Finding and Exporting Data

Not sure what tool to use to find and export data? BioMart is used to retrieve data for complex queries, involving a few or many genes or even complete genomes.

BioMart

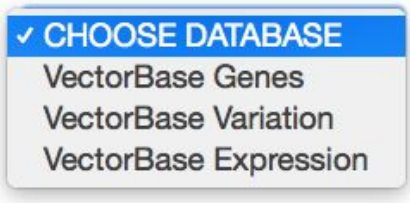
Contents

1. BioMart basics

BioMart is a biological data mining tool. It allows users to look up specific data of different types using a user friendly interface, just with a few click you can perform complex queries

across organisms. In brief, to set up a query you need to apply "Filters" and declare "Attributes", which are the reported data fields. The results can be visualized on screen and also and download in different file formats, e.g., HTML, XLS. To call the script, in addition to the web graphical user interface (GUI), you can also use web services, Perl API and the URL.

These are the different data types available from BioMart[1]:



Genes: gene metadata, predicted molecular and bioinformatic information, sequences

Variation: Single Nucleotide Polymorphisms, SNPs

Expression: Transcript data

2. How to use the tool and interpret its output?

a. Quick start

Go to BioMart (from the Tools navigation menu or from the organism pages)

Choose the Genes database

Select *Anopheles gambiae* dataset (or gene set)

Perform a query selecting "Filters" to ask for protein-coding genes:

- Filters -> gene -> gene type -> protein_coding

[1] BioMart Genes database works with all VectorBase genomes except with *Anopheles sinensis* (China) and *An. stephensi* (Indian). Based on data availability, only a subset of specie is available for the Variation and Expression databases.

- Click on “Count” (gives total of hits) and on “Results” (shown the hits). Notice that out of the 13763 genes in the genome of Anopheles gambiae, 13024 are coding genes. By default the table shows only 10.

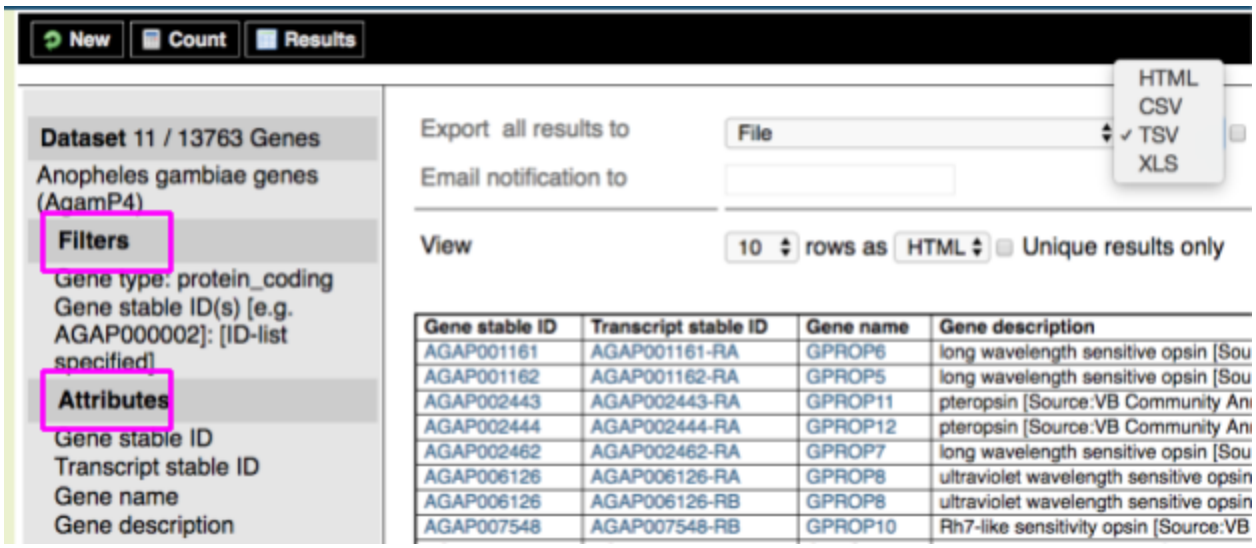
- Refine your query with an ID list. Filters -> gene -> Input external references ID list (type, paste or upload file): AGAP013149, AGAP012985, AGAP012982, AGAP001162, AGAP001161, AGAP002462, AGAP006126, AGAP010089, AGAP007548, AGAP002443, AGAP002444

Select “Attributes” to obtain the following gene metadata

- Attributes -> Features -> Gene -> Ensembl -> Gene name -> Gene description

Click on ‘Count’ and ‘Results’

Export results to -> File -> XLS -> GO



b. Sample queries

Sample query #1

How to export sequence of your genes of interest?

1. Create a text file (*.txt) with these gene ID’s: AGAP001856, AGAP008288, AGAP004261 AGAP006376, AGAP005655. A sample file provided in the tutorial’s page.

