



## 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

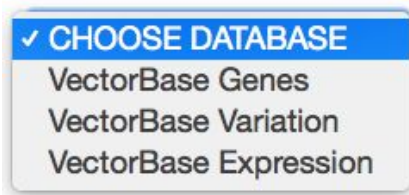
1. BioMart basics
2. How to use the tool and interpret its output?
  - a. Quick start
  - b. Five sample queries:
    - How to export sequence of your genes of interest?
    - How to find all the homolog genes for your family of interest?
    - How to find protein-coding genes with transmembrane domains?
    - List of orthologous genes between two species
    - How to join two different datasets?
3. Questions and practice exercises
  - a. Gene content - using "Filters": number of genes on a specific genomic location, gene type, query with a list of genes
  - b. Gene information - defining "Attributes" : **gene characteristics (symbol and function)**, **Gene Ontology enrichment**, genes with PubMed IDs, protein domains, ortholog assignments
  - c. Other outputs - Attributes → Sequences: gene sequence
  - d. Advanced Practice Exercises: gene type (protein-coding) -> specific genomic location -> homology relationships (orthologous), **download splice variant sequences for a specific gene**

### 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 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<sup>1</sup>:



**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

<sup>1</sup> 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

Gene stable ID	Transcript stable ID	Gene name	Gene description
AGAP001161	AGAP001161-RA	GPROP6	long wavelength sensitive opsin [Sou
AGAP001162	AGAP001162-RA	GPROP5	long wavelength sensitive opsin [Sou
AGAP002443	AGAP002443-RA	GPROP11	pteropsin [Source:VB Community Ani
AGAP002444	AGAP002444-RA	GPROP12	pteropsin [Source:VB Community Ani
AGAP002462	AGAP002462-RA	GPROP7	long wavelength sensitive opsin [Sou
AGAP006126	AGAP006126-RA	GPROP8	ultraviolet wavelength sensitive opsin
AGAP006126	AGAP006126-RB	GPROP8	ultraviolet wavelength sensitive opsin
AGAP007548	AGAP007548-RB	GPROP10	Rh7-like sensitivity opsin [Source:VB

## 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.

The screenshot shows the BioMart interface. On the left, the 'Dataset 5 / 13763 Genes' section is highlighted. Under 'Filters', 'Gene stable ID(s) [e.g. AGAP000002]: [ID-list specified]' is selected. Under 'Attributes', 'cDNA sequences' is selected. On the right, the 'Export all results to' dropdown is set to 'File', and the 'FASTA' format is selected. The 'Go' button is highlighted. Below the export options, the 'View' section shows '10 rows as FASTA Unique results only'. The main content area displays a FASTA sequence for AGAP006376, starting with 'ATGAACATGGAATTCCAACCCCTTGCCCGCGGTGTCGACGATGAACATTGGGCCAC'.

## 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”<sup>2</sup> in the provided box.

<sup>2</sup> 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”.

- Click on Attributes and select Features -> GENE -> Ensembl -> Chromosome/scaffold name, gene start (bp), gene end (bp), gene name, and gene description.
- Click on “Count” and “Results”. There are 122 genes with this “Olfactory receptor, insect” protein domain in *A. aegypti* genome.
- Export the file XLS format.

**Dataset 122 / 16955 Genes**  
Aedes aegypti genes (AaegL3 (AaegL3.4))

**Filters**  
InterPro ID(s): [ID-list specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
Gene name  
Gene description  
Chromosome/scaffold name  
Gene start (bp)  
Gene end (bp)

Export all results to: **File** (dropdown menu shows HTML, CSV, TSV, XLS)  
Email notification to:   
☐ Unique results only **Go**

View: 10 rows as **HTML** (dropdown menu shows HTML, CSV, TSV, XLS) ☐ Unique results only

