Finding and Exporting Data

Not sure what tool to use to find and export data? VectorBase Search is used to find data for simple queries for one or few genes. In contrast, Advanced Search and BioMart are used to retrieve data for complex queries, involving a bigger number of genes or complete genomes.

Search

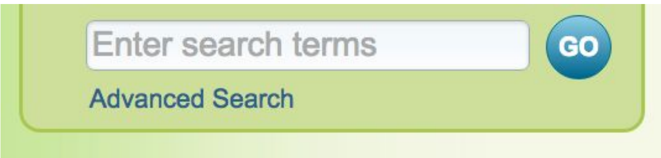
Contents

1. BLAST basics

Note that VectorBase Search allows the use of:

- *AND, OR & NOT (boolean operators)*
- *+ & -*
- *wild card (asterisk), exact searches (quotation marks) & grouped searches (parenthesis).*

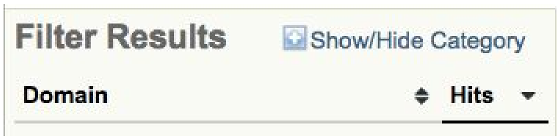
Find the search box. It's in the top right corner of every VectorBase web page.



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

Search for **actin**

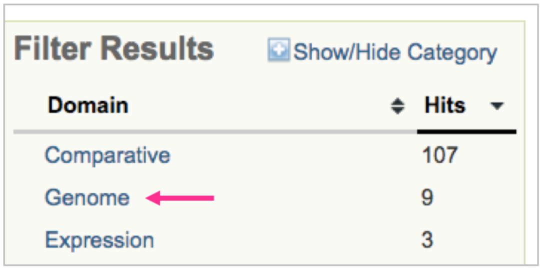
The **Filter Results** section summarizes how many search results were found in each major data section (**Domain**) and species within VectorBase. Click on the **headings** to sort alphabetically or by the number of hits. Click on the **plus sign** to show or hide categories (available options depend on which Domain you have selected).



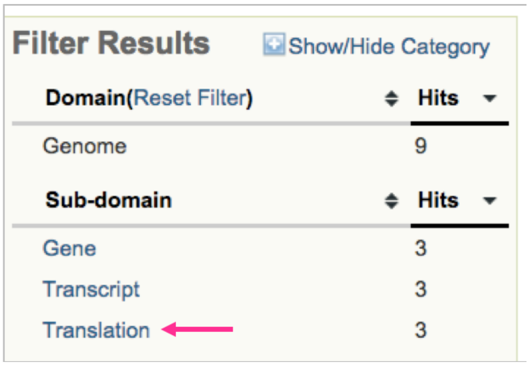
The gene metadata, **symbol** (or name) *e.g.*, srp7 or rpl2 and **description** (or function) *e.g.*, serpin 7 or ribosomal protein L2, can be used to search for genes. Note that this type of search has some caveats:

Finding genes based on their functional description or symbol is simple and intuitive. It should work well so long as the metadata exists in the database. Unfortunately this is not always the case and searches may give you no or incomplete results because not every gene has metadata associated with it. Other search strategies to find genes will be discussed later in this document.

Search for AALB002800, AALB002801, AALB002802**,** these are three **gene IDs** for *Anopheles albimanus***.** With a click filter with Genome (Domain) and Translation (Sub-Domain).



Domain	Hits
Comparative	107
Genome	9
Expression	3

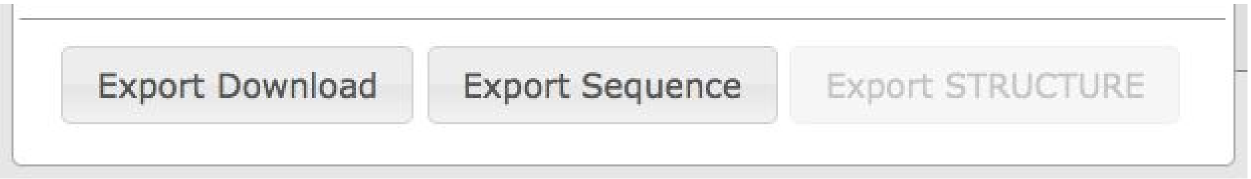


Domain(Reset Filter)	Hits
Genome	9

Sub-domain	Hits
Gene	3
Transcript	3
Translation	3

Click on the button Export.

