N.sp. Gene ERTm2	tket I	Nos Nos N.bombycis CQ1	ema N.ceranae BRL01	
Z4 Str Res	5 Gen rategy Click Ul Ori sults Gr	tholog 116	rom Ste nes with umber in t Edhazardia E.aedis USNM 41457 0	p 4 BAM his tat EC1 35	1H1 : ble to Gene EC2 32	site up limit/filt ul/ (nr s: 37) EC3 (32	- and er your E/ ((HB- A1 50 34	d down results incephal E hellen (nr Geni 21) TCC 5504 St 18	nstream but n itozoon s: Eintestinalis wiss ATCC 50506 15 23	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Gene ERTm1 2	45 Gen Nema sil (nr sil (nr sil 2) ERTm3 1	es to B	tket I 1 (nr ts: 6) ERTm6	Nos N.bombycis CQ1 10	ema N.ceranae BRL01 0	5 F-1
E A Res	5 Gen rategy Click Ulls Orf Gl	tholog roups	rom Ste nes with umber in t Edhazardia E.aedis USNM 41457 0	p 4 BAN his tat EC1 35	1H1 : ble to Gene EC2 32	site up limit/filt wi (nr s: 37) EC3 (32	- and er your Ei (18- A1 A1 50 34	d down r results incephal E helien (nr Geni 21) TCC 0504 Si 18	nstream but n itozoon ss: E intestinalis wiss ATCC 50506 15 23	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Geni ERTm1 2	45 Gen Nema si/ (nr ts: 2) ERTm3 1	es to B	t (nr t; 6) ERTm6 3	Nos N.bombycis CQ1 10	ema N.ceranae BRL01 0	
E Res	5 Gen rategy Click dl Ori sults Gr 45	tholog roups 116	rom Ste nes with umber in t Edhazardia E.aedis USNM 41457 0	p 4 BAM his tat EC1 35	1H1 : ble to Gene EC2 32	site up limit/filt wi (nr s: 37) EC3 (32	- and er your E (18- A1 A1 50 84 1	d down r result: Encephal E heilen (nr Geni 21)) TCC 0504 Si 18	Instream but n itozoon ss: E intestinalis viss ATCC 50506 15 23	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Geni ERTm1 2	45 Gen Nema sii (nr 15:2) ERTm3	atocida N.sp. Gene ERTm2 3	(nr (nr (5:6) ERTm6	Nos N.bombycis CQ1 10	ema N.ceranae BRL01 0	
Z4 Sti A Res 2 2 G Fi	5 Gen rategy Clice Ull Ori Gi Sills Ori Gi Sills Ori Gi Sills Ori Gi Sills Ori Gi Sills Ori Gi Sills Ori Gi Sills Ori Sills Ori Cori Cori Sills Ori Sills Ori Cori Sills Ori Sills Ori Sil	tholog roups 116 2 3 4 5	rom Ste nes with umber in ti Edhazardia E.aedis USNM 41457 0 Genome Next Las	p 4 BAN his tab EC1 35 View	1H1 : ble to Gene EC2 32	site up limit/filt wi (nr s: 37) EC3 (32 Adva	- and er your E (18- A1 A1 50 34 1 nced F	d down r result: incephal E hellen (nr Gen 21) 10504 18	nstream but n itozoon ? Eintestinalis S0506 15 23	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Genu ERTm1 2	45 Gen Nema si∛ (nr ss: 2) ERTm3 1	es to B atocida N.sp. Gene ERTm2 3	t (nr ts:6) ERTm6 3	Nos N.bombycis CQ1 10 Add	ema N.ceranae BRL01 0 > Columns	
Z4 Str Res 2 C	5 Gen ategy Click dl or suits or sine Re rst 1 2	tholog tholog 2 3 4 5 ene ID	rom Ste nes with Edhazardia E.aedis USNM 41457 0 Genome Next Las	p 4 BAA his tat EC1 35 View	International Action Control of C	site up limit/filt wi (nr s: 37) EC3 (32 Adva cation	- and Fryour E (1 (1 (1 1 50 34 - - - - - - - - - - - - -	d down r result: incephal E hellen (nr Gen 21) TCC 5504 Si 18 Paging	nstream but n tozoon se: Eintestinalis Miss ATCC S0506 15 23 Product Des	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Geni ERTm1 2	45 Gen Nema Si (nr s: 2) ERTm3 1	es to B	Eket 1 1 (nr 15:5) ERTm6 3	Nos Nos Nombycis CQ1 10 Add	245 Genes ema N.ceranae BRL01 0 > Columns	5
Z4 Str Res 2 C G	S Gen ategy Click ults Ori uuts G sene Res rst 1 2 S G S G S G S G S G S G S G S G S G S G	tholog roups 1116 2 3 4 5 2 24411	rom Ste nes with umber in ti Edhazardia E.aedis USNM 41457 0 Genome Next Las	p 4 BAN his tat EC1 35 View t nomi	IHI : ble to conic EC2 32 ic Lot 099: 4	site up limit/filt wii (nr s: 37) EC3 (32 Adva ation (38 - 726	- and er your E ((iB- A1 50 34 1 50 34 1 50 34 1 50 34 1 50 50 50 50 50 50 50 50 50 50 50 50 50	d down result: incephal E helien (nr Gen 21) TCC 5504 18 Paging	stream but n stre	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneus/ H348 12	Add 2 N.pari Gen ERTm1 2	45 Gen Nema si (nr s: 2) ERTm3 1	es to B	t (nr t: 6) ERTm6	Nose N. bombycis CQ1 10 Add	ema M.ceranae BRL01 0 Columns	
Z4 Str Res 2 C G G FF	S Gerr ategy Click I orf Gi Suits Orf Charl Gi Suits Orf Suits Orf Coi Suits Orf Suits Orf Coi Suits Orf Suits Orf Coi Suits Orf Suits Orf Suits Orf Suits Orf Suits Orf S	tholog roups 1116 2 3 4 5 ene IC 24411 27581	rom Ste nes with umber in ti Edhazardia E aedis USNM 41457 0 Genome Next Las ABGB ABGB	p 4 BAN his tat EC1 35 View it 01000 01000	1H1 : cunic Gene EC2 32 ic Loo 099: 4 203: 9	site up limit/filt wi (nr s: 37) EC3 (32 Adva cation (38 - 728 76 - 1,4	- and er your El ((() BB- AT 50 94 	d down r results ncephal E hellen (nr Gen 21) TCC S504 18 Paging	stream but n tozoon ? E intestinalis wiss S0506 15 23 Product Des hypothetical prote	ot within Eromaleae SJ-2008 21	Enterocytozoon E.bieneusi H348 12	Add 2 N.pari Gen ERTm1 2	45 Gen Nema si/ (nr s: 2) ERTm3 1	es to B	t (nr t:6) ERTm6 3	Nos N. bombycis CQ1 10 Add	ema N ceranae BRL01 0 > Columns	