Gene stable ID	Transcript stable ID	Gene name	Gene description	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)
AAEL001510	AAEL001510-RA	Or23	Odorant receptor [Source:UniProtKB/TrEMBL;Acc:Q17KY9]	supercont1.35	1899493	1918162
AAEL001510	AAEL001510-RA	Or23	Odorant receptor [Source:UniProtKB/TrEMBL;Acc:Q17KY9]	supercont1.35	1899493	1918162
AAEL006202	AAEL006202-RA	Or58	odorant receptor (Or58) [Source:VB Community Annotation]	supercont1.194	35194	36437
AAEL006202	AAEL006202-RA	Or58	odorant receptor (Or58) [Source:VB Community Annotation]	supercont1.194	35194	36437

### Sample query #3

How to find protein-coding genes with transmembrane domains?

- Go to BioMart. Choose genes as your database and *Culex quinquefasciatus* as your dataset.
- 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<sup>3</sup>
- Select attributes
  - GENE -> Ensembl -> Gene name -> Gene description
  - PROTEIN DOMAINS -> Protein features -> transmembrane helices start & end
- Click on “Count” and “Results”. There are 107 genes with “GPCR, rhodopsin-like, 7TM” protein domain in *C. quinquefasciatus* genome.
- Select an output format to download.

<sup>3</sup> IPR004117 comes back as “GPCR, rhodopsin-like, 7TM”

**Dataset 107 / 19796 Genes**  
Culex quinquefasciatus genes (CpipJ2 (CpipJ2.3))

**Filters**  
Gene type: protein\_coding  
InterPro ID(s): [ID-list specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
Gene name  
Gene description  
Transmembrane domain start  
Transmembrane domain end  
Transmembrane (TMHMM) domain

Export all results to  TSV ☐ Unique results only

Email notification to

View  rows as  ☐ Unique results only

Gene stable ID	Transcript stable ID	Gene name	Gene description	Transmembrane domain start	Transmembrane domain end	Transmembrane (TMHMM) domain
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	52	74	TMhelix
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	86	108	TMhelix
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	123	145	TMhelix
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	165	187	TMhelix
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	213	235	TMhelix
CPIJ020021	CPIJ020021-RA	GPROP13	long wavelength sensitive opsin [Source:VB Community Annotation]	278	300	TMhelix

## Sample query #4

List of orthologous genes between two species

- Go to BioMart. Choose genes as your database and *A. gambiae* as your dataset.
- 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]<sup>4</sup>
- 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.

**Dataset 10069 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**  
Orthologous Anopheles albimanus Genes: Only

**Attributes**  
Gene stable ID  
Transcript stable ID  
Anopheles albimanus gene stable ID  
Anopheles albimanus chromosome/scaffold name  
%id. target Anopheles

Export all results to  TSV

Email notification to

View  rows as  ☐ Unique results only

Gene stable ID	Transcript stable ID	Anopheles albimanus gene stable ID	Anopheles albimanus chromosome/scaffold name	%id. target Anopheles identical to query
AGAP004677	AGAP004677-RB	AALB007452	3R	59.8039
AGAP004677	AGAP004677-RA	AALB007452	3R	59.8039
AGAP004678	AGAP004678-RA	AALB007454	3R	43.314
AGAP004679	AGAP004679-RB	AALB007455	3R	48.8679
AGAP004679	AGAP004679-RA	AALB007455	3R	48.8679

<sup>4</sup> 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?

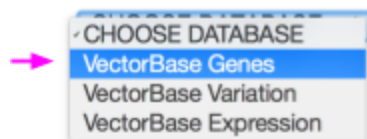
BioMart queries are run against a single species database, but if you click on the lower “Dataset” link, queries can “join” a second dataset. The results of the second dataset are connected and depend on the first one.