2. Go to BioMart.

3. Choose “Genes” as your DataBase and *A. gambiae* as your “Dataset”.

4. Click on “Filters” -> Gene

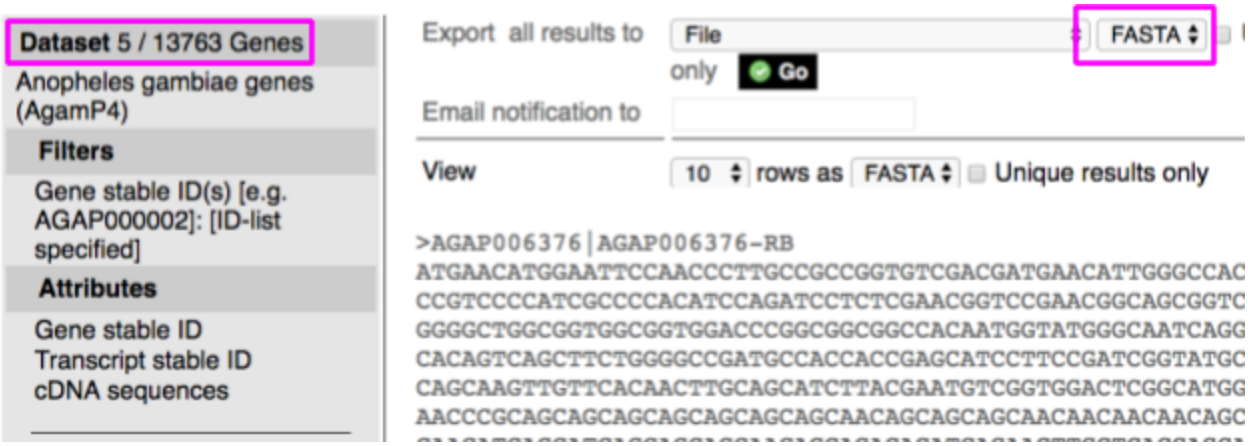
5. Click on the check box for “Input external references ID list”. By default you get “Gene stable ID(s)” in the drop down menu. Click on “Choose File” and upload the text file from step 1 or type/copy the gene IDs in the box provided.

6. Click on “Attributes” -> Sequences -> SEQUENCES

7. Select “cDNA sequences”.

8. Click on “Count” and later on “Results”. You have selected 5 out of the 13,763 *A. gambiae* genes.

9. By default the output is in FASTA format. Click on “Go” to download the results.



Sample query #2

How to find all the homolog genes for your family of interest?

1. Go to BioMart. Choose genes as your database and *A. aegypti* as your dataset.

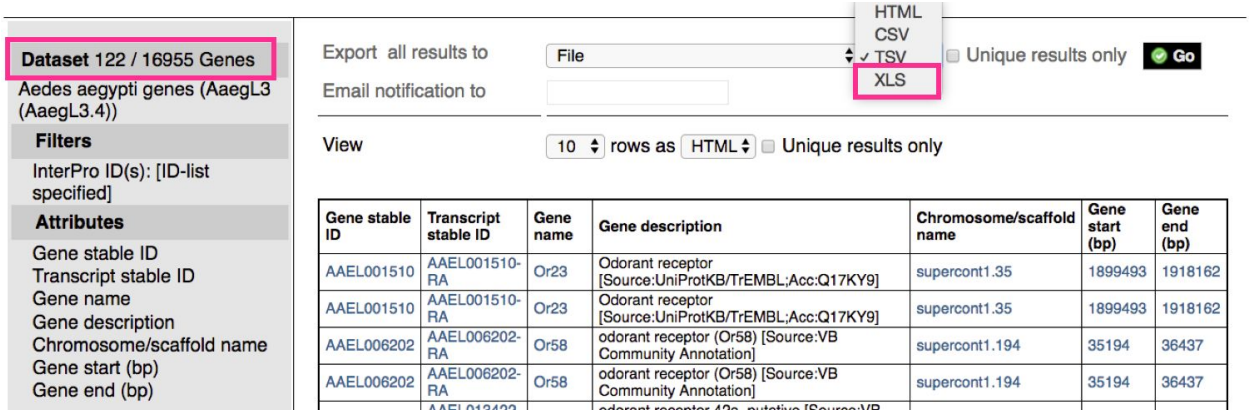
2. Click on Filters and expand the “Protein domains and families” section. Check the box “Limit to genes with these family or domain IDs”, select InterPro ID(s) and type “IPR004117”[2] in the provided box.

[2] You can obtain this information in the Genome Browser from your gene of interest under Transcript-based displays > Domains & features. Use the InterPro accession in the InterPro website, https://www.ebi.ac.uk/interpro/ . IPR004117 comes back as “Olfactory receptor, insect”.

3. Click on Attributes and select Features -> GENE -> Ensembl -> Chromosome/scaffold name, gene start (bp), gene end (bp), gene name, and gene description.

4. Click on "Count" and "Results". There are 122 genes with this "Olfactory receptor, insect" protein domain in *A. aegypti* genome.

5. Export the file XLS format.



Sample query #3

How to find protein-coding genes with transmembrane domains?

1. Go to BioMart. Choose genes as your database and *Culex quinquefasciatus* as your dataset.

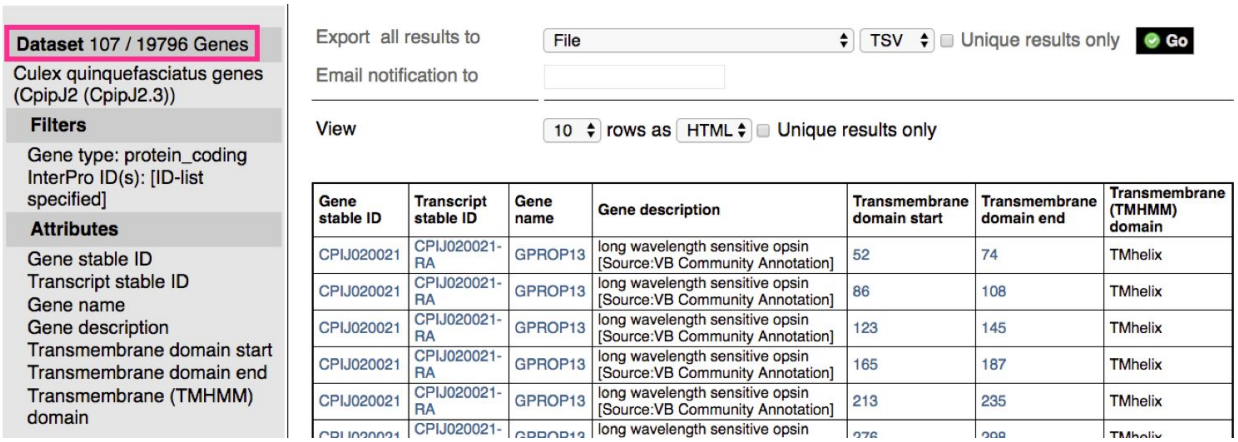
2. Select the filters
   - GENE -> Gene type -> protein_coding
   - PROTEIN DOMAINS AND FAMILIES -> Limit to genes with these family or domain IDs -> InterPro ID(s): IPR017452[3]

3. Select attributes
   - GENE -> Ensembl -> Gene name -> Gene description
   - PROTEIN DOMAINS -> Protein features -> transmembrane helices start & end

4. Click on "Count" and "Results". There are 107 genes with "GPCR, rhodopsin-like, 7TM" protein domain in *C. quinquefascaitus* genome.