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.

Download 245 Genes from the search:
Combine Gene results
Please select a format from the dropdown list to create the download report.
FASTA (sequence retrieval, configurable)
**Note: IDs will automatically be included in the report and the report will be sorted by ID.
This reporter will retrieve the sequences of the genes in your result.
Choose the type of sequence: genomic O protein O CDS O transcript
Choose the region of the sequence(s):
begin at Transcription Start *** 🗸 💽 500 nucleotides
end at Transcription Stop *** V + V 0 nucleotides
Download Type. O save to File I Show in Browser
Get Sequences
Note: If U I Rs have not been annotated for a gene, then choosing "transcription start" may have the same effect as choosing "translation start".
neih
transcriptional ATG stop codon polyA
SUTR 3'UTR
exon exon
costing sequence nt
protein: (33)

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.

Tools:
BLAST Identify Sequence Similarities Sequence Retrieval Retrieve Specific Sequences using IDs and coordinates PubMed and Entrez View the Latest Pubmed and Entrez Results

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: <u>http://piroplasmadb.org</u>

Files are available from the Download section of all EuPathDB sites Hint: select "Data Files" under the "Download" menu in the grey tool bar.

Home New Search - My Str	Version 5.0 12 May 14 nics Resource	Gene ID: TA14985 Q About PiroplasmaDB I a Summary T Downloads T Communi	A EurPathiDB Project s Gene Text Search: synth* Q Help Login Register Contact Us te new ty → ☆ My Favorite
Data Summary News and Tweets	The Working with Parasite Database workshop is June 20th, 2014. To appl	Understanding Downloads Resources works y and for more de Sequence Retrieval Upload Community Files	24, 2014 in Hiroton, UK. The deadline for this <u>expression page</u> .
Community 2014 Event stress and the second stress and the sec	Identify Genes by: Expand All Colleges All Text, Ds, Organism B Geno-Mitric Position B Gene Attributes B Protein Ambrutes B SimilarityPattern B Transcript Expression B Protein Expression B Protein Logression B Protein Logression B Protein Logression B Protein Logression B Protein Logression B Protein Logression	Identif EuPathDB Publications EspantibB Publications Espantib equeces Economic Segments (DNA Moth) ESTS ORFS	Tools: Joint Join
Education and Tutorials expand for 8 new items About PiroplasmaDB	PiropiasmaDB 5.0 12 May 14 ©2014 The EuPanDB Project Team	EuPathDB Please	Contact Us with any questions or comments Strategies WDK

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.