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

<p>Anopheles albimanus chromosome/scaffold name %id. target Anopheles albimanus gene identical to query gene %id. query gene identical to target Anopheles albimanus gene Anopheles albimanus orthology confidence [0 low, 1 high]</p> <p><b>Dataset 9080398 / 9080398 SNPs</b></p> <p>Anopheles gambiae Short Variants (SNPs and indels excluding flagged variants) (AgamP4)</p> <p><b>Filters</b></p> <p>[None selected]</p> <p><b>Attributes</b></p> <p>Variant name Variant source Chromosome/scaffold name Chromosome/scaffold position start (bp) Chromosome/scaffold position end (bp)</p>	Gene stable ID	Transcript stable ID	Anopheles albimanus gene stable ID	Anopheles albimanus chromosome/scaffold name	%id. target Anopheles albimanus gene identical to query gene	%id. query gene identical to target Anopheles albimanus gene	Anopheles albimanus orthology confidence [0 low, 1 high]	Variant name	Variant source	Chromosome
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-291799	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-244262	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-289726	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-154111	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	rs3573627 dbSNP	dbSNP	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-238489	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-255195	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0	WTSI-Ag-GVP-0.1-SNP-2L-282353	WTSI-Ag-GVP-0.1	2L
	AGAP004678	AGAP004678-RA	AALB007454	3R	43.314	41.9718	0			

## 3. Questions and practice exercises

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

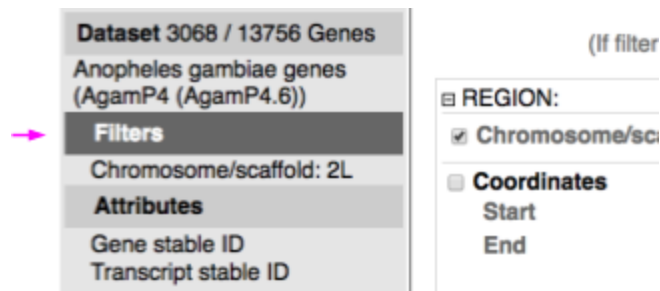


To find the number of results matching the filters you have provided, click on “Count” (see below) - to see the first few rows of the results and the attributes you have selected, click on “Results”.



### 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)



<b>Dataset</b> <input type="text"/> / 13763 Genes Anopheles gambiae genes (AgamP4)	
<b>Filters</b>	<b>REGION:</b>
Chromosome/scaffold: 2L	<input checked="" type="checkbox"/> Chromosome/scaffold
Gene type: protein_coding	
Start: 9000000	<input checked="" type="checkbox"/> <b>Coordinates</b>
End: 10000000	Start
<b>Attributes</b>	End
Gene stable ID	
Transcript stable ID	

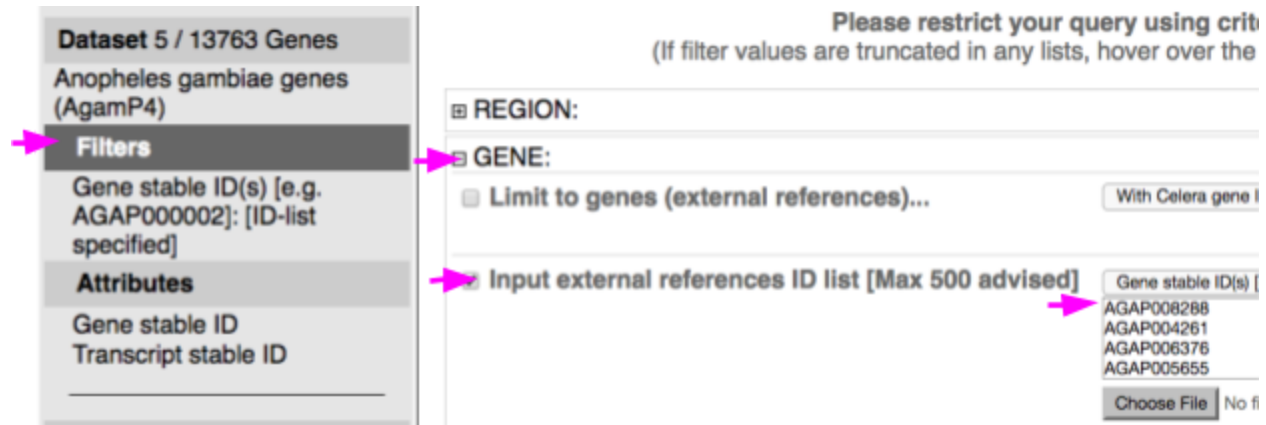
--

- 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:

(If filter values are truncated in any lists, hover over the text to see the full list)

**Dataset 10 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**

Gene Name(s) [e.g. CPF3]:  
[ID-list specified]

**Attributes**

Gene stable ID  
Transcript stable ID

**REGION:**

**GENE:**

☐ Limit to genes (external references)...

**Input external references ID list [Max 500 advised]**

Gene Name(s) [e.g. obp7, obp8, obp9, obp10]

- 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”

**Dataset 10 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**

Gene Name(s) [e.g. CPF3]:  
[ID-list specified]

**Attributes**

Gene stable ID  
Transcript stable ID

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

Dataset 10 / 13763 Genes  
Anopheles gambiae genes (AgamP4)

Filters  
Gene Name(s) [e.g. CPF3]: [ID-list specified]

Attributes  
Gene stable ID  
Transcript stable ID  
Source of gene name  
Gene description

Please select columns to be included in the output and hit 'Results'

Features  
Structures  
Homologues

Variant (Germline)  
Sequences

GENE:  
Ensembl  
External synonym  
Gene stable ID  
Transcript stable ID  
Protein stable ID  
Exon stable ID  
Gene description  
Chromosome/scaffold name

Transcript length (incl  
Gene name  
Source of gene name  
Transcript count  
% GC content  
Gene type  
Transcript type

- 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.	

<b>Dataset</b> Anopheles gambiae genes (AgamP4)	<b>Features</b> <input checked="" type="radio"/> Features <input type="radio"/> Structures <input type="radio"/> Homologues
<b>Filters</b> Gene Name(s) [e.g. CPF3]: [ID-list specified] With GO ID(s): Only	<b>Variation</b> <input type="radio"/> Variar <input type="radio"/> Sequen
<b>Attributes</b> Gene stable ID Transcript stable ID Source of gene name Gene description GO term accession GO term name	<b>GENE:</b> <b>Ensembl</b> <input type="checkbox"/> External synonym <input checked="" type="checkbox"/> Gene stable ID <input checked="" type="checkbox"/> Transcript stable ID <input type="checkbox"/> Protein stable ID <input type="checkbox"/> Exon stable ID <input checked="" type="checkbox"/> Gene description <input type="checkbox"/> Chromosome/scaffo <input type="checkbox"/> Gene start (bp) <input type="checkbox"/> Gene end (bp) <input type="checkbox"/> Strand <input type="checkbox"/> Karyotype band <input type="checkbox"/> Transcript start (bp) <input type="checkbox"/> Transcript end (bp) <input type="checkbox"/> Transcription start s
<b>Dataset</b> [None Selected]	<b>Phenotype</b> <input type="checkbox"/> Phenotype descripti <input type="checkbox"/> Source name <input type="checkbox"/> Study external refere
	<b>EXTERNAL:</b> <b>GO</b> <input checked="" type="checkbox"/> GO term accession <input checked="" type="checkbox"/> GO term name <input type="checkbox"/> GO term definition

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

<b>Dataset</b> Anopheles gambiae genes (AgamP4)	<b>Please restrict your query using cr</b> (If filter values are truncated in any lists, hover over th
<b>Filters</b> Gene Name(s) [e.g. CPF3]: [ID-list specified] With GO ID(s): Only	<b>REGION:</b> <b>GENE:</b> <input checked="" type="checkbox"/> Limit to genes (external references)... <div> <input checked="" type="radio"/> Only  <input type="radio"/> Excluded         </div>

- 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.	