5. Select an output format to download.

[3] IPR004117 comes back as "GPCR, rhodopsin-like, 7TM"



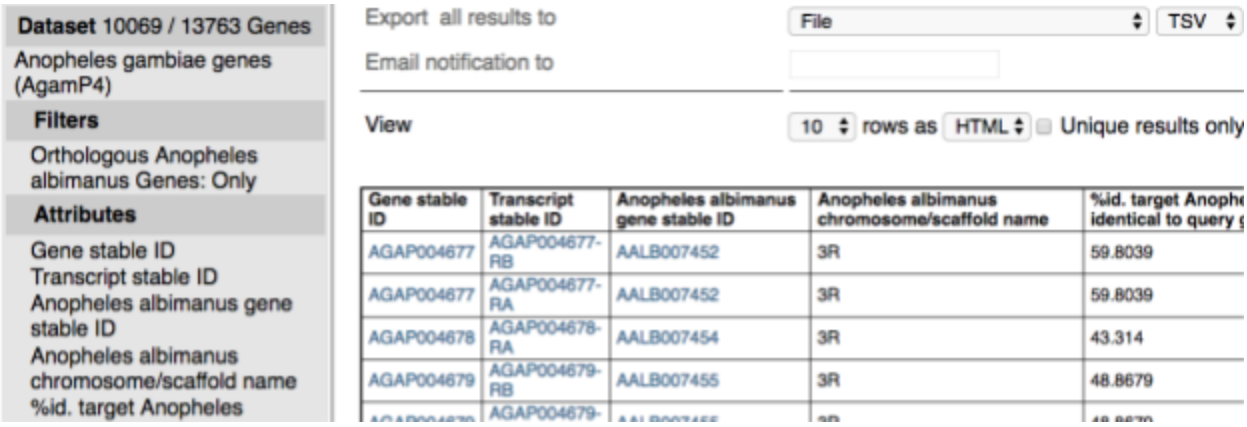
Sample query #4

List of orthologous genes between two species

1. Go to BioMart. Choose genes as your database and *A. gambiae* as your dataset.

2. Select the filters
   - Filters: multi-species comparisons ---> Homolog filters ---> Orthologous *An. albimanus*
   - Attributes:
     Homologs ---> Orthologs ---> *A. albimanus* ---> gene stable ID, chromosome/scaffold, % target & % query, Ortholog confidence [0 low, 1 high][4]

3. Click on Count and Results. There are 10,069 orthologous genes between *A. gambiae and A. albimanus*. The file can be download to sort the genes based on the orthology confidence.



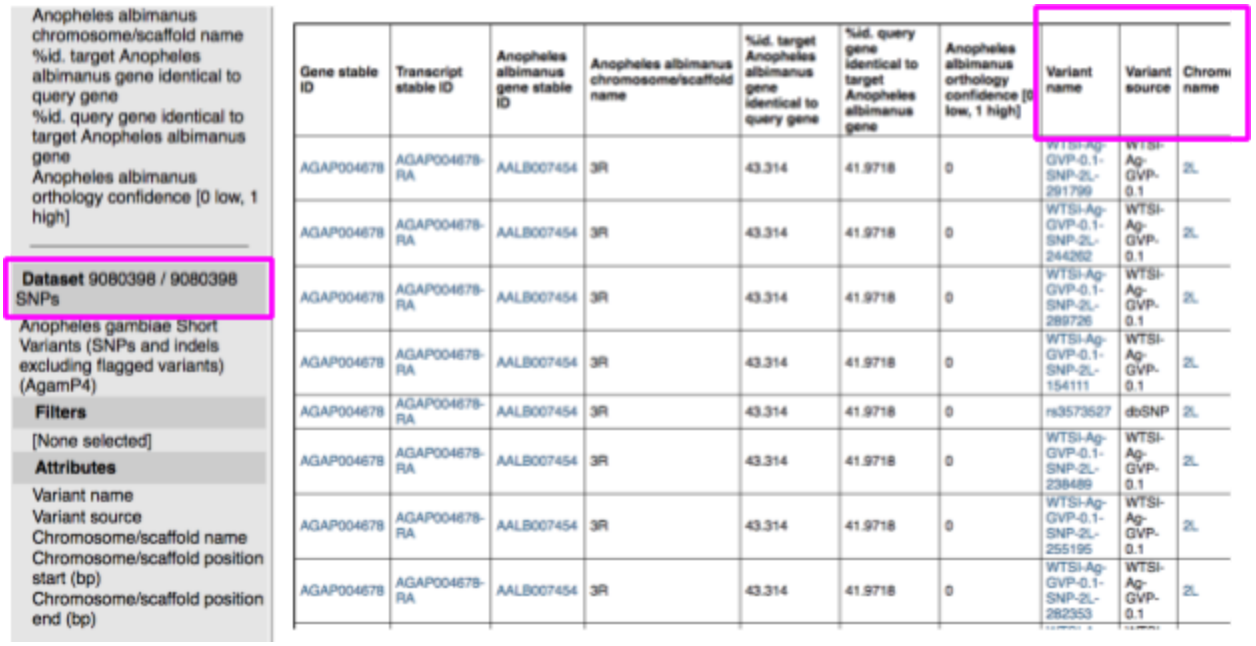
[4] FAQ: How are "high confidence" orthologs defined?, https://www.vectorbase.org/faqs/how-are-high-confidence-orthologs-defined

Sample query #5

How to join two different datasets?