For this especifi query, two of its options are activated, Download and Sequence. Try both.



The Export options are automatically activated/deactivated under the following criteria:

- Single Domain and single Sub-domain should be selected.
- Export **Download**: It creates a file that includes information such as gene ID, symbol, species, strain, gene biotype, description, protein domain, GO terms or location (supercontig or chromosomes and base pair range); the columns headers depends on the sub-domain selected. The downloaded file is in CSV format which can be loaded in a spreadsheet program like Excel.
- Export **Sequence**: Is activated when the query includes two or more sequences and the Genome (Gene, Transcript or Translation subdomain), Transcriptome or Proteome filters are selected. Depending on the query, it creates a file with nucleotide or amino acid sequence in FASTA format that can be open with a text editor software like Notepad (Windows) or TextWrangler (Mac). Activated with Domains/Sub-domains: Genome/Translation, Genome/Transcript, Genome/Gene, Proteome, Transcriptome.
- Export **STRUCTURE**: Works only with Population Biology/Sample genotype, with Genotype class = microsatellite
- Login users: a) up to 300,000 results *on the fly*; b) *background job* over 300,000 results. Anonymous users: only up to 5,000 results *on the fly*. Please login or join VectorBase to get more results.

3. Questions and practice exercises

Single choice:
Unless otherwise stated, a single correct answers is possible.

1. Finding a gene of interest

Imagine you are reading a journal article and you wish to know more about the gene(s) mentioned in it. Here is an extract from a paper:

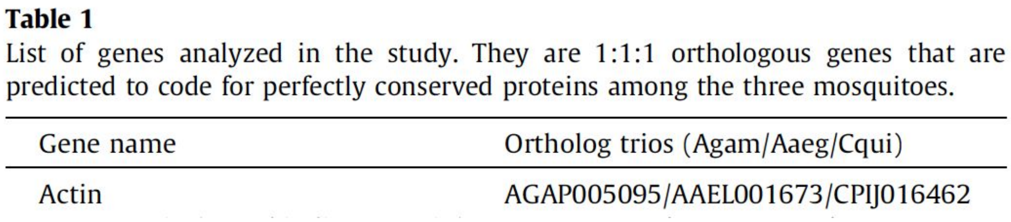


Table 1
List of genes analyzed in the study. They are 1:1:1 orthologous genes that are predicted to code for perfectly conserved proteins among the three mosquitoes.

Gene name	Ortholog trios (Agam/Aaeg/Cqui)
Actin	AGAP005095/AAEL001673/CPIJ016462

Search for **actin**

Question 1.1

Which data domain has the highest number of hits for a search for 'actin'?

	Answer
Comparative	
Transcriptome	
Population Biology	

Question 1.2

Now search using a single asterisk character * This is the wildcard search - it will find everything in VectorBase. How many different data domains are there in VectorBase?

	Answer
9	
10	
21,289,169	

Question 1.3

When you click on a domain or species filter, the search is repeated with the filter condition/restriction added to your query.

Search again for **actin** and, using the filters, answer this:

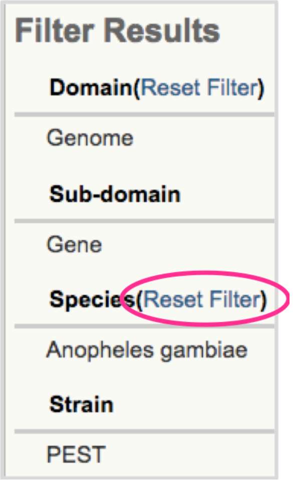
How many *Anopheles gambiae* Genome/Gene hits are there in VectorBase?

