

# VectorBase Hands-on Workshop

answer key (only sample use cases and questions in a separate file)



**VectorBase**

Bioinformatics Resource for  
Invertebrate Vectors of Human Pathogens

VectorBase has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases and National Institutes of Health, NIAID/NIH

Updated on May 2019

# Instructor

Gloria I. Giraldo-Calderón, PhD

VectorBase Scientific Liaison/Outreach Manager

Please remember that the answers to the exercises may change because VectorBase data, tools and resources are updated every two months. For details about these changes visit:

<https://www.vectorbase.org/releases>

## Table of Contents

<b>Lecture Notes</b>	<b>4</b>
<b>Case study #1: Odorant receptor (OR) genes</b>	<b>5</b>
Search	5
Critical thinking: Gene metadata submissions	6
Genome Browser > Gene tab	7
Genome Browser > Location tab > RNAseq data	9
Critical thinking: external IDs for gene Search	12
Expression Browser > Expression Report	13
Critical thinking: Why do we need a genome browser?	15
<b>Case study #2: Opsin (GPRop) genes</b>	<b>15</b>
BLAST	15
Galaxy > BLAST	16
Genome Browser > Gene Tab > Comparative Genomics	19
ClustalW > HMMER	19
Critical thinking: BLAST vs Hmmer vs Comparative Genomics	21
Apollo	22
Critical thinking: Gene manual annotation submissions	22
<b>Case study #3: Transcript Expression</b>	<b>23</b>
Advanced Search > Expression	23
Expression Browser	25
Critical thinking: in vivo differential expression data	26
<b>Case study #4: Insecticide resistance from Ethiopia</b>	<b>26</b>
Population Biology > Insecticide Resistance	26
Critical thinking: open access data	33
<b>Case study #5: Variation data of the voltage-gated sodium channel gene</b>	<b>33</b>
Genome Browser > Variation tab	34
Find data display > Variations	39
Critical Thinking: VectorBase data tools and resources	40

## Lecture Notes

## Case study #1: Odorant receptor (OR) genes

- 5 min video: <https://youtu.be/kj9HYBD-knQ>
- Kelly, M., Su, C.-Y., Schaber, C., Crowley, J. R., Hsu, F.-F., Carlson, J. R., & Odom, A. R. (2015). Malaria Parasites Produce Volatile Mosquito Attractants. *mBio*, 6(2), e00235–15. <http://doi.org/10.1128/mBio.00235-15>

1. Which *Anopheles* genes are mentioned in the paper Kelly et al. 2015? **Hint:** go to the section called ‘*Anopheles* odorant receptors respond to malaria parasite-produced terpenes’

Gene Metadata	
Symbol or Name e.g., srp7 or rpl2	Description or Function e.g., serpin 7 or ribosomal protein L2
AgOR	<i>Anopheles gambiae</i> odorant receptor
AgOR 21	“ 21
AgOR 50	“ 50
AgOR 75	“ 75

### Search

2. Go to VectorBase and use **Search** to find these genes

Query keyword(s)	Total number of results	Filter with*: Genome > Gene > <i>Anopheles gambiae</i>
odorant receptor	19.963	530
AgOR	One and it automatically opens. External > Publication: click on ‘Search VectorBase’	Or5 (AGAP011467) Or4 (AGAP011468) OR3 (AGAP011469) S7 (AGAP010592)
AgOR21 Ag OR21 Ag OR 21	9.680 1.026.392 1.483.035	NOT including genome domain 1 hit, Or21 (AGAP029499) 53 hits, including Or21
AgOR50 Ag OR50 Ag OR 50	19.520 1.036.337 1.303.980	NOT including genome domain 1 hit, Or50 (AGAP029705) 1.781 hits, includes Or50
AgOR75 Ag OR75 Ag OR 75	8.021 1.023.747 1.034.068	NOT including genome domain 1 hit, Or75 (AGAP002045) 8, includes Or75

\*the category headings and number of results can be sorted alphabetically and in decreasing/increasing order, respectively

3. Use 'odorant receptor' (without the quotation marks) as a keyword. Filter with Genome > Gene > *Anopheles gambiae*. Export the results. There are three formats; which ones are available for this specific query? Briefly describe the content of the available ones:

a) Download:

Available. A \*csv file, which can be opened in Excel. It has 12 columns. It Contains the VectorBase gene ID (or accession) and other data such as symbol, species, strain, biotype (e.g., protein coding), description, domain, GO and localization (chromosome, base pair range start and end)

b) Sequence<sup>1</sup>:

Available. The gene protein sequence in fasta format, with accession used as name.

c) STRUCTURE<sup>2</sup>:

Not available

4. Go back to the \*csv file

- This file only contains odorant receptor genes. True \_\_\_ False X
- Locate odorant receptors 21, 50 and 75
- Search them again in VectorBase but using the gene IDs. **Note**: VectorBase Search box receives up to 10 gene IDs simultaneously
- Briefly describe the behavior of this type of query and the obtained results

	Query keyword(s): <b>VectorBase ID</b>	Results <b>Are filters necessary?</b> e.g., Genome > Gene > <i>A. gambiae</i>
75	AGAP002045	These three genes are the the top hits in the results page, used individually or together (separated with spaces or commas)
21	AGAP029499	
50	AGAP029705	Once filters are applied, very few hits should appear (~< 10 or one or two) and the genes of interest stay at the top.

### Critical thinking: Gene metadata submissions

5. VectorBase does not have personnel and/or funding to do literature curation, to include gene metadata (names/symbols and descriptions/functions), it is up to the research authors to submit their finding to VectorBase. Consider this hypothetical case.

<sup>1</sup> From the Genome Browser (> Export data > Next) you can also download both protein and nucleotide gene sequences.

<sup>2</sup> Sample query for a STRUCTURE file: Population Biology > Sample genotype > *Aedes aegypti*

Research group 'a' works on insecticide resistance genes in one mosquito species. They assign metadata to ~100 genes and the data is documented in the corresponding paper text, tables and figures, published in year 2018. Without knowing about group 'a' research, group 'b' works on some of the same genes and also assigns metadata. Group 'b' submits their metadata to VectorBase, provide the VectorBase gene IDs in the paper and publish in year 2019. You are conducting a new research about the same genes and found the data in VectorBase and the two papers.

Which metadata you use in your paper? The ones from research group a \_\_\_\_ or group b \_\_\_\_  
Briefly justify your answer.

### Genome Browser > Gene tab

6. For these three genes, AGAP029499 AGAP029705 AGAP002045, go to the **Genome Browser > gene pages**.

- Look at 'Gene > Literature' (in the left hand menu)
- Look for the 'Synonyms' information (in the central part of the page)
- What information are these sections providing and how this relates to your query and results?

Literature: gene metadata might come from this paper. Here is a link to submit your papers to VectorBase: <https://www.vectorbase.org/content/submit-data> (submit gene data > PubMed links to genes)

Synonym: different scientist/community members have given the genes different metadata.

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

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

8. Perform two searches in two browser windows or tabs for **odorant receptor** (without quotes) and **"odorant receptor"** (with double quotes) filtering with Genome > Gene > *Anopheles gambiae*. **Note:** single quotes are not recognized.

	<b>odorant receptor</b>	<b>"odorant receptor"</b>
Number of results	530	76

9. Now search using a single asterisk character \* or with a click on 'Advanced Search'. This is the wildcard search - it will find everything in VectorBase. How many different Search domains are there in VectorBase?

- a) 20
- b) 11 **X**
- c) 44.816.571

### Non-odorant genes

10. Search for this gene: AGAP004707. From the genome browser provide the following information:

	Answer
Synonyms	kdr, VSC
Number of homologous genes	( 40 ) orthologues and ( 9 ) paralogues
Number of splice variants or transcripts	13

11. Try this query: GPRop\* (notice the asterisk at the end of the gene name/symbol). Filter with Genome > Gene domain. Do not filter for species.

- Is a specific query only for opsins and pteropsins True **X** False \_\_\_\_\_
- From all (41) VectorBase genomes, only seven species have opsins and/or pteropsins. True \_\_\_\_\_ False **X** Justify your answer

Only to these seven species a gene name/symbol has been assigned by members of the user community

- The number of opsins in each species is different True **X** False \_\_\_\_\_

12. Try this query: AGAP001374 and filter with Proteome. This filter allows you to look for peptide sequences derived from mass spectrometry (MS) experiments.