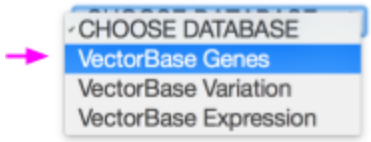
BioMart queries are run against a single species database, but if you click on the lower "Dataset" link, queries can "join" a second dataset. The results of the second dataset are connected and depend on the first one.

This second dataset could be in the gene, variation or expression databases. For query #4 add *A. gambiae* variation (short variants, SNPs and INDELs) as the second dataset



3. Questions and practice exercises

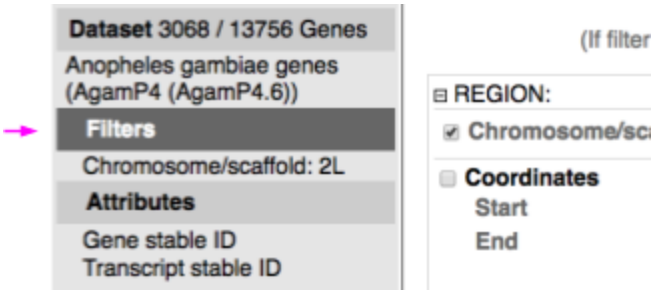
Instructions: Please take note of how do you reach your answerer including the specific "Filters" and "Attributes" used in each case. Use BioMart "Genes" database, to solve these exercises.



To find the number of results matching the filters you have provided, click on "Count" (see below) - to see the first few rows of the results and the attributes you have selected, click on "Results".

a. **Gene content** - using “Filters”

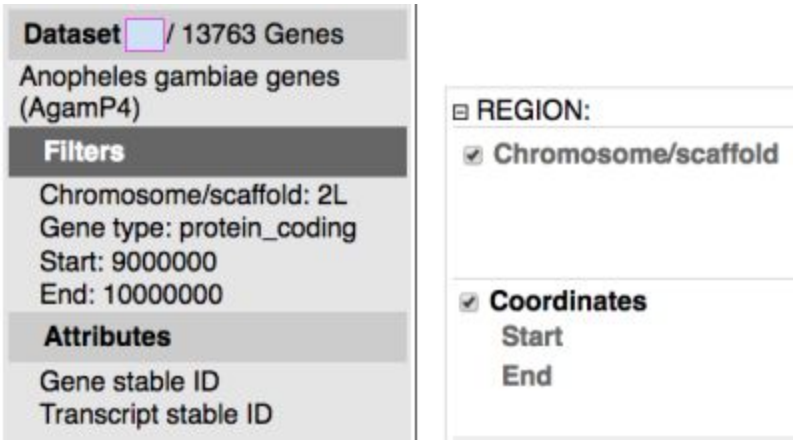
- Choose Database: VectorBase Genes
- Choose Dataset: *Anopheles gambiae* genes
- Click on “Filters”
- Explore the different filter types by expanding and contracting each section of the form.



- How many genes are there on *An. gambiae* chromosome arm 2L?

- Modify this using the "Gene->Gene type" filter to find out how many **protein coding** genes are located in chromosome arm 2L:

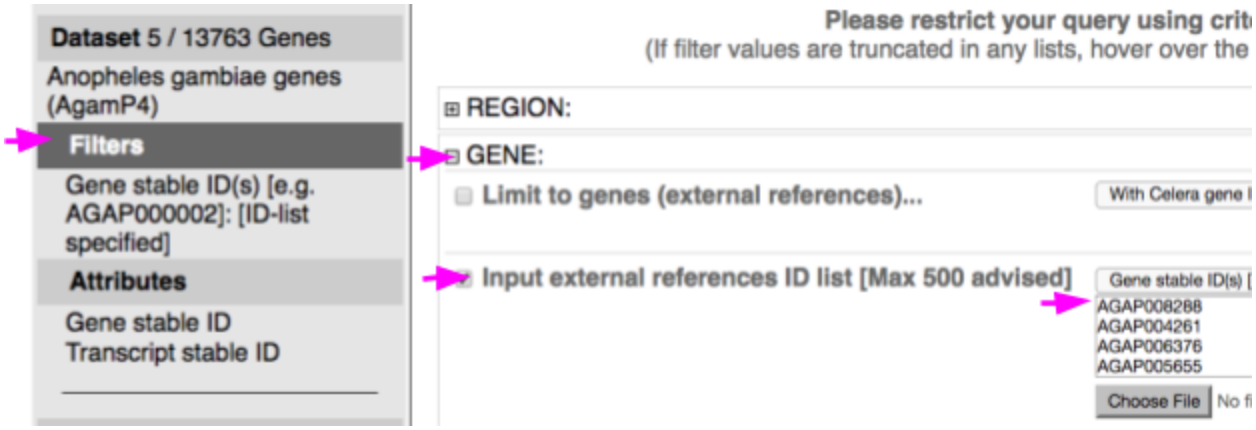
- Add one more filter to find all protein coding genes are between nucleotides 9,000,000 -10,000,000 of 2L? (**Hint**: you can not use periods or commas, see below)



- To start a new session click on “New”

- Now you are going to filter your results using a list of *An. gambiae* gene IDs. Use the sample genes located in Tutorial page: VectorBase_BioMart_SampleGenes_2017.txt

  You can copy-paste in the IDs (see below) or upload the file containing IDs from your computer:

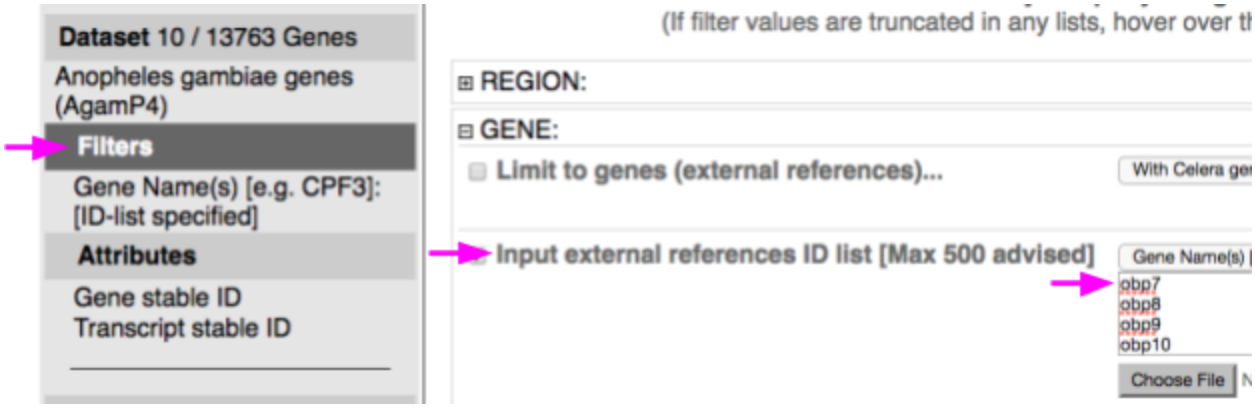


- Click on "Count" and "Results". Results are presented in a Table format. What are the (default) columns of the table?

- Click on "New"

- Again, choose *An. gambiae* genes

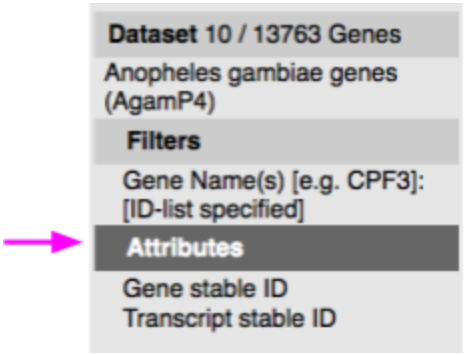
- Filter using the Gene Names (symbols) OBP1, OBP2, … OBP10
  Type them in the provided box. **Hint**: See the screenshot below for guidance when using the ID list limit filter:



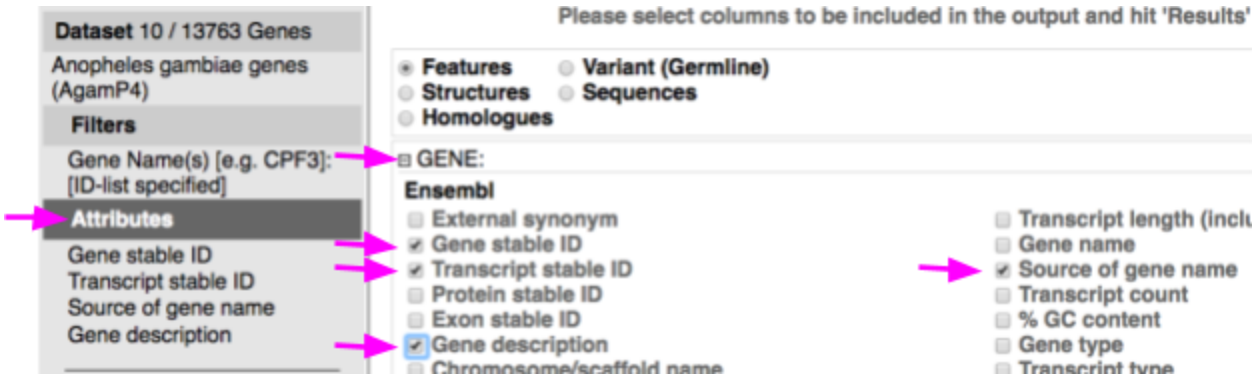
- Think about filters that may be relevant to you.

- Keep the query using 10 OBP gene symbols to start the next section

b. Gene information - defining “Attributes”



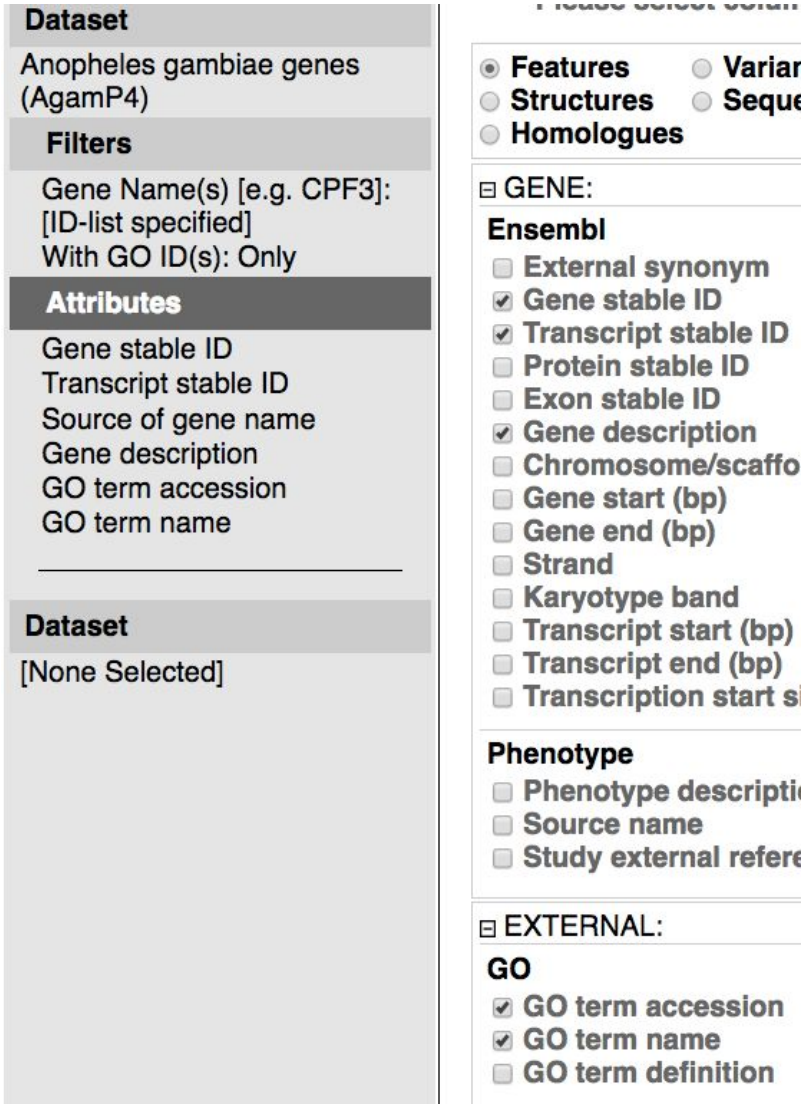
- Show “Gene name” (or symbol) and “Gene description” (or function)