	Answer
3,441	
31	
787	

Question 1.4

Practice the following:

- Adding and remove filters with the "Reset Filter" links



Filter Results

Domain(Reset Filter)

Genome

Sub-domain

Gene

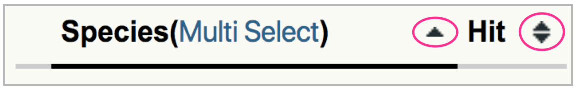
Species(Reset Filter)

Anopheles gambiae

Strain

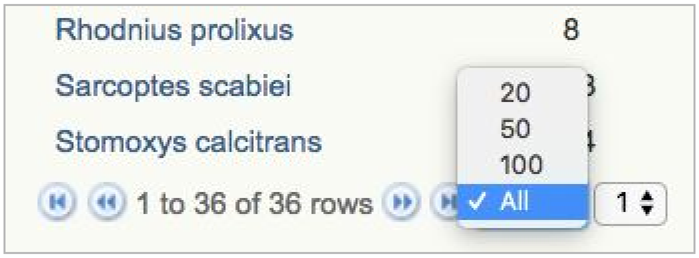
PEST

- Sort species alphabetically and by number of hits, with the arrow heads



Species(Multi Select) ▲	**Hit** ⬍

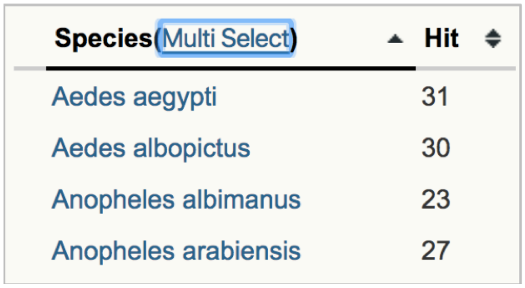
- Show partial and complete list of species



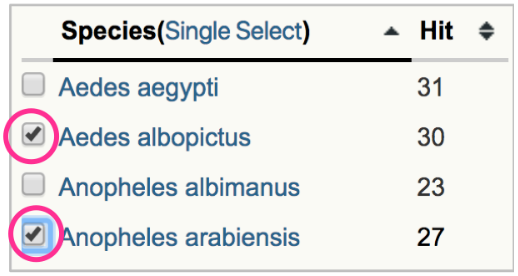
Rhodnius prolixus	8
Sarcoptes scabiei	3
Stomoxys calcitrans	4

1 to 36 of 36 rows

- Select single or multiple species



Species(Multi Select) ▲	**Hit** ⬍
Aedes aegypti	31
Aedes albopictus	30
Anopheles albimanus	23
Anopheles arabiensis	27



Species(Single Select) ▲	**Hit** ⬍
□ Aedes aegypti	31
☑ Aedes albopictus	30
□ Anopheles albimanus	23
☑ Anopheles arabiensis	27

How many species have *more* Genome/Gene hits than *Aedes albopictus* for the query **actin**?

Multiple choice:

Unless otherwise stated, multiple correct or true answers are possible.

Question 1.5

Filter for just the Genome domain hits from *Aedes aegypti*

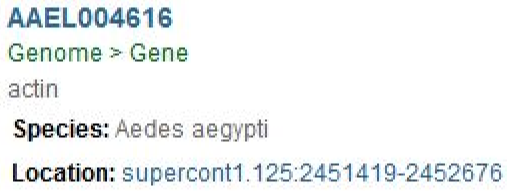
Use next/last page & first/previous controls to inspect the list, and read some of the gene descriptions below.

Which of the following statements are true (there may be more than one)?

	True	False
Results for genes such as 'attractin' and 'gliotactin' are listed because they end in 'actin'.		
Four genes have **symbols**, i.e., 'Act-4, Arp5, Arp8 and Act1', all other genes have VectorBase gene ID only (which start with AAEL).		
Some non-actin genes, such as one described as 'actin-binding' are shown because their **descriptions** contain the word actin.		
Only true actin genes are shown.		
All actin genes are located on chromosome 3		

Question 1.6

What happens when you click on the main dark-blue links, e.g. AAEL004616 gene ID in the image below?



AAEL004616

Genome > Gene

actin

Species: Aedes aegypti

Location: supercont1.125:2451419-2452676

	True	False
A search is performed for that gene ID.		