- This gene has MS peptides from three experiments Yes **X** No \_\_\_\_\_
- This gene has 75 peptides Yes **X** No \_\_\_\_\_
- Select the top hit. Click on 'Protein matches' and go to the Genome Browser. What is the protein length in amino acids? **391**



- Click on 'Configure this image'



- How many MS peptide tracks are available? 14
- They are all turn on by default. How many of these MS tracks have peptides for this gene? **Hint:** hover with the mouse above the figure to highlight each track and facilitate the count 5

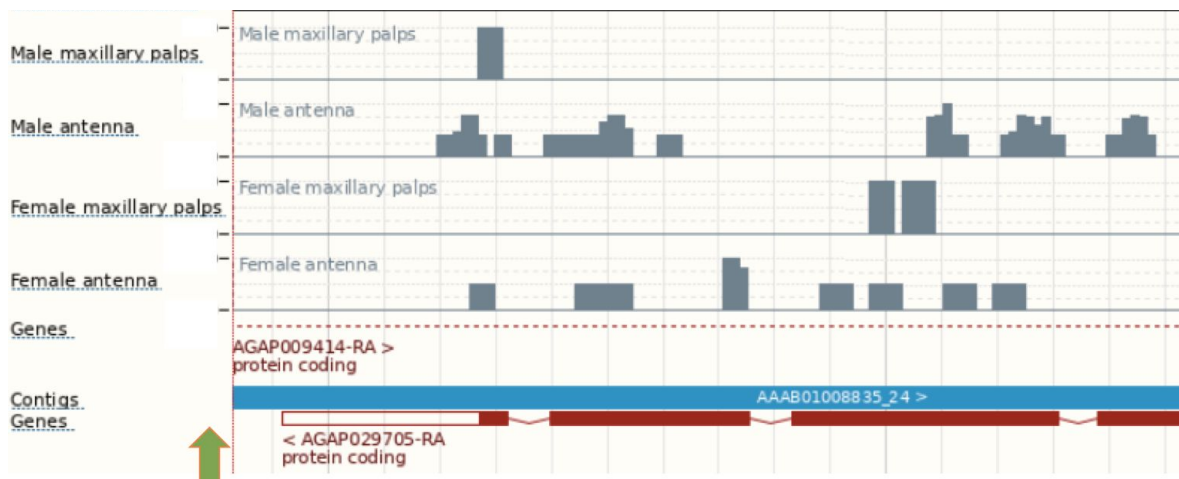


#### Genome Browser > Location tab > RNAseq data

- Use Search to query for AGAP029705 Or50 > Genome Browser > select the location tab.
  - Click Configure this page > Personal data > **Track hub registry search** > 'Search', this will show you all available **RNAseq data**. Select 'Chemosensory appendages (Pitts 2011)' > 'Attach this hub' > Configure your hub.
  - Are there 6 or 12 tracks? Are they duplicated?

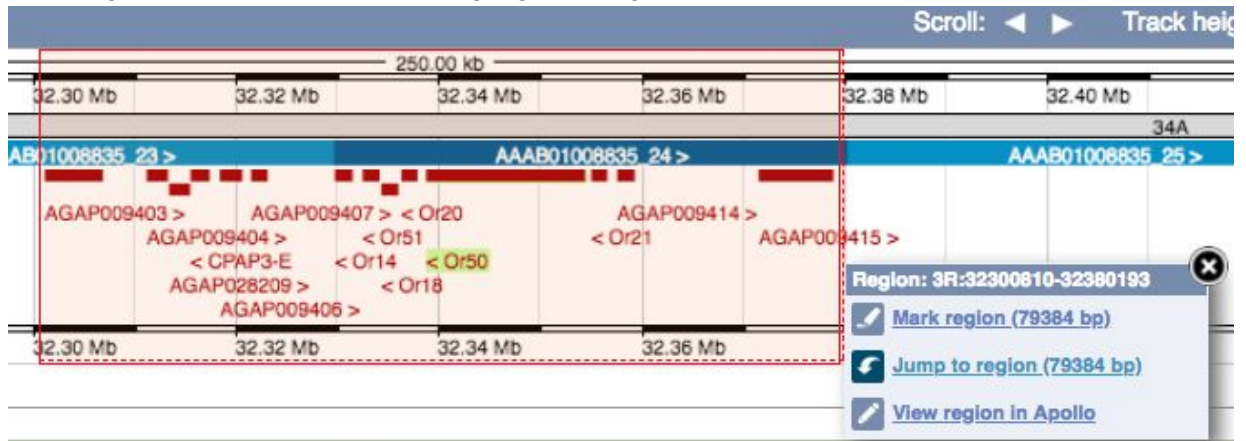
There are 6 tracks in two different formats, bigwig and bam.