**Dataset 10 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**  
Gene Name(s) [e.g. CPF3]:  
[ID-list specified]  
With GO ID(s): Only

**Attributes**  
Gene stable ID  
Transcript stable ID  
Source of gene name  
Gene description  
GO term accession  
GO term name  
Sequence **Publications ID**

☐ EXTERNAL:

**GO**  
☒ GO term accession  
☒ GO term name  
☐ GO term definition  
☐ GO term evidence code  
☐ GO domain

**GOSlim GOA**  
☐ GOSlim GOA Accession(s)  
☐ GOSlim GOA Description

**External References (max 3)**  
☐ Celera gene ID  
☐ Celera peptide ID  
☐ Celera transcript ID  
☐ ChEMBL ID  
☐ European Nucleotide Archive ID  
☐ European Nucleotide Archive ID  
☐ GO ID  
☐ GOSlim GOA ID  
☐ ImmunoDB ID  
☐ NCBI gene ID  
☐ PDB ID  
☐ RefSeq DNA ID  
☐ RefSeq peptide ID  
☐ RFAM ID  
☐ Ribosomal Protein Gene DB ID  
☒ Sequence **Publications ID**  
☐ STRING ID  
☐ tRNAScan-SE ID

- Query these 10 proteins Pfam domains

☒ Features
 ☐ Variant (Germline)
 ☐ Structures
 ☐ Sequences
 ☐ Homologues

---

☐ GENE:

---

☐ EXTERNAL:

---

☐ PROTEIN DOMAINS AND FAMILIES:

---

**Domains**

<input type="checkbox"/> CDD ID	<input type="checkbox"/> MS head cast ID
<input type="checkbox"/> CDD start	<input type="checkbox"/> MS head cast start
<input type="checkbox"/> CDD end	<input type="checkbox"/> MS head cast end
<input type="checkbox"/> MS larva ID	<input type="checkbox"/> MS pupa ID
<input type="checkbox"/> MS larva start	<input type="checkbox"/> MS pupa start
<input type="checkbox"/> MS larva end	<input type="checkbox"/> MS pupa end
<input type="checkbox"/> MS head ID	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS head start	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS head end	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS male reproductive ID	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS male reproductive start	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS male reproductive end	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS malpighian tubule ID	<input type="checkbox"/> MS salivary gland
<input type="checkbox"/> MS malpighian tubule start	<input type="checkbox"/> PANTHER ID
<input type="checkbox"/> MS malpighian tubule end	<input type="checkbox"/> PANTHER start
<input type="checkbox"/> MS midgut ID	<input type="checkbox"/> PANTHER end
<input type="checkbox"/> MS midgut start	<input checked="" type="checkbox"/> Pfam ID
	<input type="checkbox"/> Pfam start

- 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?



**Dataset 3069 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**  
Chromosome/scaffold: 2L

**Attributes**  
Gene stable ID  
Transcript stable ID  
Aedes aegypti orthology confidence [0 low, 1 high]  
Aedes albopictus orthology confidence [0 low, 1 high]

**Dataset**  
[None Selected]

**Features** **Variant (Germline)**  
**Structures** **Sequences**  
**Homologues**

**GENE:**

**ORTHOLOGUES (Max select 6 orthologues):**

**Aedes aegypti Orthologues**

<input type="checkbox"/> Aedes aegypti gene stable ID	<input type="checkbox"/> Aedes aegypti homology type
<input type="checkbox"/> Aedes aegypti gene name	<input type="checkbox"/> %id. target Aedes aegypti gene identical to query gene
<input type="checkbox"/> Aedes aegypti protein or transcript stable ID	<input type="checkbox"/> %id. query gene identical to target Aedes aegypti gene
<input type="checkbox"/> Aedes aegypti chromosome/scaffold name	<input type="checkbox"/> Aedes aegypti Gene-order conservation score
<input type="checkbox"/> Aedes aegypti chromosome/scaffold start (bp)	<input type="checkbox"/> Aedes aegypti Whole-genome alignment coverage
<input type="checkbox"/> Aedes aegypti chromosome/scaffold end (bp)	<input type="checkbox"/> dN with Aedes aegypti
<input type="checkbox"/> Query protein or transcript ID	<input type="checkbox"/> dS with Aedes aegypti
<input type="checkbox"/> Last common ancestor with Aedes aegypti	<input checked="" type="checkbox"/> Aedes aegypti orthology confidence [0 low, 1 high]