A filter is applied to restrict results to that gene ID.		
An error of type 5 occurs.		
The gene's 'home page' in the genome browser is opened.		

Question 1.7

Go back again to the Search results page. Now try clicking on the blue link next to Location: (see image in previous question). Where does this take you?

	True	False
The gene's 'home page'		
The graphical genomic region/location browser		
The gene tree page		
The genomic variation page		

Question 1.8

Still using the search for **actin** in the Genome domain, answer the following. How are VectorBase gene IDs for *Anopheles farauti* constructed?

	Answer
The letters AFARAUTI followed by some digits	
The letters AFAF followed by 6 digits	
The letters AFAF followed by 7 digits	

Match these pairs:

Species	**VectorBase gene IDs**
(a) *Glossina morsitans* Yale	( ) PHUM*
(b) *Aedes albopictus* Foshan	( ) BGLB*
(c) *Lutzomyia longipalpis* Jacobina	( ) AALB*
(d) *Anopheles albimanus* STECLA	( ) GMOY*

(e) *Pediculus humanus* USDA	( ) AALF*
(f) *Biomphalaria glabrata* BB02	( ) LLOJ*

Question 1.9

Now make a fresh new search[1] with the VectorBase gene ID **AGAP005095**.
Note how many hits there are for the Genome domain before you click on the Genome domain filter. When you click on the Genome domain filter, what happens, and why?

	True	False
There are three hits in that domain but they all belong to the same gene. Each takes me to a different page in the genome browser, the gene, transcript and protein pages.		
The genome domain filter always redirects the user to the genome browser.		
There is no logic to describe the behaviour of links in VectorBase search.		

Question 1.10

Given your experience with search in the previous questions, which of the following statements are true?

	True	False
To find a single gene of interest, it is best to use the VectorBase stable ID if it is provided in the paper.		
Keyword-based searches are likely to find multiple genes, some of which will be relevant, others not.		
VectorBase search filters allow you to refine your search without typing.		

[1] When you use the top-right search box, a new search with no filters is always performed for you.

Optional exercises

- Find the VectorBase stable gene ID for the gene in this figure legend

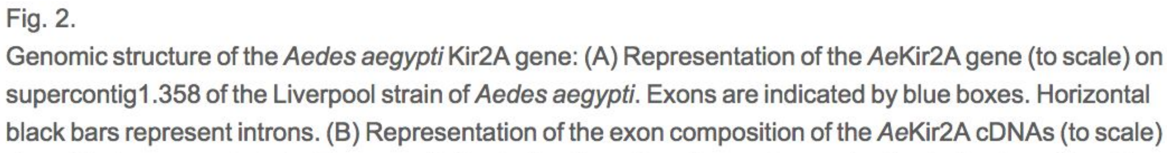


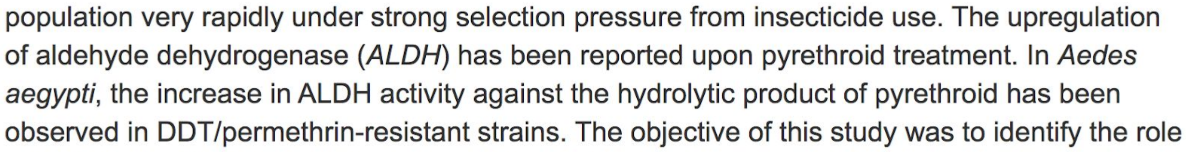
Fig. 2.
Genomic structure of the *Aedes aegypti* Kir2A gene: (A) Representation of the *Ae*Kir2A gene (to scale) on supercontig1.358 of the Liverpool strain of *Aedes aegypti*. Exons are indicated by blue boxes. Horizontal black bars represent introns. (B) Representation of the exon composition of the *Ae*Kir2A cDNAs (to scale)

- How many 'cecropin' genes can you find in the different *Anopheles* species using keyword search? **Hint**: use a boolean operator

- Do you think *Anopheles melas* really only has one cecropin gene? Think about what might explain this...