- Save and close with a click outside the popup window.
- There are three panels on this page. In the bottom one, you should see the data added against the gene of interest



Write axis numbers as you see them in your screen

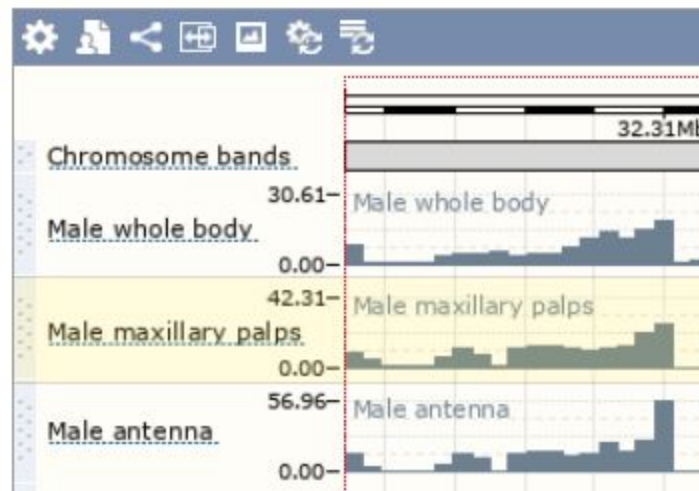
- e. In the middle panel (Region in detail) is visible a region of 250 kb. Select a region of ~80 kb including the AGAP029705 Or50, highlighted in green.



f. Look at the bottom panel.

g. How many genes are you looking at now? 14

h. Notice the scale is different for each RNAseq track. Customize the scale with a click on each track.



i. What happened to the expression of the gene AGAP029705 Or50? Compared to its neighbor genes is much lower.

j. Identify the paper where the data was published provide its author and year. Where can you find this information? **Hint:** Go back to configure this page. Click on the one track information icon.

Pitts et al 2011. Track info, follow the SRS\* links to ENA > Study > Publications

k. Go to the paper (NCBI > PMC). Open the 'Additional file 1'. Find the column with the RPKM (Reads Per Kilobase Million) and fill this table with the level of expression in each track/experiment. Work in **pairs**, each partner works with one gene in one sex:

Gene ID	Female			Male		
	Antenna fa	Bodies fb	Palps fp	Antenna ma	Bodies mb	Palps mp
AGAP009405	26.45	21.55	58.15	5.32	16.38	59.37
AGAP029705 Or50	0.99	0.10	0.16	0.92	0.11	0.07

I. what is the correlation between the paper data and the data in VectorBase?

The gene AGAP009405 was represented in VectorBase as with more transcripts, while AGAP02970 was shown with less transcripts. This correlates accordingly with what the authors reported in the paper.

14. Is the list of available tracks in the 'The Track Hub Registry' a comprehensive list of RNAseq experimental data? Select the correct choice:

- Yes \_\_\_\_\_, The Track Hub Registry is a comprehensive list because, it is one of the archival repositories together with DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI
- No   X  , The Track Hub Registry is not a comprehensive database

15. **Alternatively**, you can also query the same RNAseq data using [Advanced Search](#).

Expression > RNAseq track groups > *Rhodnius prolixus* > activation link 'Browse Genome' icon.  
Which of the following tissues are included:

	Answer
anterior midgut	X
antennae	
ovary	X
haed	

16. Go back to Search and find this gene: AGAP002045. From the genome browser provide the following information.

- a) Click on 'show transcript table' > 'protein' summary display. What is the main thing it shows?