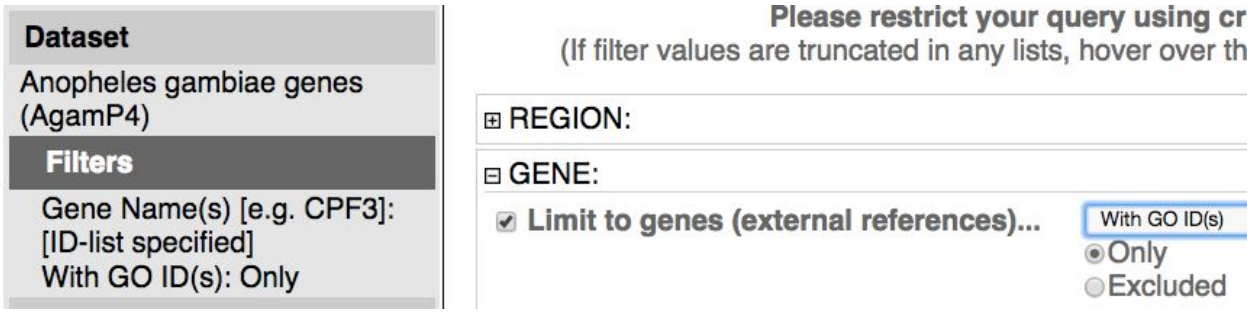
- Select one or more of the GO (Gene Ontology) attributes from the “External” section. How many of these 10 OBP genes have GO terms?

  Hint: use the "Unique results only" checkbox and remove the "Transcript stable ID" attribute.

	Answer
None of the OBP genes have GO terms associated with them.	
Some of the OBP genes have GO terms associated with them.	
All of the OBP genes have GO terms associated with them.	

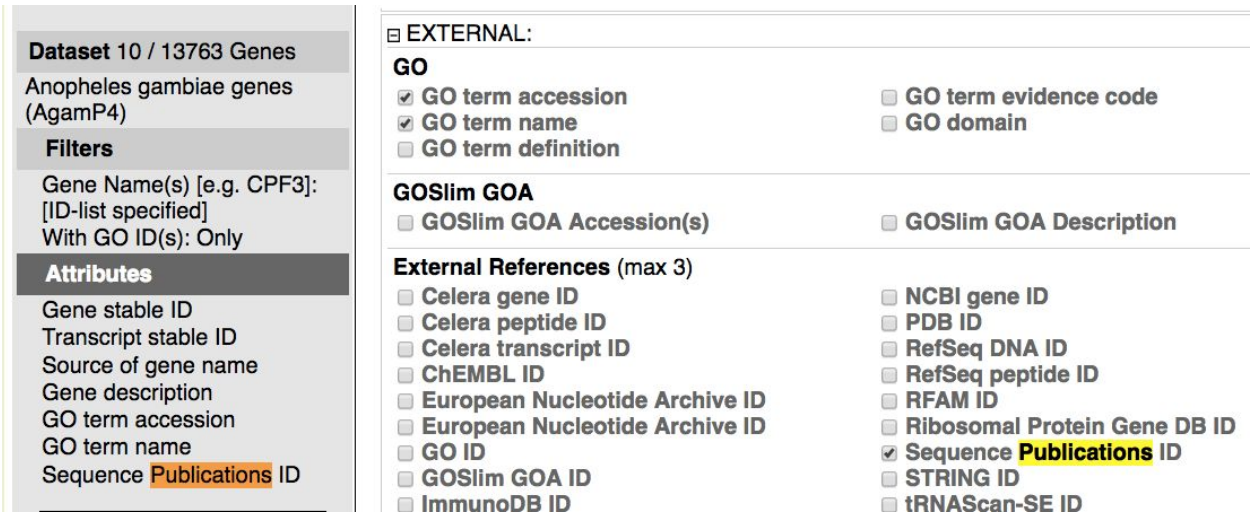


Note, you can also answer this question using a Filter
(Filter -> Gene -> Limit to genes -> with GO ID(s))