2. Gene symbol, description and quotation marks:

Here is another extract from a published paper:



population very rapidly under strong selection pressure from insecticide use. The upregulation of aldehyde dehydrogenase (*ALDH*) has been reported upon pyrethroid treatment. In *Aedes aegypti*, the increase in ALDH activity against the hydrolytic product of pyrethroid has been observed in DDT/permethrin-resistant strains. The objective of this study was to identify the role

Perform two searches in two browser windows or tabs for **aldehyde dehydrogenase** (without quotes) and **"aldehyde dehydrogenase"** (with double quotes[2]) **within the Genome domain** (click the Genome domain filter).

[2] Different results can be obtained with single quotes.

Question 2.1

Compare the results obtained with the two searches. Do you notice a different sub-domain within the Genome domain?

	Answer
No	
Yes, there's a new one called "Transcript"	
Yes, there's a different one called "Mitochondrial genome"	
Yes, there's a new one called "Gene expression"	

Question 2.2

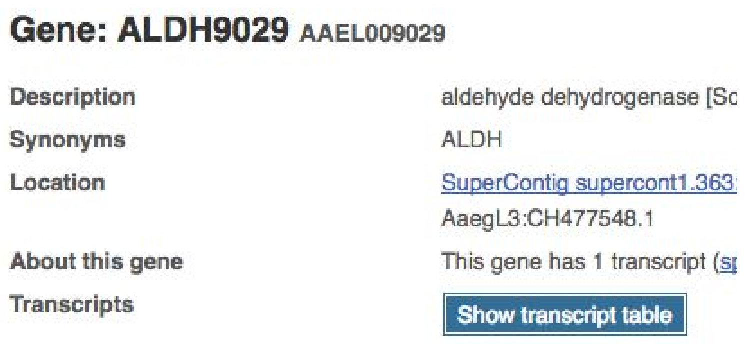
Which one of the following statements are true?

	True	False
The quoted query has more hits for *Aedes aegypti* than the non-quoted query because the quoted query is more specific.		
The non-quoted query has more hits for *Aedes aegypti* than the quoted query because the non-quoted query finds genes with either the words aldehyde OR dehydrogenase		
With the non-quoted query, the results containing both words are listed first		
Every gene described as alcohol dehydrogenase has a symbol.		

Question 2.3

Continue using only the quoted query: "aldehyde dehydrogenase" within Genome > Gene and *Aedes aegypti*. Make a note of the number of hits

Click on hit AAEL009029. In the new page that open click "Show transcript table".



Gene: ALDH9029 AAEL009029

Description	aldehyde dehydrogenase [Sc
Synonyms	ALDH
Location	SuperContig supercont1.363: AaegL3:CH477548.1
About this gene	This gene has 1 transcript (sp
Transcripts	Show transcript table

Look at the transcript table and provide the information required:

	Answer
One column links out to VectorBase protein summary page, which one is?	
What is the protein length in amino acids?	

Go back to the gene page and look for the following information:

	Answer
Synonyms	
Number of homologous genes	( ) orthologues and ( ) paralogues
Number of splice variants	