	The correct answer is:
Differentially spliced transcript isoforms	
Homologous proteins in VectorBase species	
2D electrophoresis spot identifications	
Homology matches to conserved protein domains	X

- b) Click on the grey bar representing the Pfam domain. You should see a pop-out info box. You can reposition the box by dragging the title bar and close it when no longer needed. In the pop-out click on both of the links to open new web browser tabs for each. Briefly describe what you find in each

What is Pfam?	What is InterPro?
<p><b>Pfam</b> is a database of conserved domains curated by humans.</p> <p><b>PANTHER</b> is another similar database, and there are several more. Each database has a different biological focus and/or technical approach.</p>	<p><b>Interpro</b> 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 <b>olfactory receptor, insect 'family'</b> has a section top-right which attributes the source database(s), in this case, Pfam and PANTHER.</p>

- c) Customize Pfam and InterPro in VectorBase. Click on configure the protein domain diagram, by clicking on the gear icon -"configure this image". How many track/options are available for protein domains?

14
----

## Critical thinking: external IDs for gene Search

17. *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 query with IDs from InterPro, Pfam and the other domain databases.

- Search VectorBase with PF02949 and IPR004117 and filter into Genome > Gene > *Anopheles gambiae*. What is the number of hits obtained with each?





"odorant receptor"	PF02949	IPR004117
76	77	80

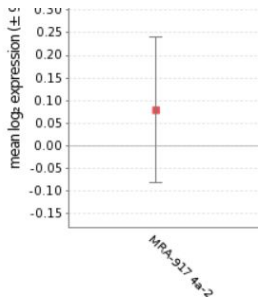


- Compare the double quoted "odorant receptor" and the InterPro IPR004117 results. Why do you think there is a difference in the number of results obtained with both queries?

	Answer
Domain databases contain evolutionary information	
Automated domain assignment via homology methods has better coverage than the human (community) annotation of gene descriptions	X
Protein similarities are a useful method for annotating genes in newly sequenced genomes.	

## Expression Browser > Expression Report

18. Use Search to query for AGAP004356 Or56. In the left hand menu select '**Expression Report**'. These are all the different experimental conditions under which this gene has experimental data in VectorBase. Answer the following about this report:

	True	False
<p>There are experiments reporting either RNAseq or microarray data</p> <div> <p><b>Experiment</b></p> <p><b>MR4 cell-lines (AVCL consortium, 2014)</b></p> <p> RNA-Seq experiment info</p> <p> Plots and data</p> <p><b>Female lower reproductive tract post-mating time-series (Gabrieli et al., 2014)</b></p> <p> Microarray experiment info</p> <p> Plots and data</p> </div>	X	

<p>There is not a link or option to obtain raw data or plots about each experiment</p> <div><p>Export option: <a href="#">PDF</a></p><p>Plot data in tabular form:</p><table><thead><tr><th>StrainOrLine</th></tr></thead><tbody><tr><td>MRA-917 4a-2</td></tr><tr><td>MRA-918 4a-3A</td></tr><tr><td>MRA-919 4a-3B</td></tr><tr><td>MRA-920 L35</td></tr><tr><td>MRA-921 Sua 4.0</td></tr><tr><td>MRA-922 Sua 5.1*</td></tr><tr><td>MRA-924 SuaE1</td></tr></tbody></table><p>Export options: <a href="#">CSV</a>   <a href="#">XML</a></p></div>	StrainOrLine	MRA-917 4a-2	MRA-918 4a-3A	MRA-919 4a-3B	MRA-920 L35	MRA-921 Sua 4.0	MRA-922 Sua 5.1*	MRA-924 SuaE1		X
StrainOrLine										
MRA-917 4a-2										
MRA-918 4a-3A										
MRA-919 4a-3B										
MRA-920 L35										
MRA-921 Sua 4.0										
MRA-922 Sua 5.1*										
MRA-924 SuaE1										
<p>The experimental factor<sup>3</sup> icons allows you to locate the experiments of interest without having to read the list of experiments</p> <div><div>Experimental factor</div><div><div>Organism part and sex</div><div></div></div></div>	X									
<p>p-value color code: The strongest red indicates the less significant result gradating through pink to white which indicates strongest significance at the 0.05 level.</p> <div><div>P-value</div><div><div>0.00054</div></div></div>		X								
<p>The p-value shows if the differential expression is statistically significant</p>	X									