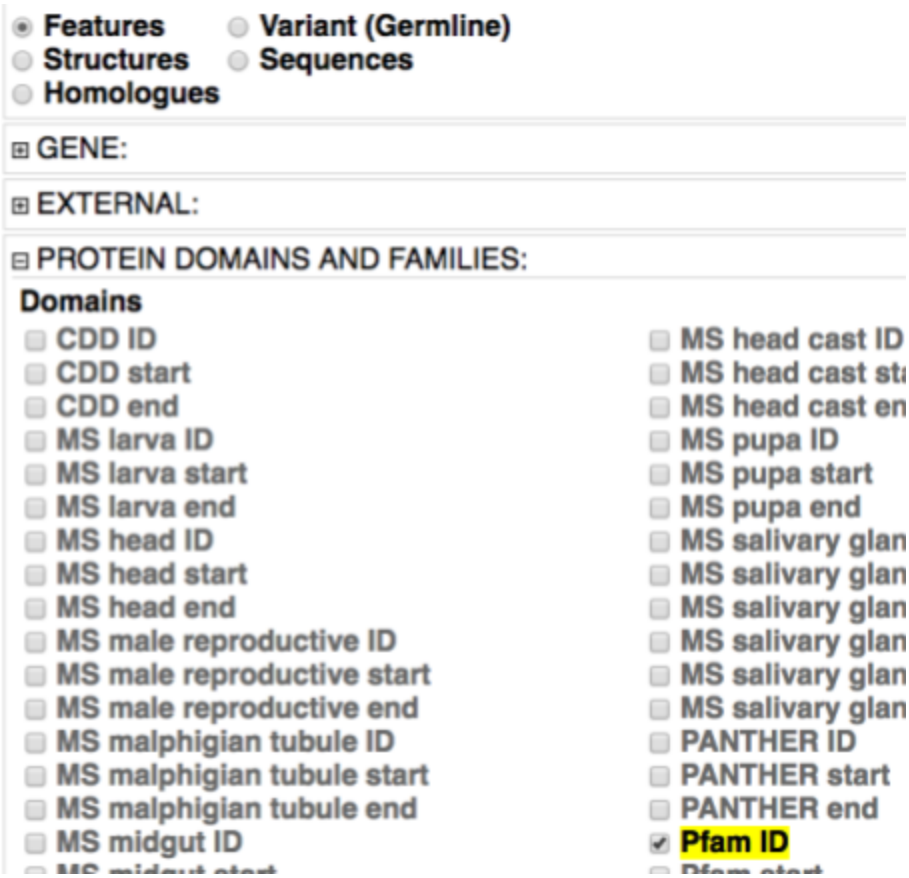


- Do these genes have literature citations (Sequence Publications ID)?

	Answer
None of the OBP genes have publications associated with them.	
Some of the OBP genes have publications associated with them.	
All of the OBP genes have publications associated with them.	

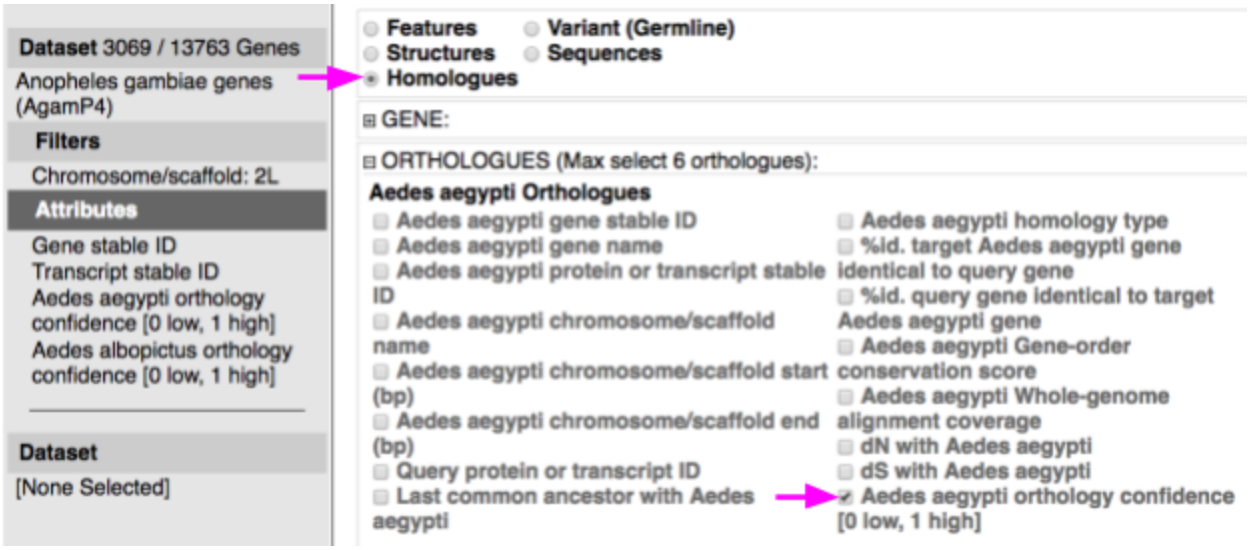


- Query these 10 proteins Pfam domains



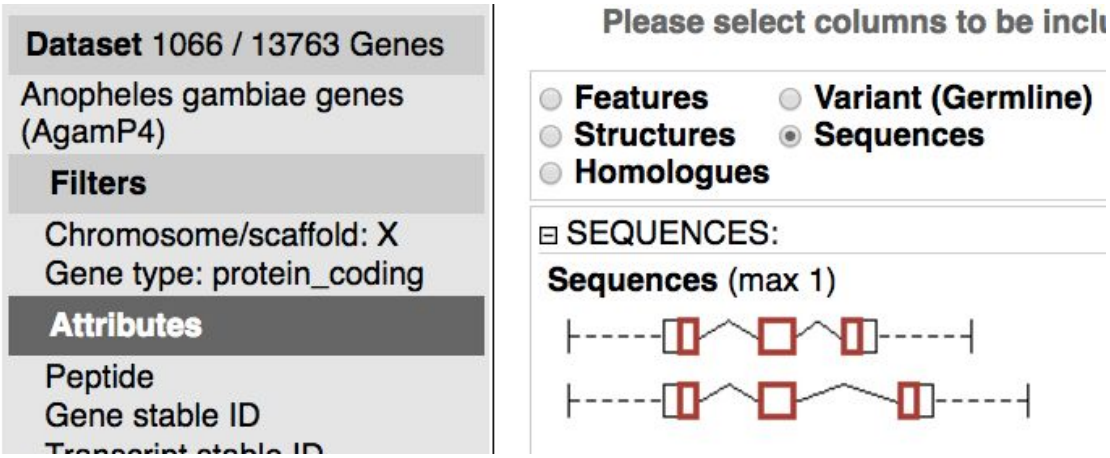
- Start a new session and select *An. gambiae* chromosome arm 2L.