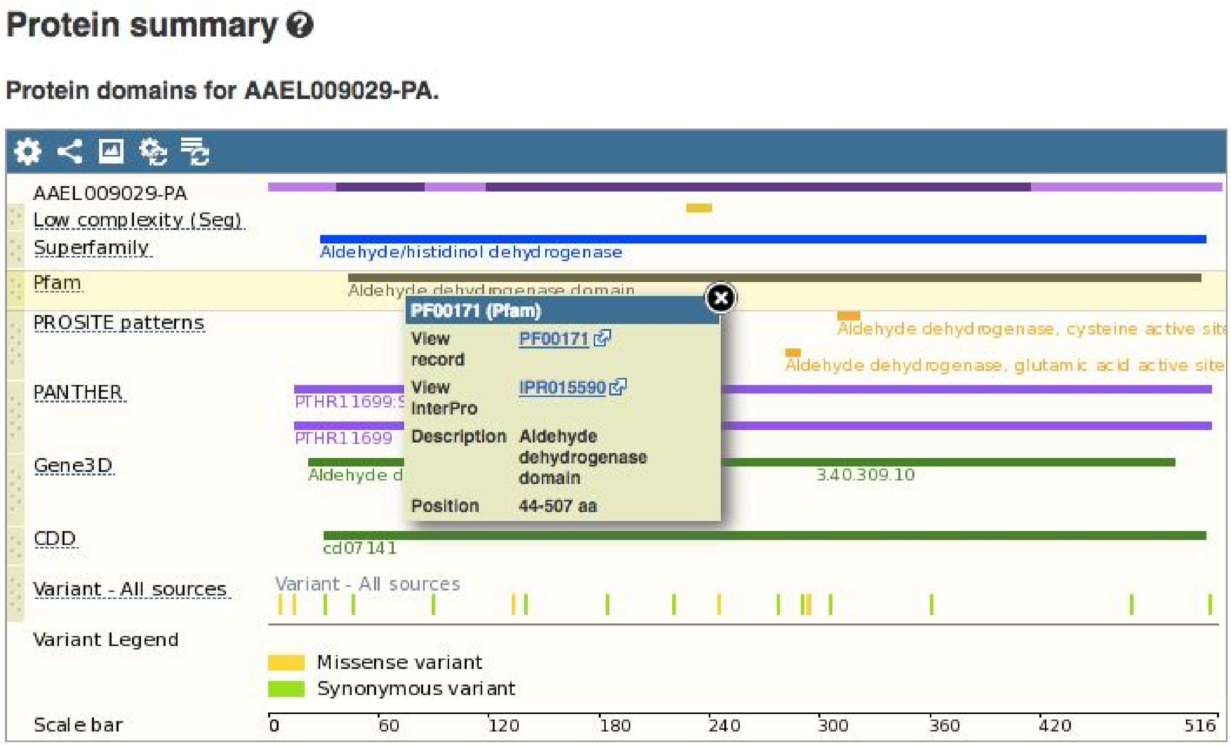
Question 2.4

Go back to the protein page. Take a good look at the Protein Summary diagram. What is the main thing it shows?

	Answer
Differentially spliced transcript isoforms	
Homologous proteins in VectorBase species	
2D electrophoresis spot identifications	
Homology matches to conserved protein domains	

Question 2.5

Click on the grey bar representing the Pfam domain. You should see a pop-out info box as shown below. You can reposition the box by dragging the title bar and close it when no longer needed.



Protein summary

Protein domains for AAEL009029-PA.

PF00171 (Pfam)

View record: PF00171

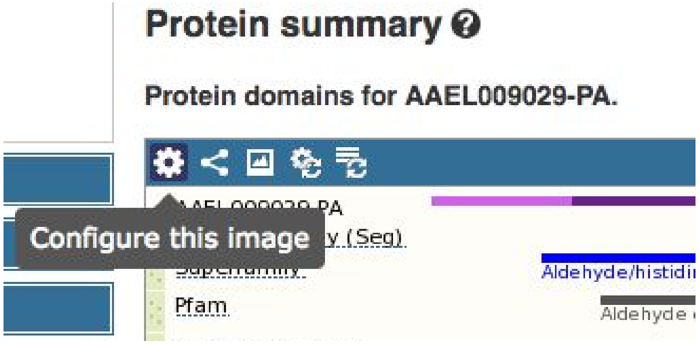
View InterPro: IPR015590

Description: Aldehyde dehydrogenase domain

Position: 44-507 aa

In the pop-out click on both of the links to open new web browser tabs for each.