<sup>3</sup> <https://www.vectorbase.org/documentation/expression-browser-user-guide>

19. **Alternatively**, you can look for each gene expression via Tools > Expression Browser > Go directly to gene and click on the sample gene, AGAP001111

**Go directly to:**  
**Gene or gene symbol:**  
  e.g.: AGAP001111

Which of these experiments do not have differential expression?

	Answer
Adult tissues (Baker et al. 2011)	
Developmental series (Koutsos et al. 2007)	
Circadian rhythm: heads, light-dark (Rund et al. 2011)	X

### Critical thinking: Why do we need a genome browser?

20. For each gene, in each genome, there is a gene (transcript and location) page in the **Genome Browser**. In your own words and based on the instructor lecture and the previous questions, why do we need a genome browser? For example, this is the paper for *Rhodnius prolixus* genome (Mesquita et al. 2015), <https://www.ncbi.nlm.nih.gov/pubmed/26627243> For this, and any other species, why is not enough with just the genome paper?

#### Case study #2: Opsin (GPRop) genes

- Data: opsins, G-protein coupled receptors or GPCRs

#### BLAST

Is a program to perform sequence similarity search

1. When planning a PCR experiment, it is necessary to know the size of the amplicon products for the gDNA and cDNA, to interpret the obtained results in the agarose gel.

- A lab colleague have designed these primers and has asked you to help double check that they are target the gene of interest
- Copy the primers sequence from this link <https://tinyurl.com/yyx7wwho>.
- Go to VectorBase Tools > **BLAST** > paste both primers
- Select BLASTn > *Culex quinquefasciatus* > Scaffolds and Transcripts > Submit

2. In the Results click on Transcripts

	Answer
To which gene both the Reverse and the Forward primer align completely?	<b>CPIJ004067</b>
Click on the best hit > Browse Genome. In which tool are you now? Which tab?	<b>Genome Browser Transcript tab</b>

3. In the Results click on Scaffolds

	Answer
Which hit (supercontig) has the lowest e-value, highest score, and higher identity?	<b>supercont3.60</b>
Click on the first hit. In the popup window click on Browse Genome. What is the green bar in the bottom display? the gene, superconting, primer (query)?	<b>primer</b>
In which tool are you now? Which tab?	<b>Genome Browser Location tab</b>

4. Repeat the previous step but for the second hit. Adjust the zoom to visualize the whole gene.

## 5. **Form pairs or groups of three**

Galaxy > BLAST

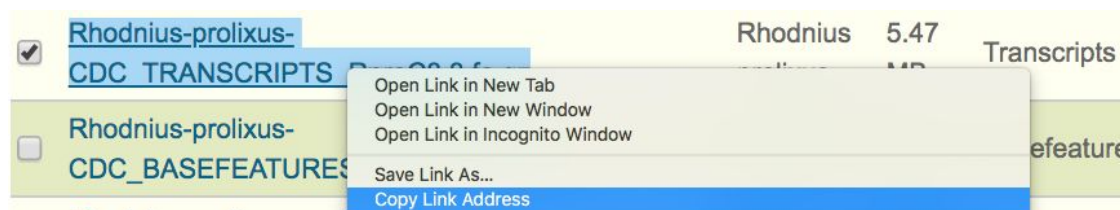
6. Now, let's BLAST the genome of *Rhodnius prolixus* against the transcriptome of *Triatoma dimidiata* using Galaxy.

