Data retrieval and download

1a Downloading a set of gene results and associated data from a search result

For this exercise, you can start with any result list you generated this morning, or use this shared strategy that returns a list of *P. vivax* genes that are likely proteases expressed in gametocytes.

http://plasmodb.org/plasmo/im.do?s=2db873c2b03b57bf

Use the Download tool to create a table with one row per gene and columns for the associated data: Genomic Location, Product Description, Transcript Length and all Curated GO Function. Which report type would you choose to create your table?

My S	trategi	es:	New	Opened (1) All (8	4) 🏦	Bas	sket	Public	Strategies (2	6) H	lelp							
(Genes) Strategy: P vivax genes that are likely proteases expressed in gametocytes (MiMB2017)																			
protease 1631 Genes Step 1		GO: 2'	GO:proteolysis 2160 Genes 2803 Genes Step 2		PfNf54 Gametocy 1699 Genes 65 Genes Step 3		Pf to Pvivax 74 Genes Step 4		Add Step					Add Description Rename Duplicate Save As Share Delete					
74 G Strat	74 Genes from Step 4 Strategy: P vivax genes that are likely proteases expressed in gametocytes (MiMB2017) □ ▼ Click on a number in this table to limit/filter your results																		
All Results	Ortholog Groups								Plasmodium										
		P.berghei	P.chabaudi	P.coatneyi	P.cynomolgi	P.falciparum (nr Genes: 0)		P.fragile	P.gaboni	P.gallinaceum	P.inui	P.knowlesi	P.malariae	P.ovale curtisi	P.reichenowi	P.relictum	P.vinck Gene	ei (nr es∶0)	P.viva (n Gen 74
		ANKA	chabaudi	Hackeri	strain B	3D7	IT	strain nilgiri	strain SY75	8A	San Antonio 1	strain H	UG01	GH01	CDC	SGS1- like	petteri strain CR	vinckei strain vinckei	P01
74	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gene Results Genome View Analyze Results First 1 2 3 4 Next Last Advanced Paging Download Add to Basket Add Columns																			
	🗘 Gene 🛛	ID 🗘 Transcript ID		‡ Org	韋 Organism 🔕			Genomic Location (Transcript)				韋 Input Ortholog(s) 🕹							
	PVX_0894	25 PVX_089425.1			P. vivax Sal-1				Pv_Sal1_chr05:541285543357(-)					PF3D7_0818900					
	PVX_0993	315 PVX_099315.1			P. vivax Sal-1			Pv_Sal1_chr07:670460672876(+)					PF3D7_0818900						
	PVX_117322 PVX_117322.1			2.1	P. vivax Sal-1				Pv_Sal1_chr12:17722701773651(-)				PF3D7_1462800						

- Tab delimited (Excel) choose columns to make a custom table create a file with one row per gene and unlimited (almost) 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.

- Tab delimited (Excel) choose a pre-configured table 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. Only one table can be downloaded at a time.
- 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.

Hint: choose the option for a 'Tab delimited (Excel) - choose columns to make a custom table' to open the tool. Under Choose Columns you can either expand every category and browse to find the data you want, or you can use the search function.

Download 74 Genes							
Results are from search: Transform by Orthology							
Choose a Report: Tab delimited (Excel) - choose columns to make a custom table Tab delimited (Excel) - choose a pre-configured table FASTA (sequence retrieval, configurable) GFF3: Gene models and optional sequences 							
Note: IDs will automatically be included in the report and the report will be sorted by ID.							
Choose Columns	Choose Rows						
select all clear all expand all collapse all	Include only one transcript per gene (the longest)						
Search Columns Q	Open Download Type						
 Search Specific Input Ortholog(s) Search Weight Gene models Annotation, curation and identifiers Link outs Genomic Location Chromosome Genomic Location (Gene) Genomic Location (Transcript) Genomic Location (Transcript) Genomic Sequence ID Y Taxonomy Orthology and syntemy Sequences Protein features and properties Protein factures and properties Function prediction select all clear all expand all collapse all 	 Text File Excel File* Show in Browser Additional Options Include header row (column names)						
	Submit						

1b Download the genomic sequences of genes in a list of results. This is a good way to get sequences for further analysis.

Use same list of results as in 1a. Choose Download again but this time choose **FASTA** (sequence retrieval, configurable). Explore the tool. What kind of sequences can you retrieve? Protein? Genomic? Coding?

Download your gene sequences in fasta format and include the 500bp upstream of the start sites.

Download 74 Genes		
Results are from search: Transform by Orthology		
Choose a Report: The delimited (Excel) - choose columns to make a custom ta Tab delimited (Excel) - choose a pre-configured table FASTA (sequence retrieval, configurable) GFF3: Gene models and optional sequences Comparison of the sequence of the se	ible 🕢	
Choose the type of sequence:		
 Genomic Protein CDS Transcript 		
Choose the region of the sequence(s):		
Begin at Transcription Start*** V + V 0 nucleotides	Use this section t	0
End at Transcription Stop*** + 0 nucleotides	configure the too	ol
Download Type:	to return the	
● Text File	500bp upstream	of
Get Sequences	the gene	
For "genomic" sequence: If UTRs have not been annotated for a gene, then choosin For "protein" sequence: you can only retrieve sequence contained within the ID(s) lis the amino acid end (last amino acid in the protein = 0).	"transcription start" may have the same e sted. i.e. from downstream of amino acid se	
transcriptional ATG start	stop codon polyA	
5'UTR exon intron	exon	
(coding sequence n)		

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

1c 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:



The tool contains several options for downloading sequences.

- 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. How will you retrieve just the gene IDs for your genes? Maybe you can use the download tool described in 1a to retrieve only the IDs.

1d 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.



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, txt? How would you download the complete genome sequence and annotation for *T* annulata Ankara?