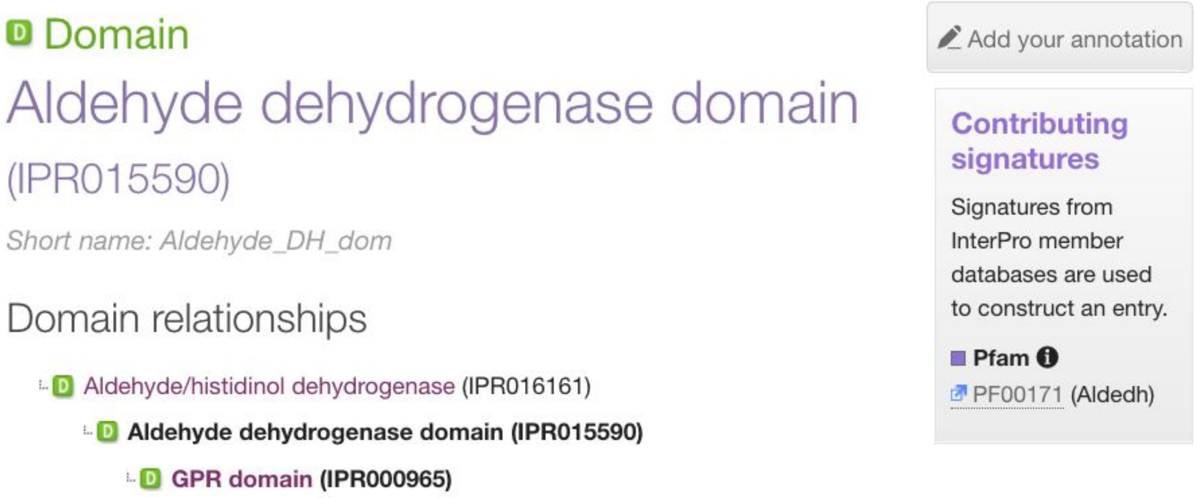
Pfam is a database of conserved domains curated by humans. Superfamily, Gene3D are two other similar databases, and there are several more. Each database has a different biological focus and/or technical approach. If you like, you can configure the protein domain diagram to show only the domain databases you want to see (by clicking on the gear icon -”configure this image”).



Protein summary

Protein domains for AAEL009029-PA.

Interpro (https://www.ebi.ac.uk/interpro/) is an aggregated database of conserved domain databases. It is like a "one-stop shop for all your conserved domain database needs". Note that the Interpro page for the **aldehyde dehydrogenase domain** has a section top-right which attributes the source database(s), in this case, Pfam:



D Domain

Aldehyde dehydrogenase domain (IPR015590)

Short name: Aldehyde_DH_dom

Domain relationships

- D Aldehyde/histidinol dehydrogenase (IPR016161)
  - D Aldehyde dehydrogenase domain (IPR015590)
    - D GPR domain (IPR000965)

Add your annotation

Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

Pfam

PF00171 (Aldedh)

Every protein from VectorBase's annotated genomes is automatically scanned against all InterPro domains (i.e. Pfam, Superfamily, Gene3D, etc) using advanced homology detection algorithms such as Hidden Markov Models. This information is also available in VectorBase search, so you can search with IDs from InterPro, Pfam and the other domain databases.

Search VectorBase with either PF00171 or IPR015590 and filter into Genome > Gene and *Aedes aegypti*. What is the number of hits obtained with each?

Compare the the double quoted "aldehyde dehydrogenase" (8) and the InterPro IPR015590 (31) hit results for *Aedes aegypti*. Why do you think there is a difference in the number of results obtained with both queries?

	Answer
Domain databases contain evolutionary information.	
Automated domain assignment via homology methods has broader coverage than the human annotation of gene descriptions.	
Protein similarities are a useful method for annotating genes in newly sequenced genomes.	
Protein 3D structure is conserved more than sequence.	

Question 2.6

For the InterPro query: Reset the species filter and select the Gene sub-domain. Examine the number of hits for the different species.

The *Anopheles* species mostly have 10 hits for the aldehyde dehydrogenase domain. *Anopheles stephensi* has 20. Explore the results and complete the table below. **Hint**: Click on the species or hit headings (or arrow heads) to sort alphabetically or by the number of hits.

	True	False
Two *Anopheles stephensi* different strains have been sequenced and assembled, but each strain has 10 predicted aldehyde dehydrogenases.		
A recent whole genome duplication has doubled the number of genes.		
The specialised feeding behaviour of *Anopheles stephensi* is suspected to have driven a dramatic expansion in this gene family.		

Note: *A. sinsensis* has a similar issue.