- Show ortholog assignments for *Ae. aegypti and Ae. albopictus*. Select ortholog confidence [0 low, 1 high]. How many *A. gambiae* genes have ortholog to these two species?



c. Other outputs – Attributes → Sequences

- Start a new session and choose *A. gambiae*
- Filter for the protein-coding genes on chromosome X. Click “Count” and “Result”
- Investigate the sequence export options (**Hint**: Click on the circle next to “Sequences” and click on the plus sign next to “SEQUENCES”)



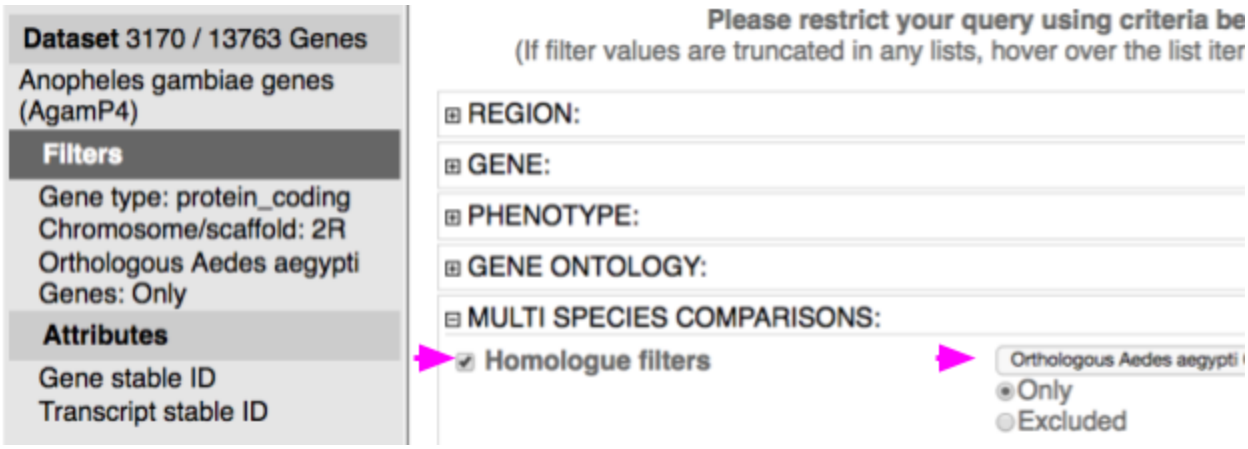
- Look at the various options for download

- Investigate the following, but please note that you don't need to actually download the files, just review the resulting output in the HTML version
    - cDNA sequence → Results
    - Coding sequence → Results
    - 1kb upstream of each locus (*e.g.* for promoter motif prediction analysis) → Results

d. Advanced Practice Exercises

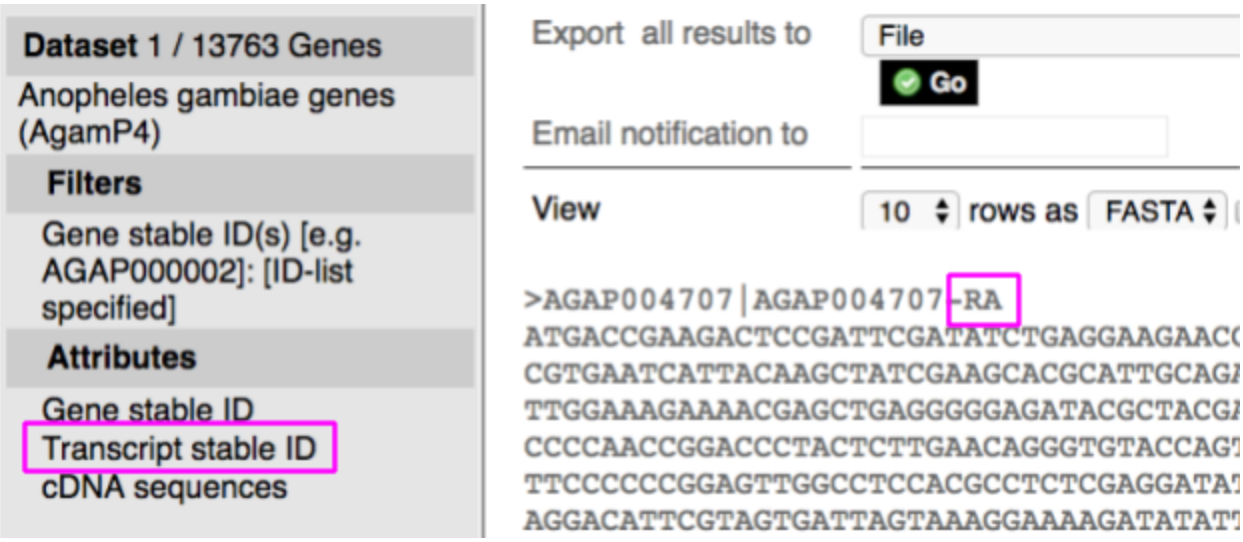
- Start a new session and choose *A. gambiae*

- Find protein-coding genes on *A. gambiae* chromosome 2R with orthologous *A. aegypti* genes.



- Can you add further species?

- Start a new session. Download the splice variants in cDNA format for the gene AGAP004707



- Come up with a meaningful query that relates to your area of interest.

If you need help with any question and its answer contact us at info@vectorbase.org. Because VectorBase data, tools and resources are updated every two months (6 release cycles per year), answers to these exercises will change too.