### 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”)

**Dataset 1066 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**  
Chromosome/scaffold: X  
Gene type: protein\_coding

**Attributes**  
Peptide  
Gene stable ID  
Transcript stable ID

**Please select columns to be included**

**Features** **Variant (Germline)**  
**Structures** **Sequences**  
**Homologues**

**SEQUENCES:**

**Sequences (max 1)**

Diagram showing two gene models with exons represented by red boxes and introns by lines. The first model has three exons, and the second model has four exons.

- Look at the various options for download

<input type="radio"/> Unspliced (Transcript) <input type="radio"/> Unspliced (Gene) <input type="radio"/> Flank (Transcript) <input type="radio"/> Flank (Gene) <input type="radio"/> Flank-coding region (Transcript) <input type="radio"/> Flank-coding region (Gene)	<input type="radio"/> 5' UTR <input type="radio"/> 3' UTR <input type="radio"/> Exon sequences <input type="radio"/> cDNA sequences <input type="radio"/> Coding sequence <input checked="" type="radio"/> Peptide
<b>Upstream flank</b> <input type="checkbox"/> Upstream flank	
<b>Downstream flank</b> <input type="checkbox"/> Downstream flank	

- 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.

**Dataset 3170 / 13763 Genes**  
 Anopheles gambiae genes (AgamP4)

**Filters**

Gene type: protein\_coding  
 Chromosome/scaffold: 2R  
 Orthologous Aedes aegypti  
 Genes: Only

**Attributes**

Gene stable ID  
 Transcript stable ID

**Please restrict your query using criteria below**  
 (If filter values are truncated in any lists, hover over the list item)

☐ **REGION:**

☐ **GENE:**

☐ **PHENOTYPE:**

☐ **GENE ONTOLOGY:**

☐ **MULTI SPECIES COMPARISONS:**

☒ **Homologue filters**

☒ Orthologous Aedes aegypti  
☐ Only  
☐ Excluded

- Can you add further species?

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

The screenshot shows the VectorBase BioMart interface. On the left, the 'Dataset 1 / 13763 Genes' section is active, displaying 'Anopheles gambiae genes (AgamP4)'. Under the 'Filters' section, 'Gene stable ID(s) [e.g. AGAP000002]: [ID-list specified]' is entered. Under the 'Attributes' section, 'Gene stable ID', 'Transcript stable ID', and 'cDNA sequences' are selected. On the right, the 'Export all results to' dropdown is set to 'File', and the 'Go' button is highlighted. The 'Email notification to' field is empty. The 'View' section shows '10 rows as FASTA'. The main content area displays the cDNA sequence for AGAP004707, with the transcript identifier 'AGAP004707-RA' highlighted in a pink box.

**Dataset 1 / 13763 Genes**  
Anopheles gambiae genes (AgamP4)

**Filters**  
Gene stable ID(s) [e.g. AGAP000002]: [ID-list specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
cDNA sequences

Export all results to **File** **Go**

Email notification to

View 10 rows as FASTA

>AGAP004707 | AGAP004707-RA  
ATGACCGAAGACTCCGATTCGATATCTGAGGAAGAACC  
CGTGAATCATTACAAGCTATCGAAGCACGCATTGCAG  
TTGGAAAGAAAACGAGCTGAGGGGGAGATACGCTACG  
CCCCAACCGGACCCTACTCTTGAACAGGGTGTACCAG  
TTCCCCCGGAGTTGGCCTCCACGCCTCTCGAGGATA  
AGGACATTCGTAGTGATTAGTAAAGGAAAAGATATAT

- 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](mailto: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.