Question 2.7

Which of these are genuinely plausible explanations for the smaller number of hits seen in *A. atroparvus* and *A. epiroticus* (9)?

	True	False
Genome sequencing and assembly issues resulting in gaps or errors in the genome sequence.		

More gene(s) were deleted since that species diverged from the last common ancestor than were created by gene duplication.		
Extreme sequence evolution has prevented the conserved domain models from recognising the sequence		

Bonus exercises:

- Discuss with your neighbour or a instructor why some *Anopheles* species have more than 10 aldehyde dehydrogenase hits.
- Go to InterPro for a protein domain you know or work on. Search VectorBase and critically interpret the results.
- Just like with Pfam and InterPro, the IDs from other databases can also be used in VectorBase:

Queries	**Database**	**Answers**
PF03028 What is the gene description (or function) of this protein domain in *Aedes aegypti*?	Pfam	dynein chain
PMID:18983277 In *An. gambiae* which gene is mentioned in this publication? Provide the VectorBase ID		
GO:0004930 What is the name of this term in the GO ontology?		
KEGG:aga00010 Which *Anopheles* species has genes in this pathway? Select its top hit. In the gene (browser) page, in the left hand side menu select 'External references'. What are the pathway descriptions (or database identifiers)?		

GO:0016620 In addition to the Ontology and Genome domains, from which other category this query provides hits		

3. Asterisk or the wild card *:

- A search for * by itself matches all field values. Type the wild character or asterisk in the Search box. How many hits you obtained as a result?

- Click on the Expression Domain. How many hits are for the sub-domain Experiment?

- Click on the Experiment sub-domain. How many experiments are per species?

Organism	**Number of experiments**
An. gambiae	
Ae. aegypti	

The asterisk (*) acts as a wildcard placeholder for any number of characters. A search for Tryp* finds any field value beginning with Tryp-. A search for *psin finds any field value ending with -psin. A search for *ryps* finds any field value that contains the -ryps-substring.

- Try these queries for *An. gambiae*, how many Genome > Gene hits you find for each?

tryp* AND *Anopheles gambiae*	*psin AND *Anopheles gambiae*	*ryps* AND *Anopheles gambiae*

4. Boolean logic: AND and OR

Try the following queries.

- OBP3 OR OBP4
- OBP3 AND OBP4

Do you obtain different results? Why?

Question 4.1

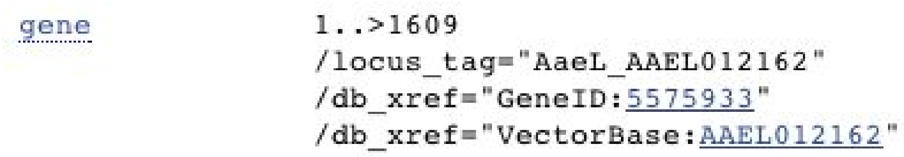
- Combine boolean logic and the asterisk to find out the number of GPROP gene**s** in *Culex quinquefasciatus*

The query you used	Number of results

5. GenBank accessions:

- When you find a gene of interest in GenBank (www.ncbi.nlm.nih.gov/genbank/), take note of its accession, *e.g.*, XM_001655875.2

- Scroll to the “gene section”, there you will find a reference to VectorBase:



```
gene            1..>1609
                /locus_tag="AaeL_AAEL012162"
                /db_xref="GeneID:5575933"
                /db_xref="VectorBase:AAEL012162"
```

- What is the VectorBase gene ID:

- What happens when you click on VectorBase gene ID?

- **Alternatively**, if you already know the GenBank accession you can use it as keyword in VectorBase Search. What is the Genome > Gene top hit for XM_001655875.1 ?
  Hint: is a VectorBase gene ID

6. Try the following queries:

Queries	**Answers**
OBP5 This gene symbol or description is assigned to how many species in the Genome > Gene category?	
GPROP3 In how many species can you find this gene symbol?	
Kinase What are the organisms with the highest and lowest number of kinase genes?	
Rhodnius prolixus Is this organism in any population biology project/experiment? Give its ID.	

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.