- Tools > Galaxy
- Get data > upload from computer

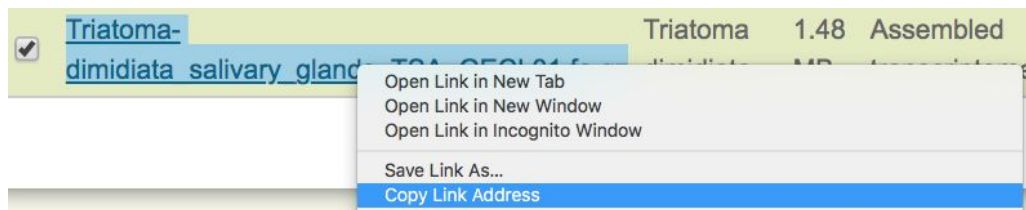




- c. Paste/fetch data. Do not close this window.
- d. **In a new window or tab.** Downloads > Data files
- e. *Rhodnius prolixus* > filter
- f. 'copy link address' from genome transcripts > paste it in Galaxy



- g. Same process to 'copy link address' from *Triatoma dimidiata* salivary glands transcriptome > paste it in Galaxy



- h. Galaxy > click start

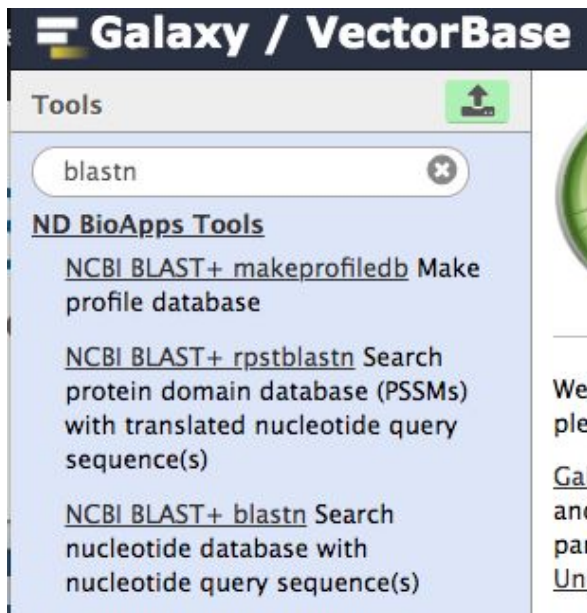


- i. Once the files are done uploading, they turn green, examine their content with 'view data'. What format are they in?

Fasta



- j. look for and select BLASTn




- k. select *Triatoma* as the query and *Rhodnius* as the subject database  
 l. select BLAST: dc-megablast  
 m. take a screenshot of the output columns names  
 n. click on Execute  
 o. this job will take time to complete, continue with the next exercises and we will get back to it later
7. Examine this <https://tinyurl.com/yd2kgjen> list of opsin genes.

To which species the genes belong to?	How many genes are there in each species?
<i>Aedes aegypti</i>	10
<i>Anopheles gambiae</i>	11
<i>Culex quinquefasciatus</i>	13

- File > Download as > Plain text (.txt)

- In the Search box click on 'switch search type'<sup>4</sup> and upload the file and click GO.



- In the Search results filter: Genome > Gene > Export > Export download
- Based on their gene metadata, there are subcategories of these genes. Open the file and complete the missing number of genes/subcategory/species:

	Long	Short	Ultraviolet	Pteropsins	Rh7-like
<i>Anopheles gambiae</i>	6	1	1	2	1
<i>Aedes aegypti</i>	6	1	1	1	1
<i>Culex quinquefasciatus</i>	8	2	1	1	1

### Genome Browser > Gene Tab > Comparative Genomics

- There are two *A. aegypti* genes with no description/function. Use the gene IDs in Search to find the genes, once you find them:
  - in the genome browser > gene tab > click on **comparative genomics > gene tree**
  - scroll to the bottom of the page and analyse the tree, your gene of interest is in red
  - what do you think is the putative subcategory of these genes?

It seems that both are long opsins

- Go back to Search and upload again the opsins\_id.txt file.
  - Filter Genome > Translation > Export > Export sequence, to download these gene sequences in amino acid format

### ClustalW > HMMER

- Use VectorBase Tools > **ClustalW** to perform a multiple sequence alignment (MSA)
  - choose file > sequence type: protein > Submit > Send to **HMMER**
  - Select these three species > Submit > provide the total number of putative opsin genes.

Species	Total number of putative opsin genes
<i>Anopheles darlingi</i>	9
<i>Lutzomyia longipalpis</i>	4
<i>Phlebotomus papatasi</i>	3

<sup>4</sup> Search 'Switch Search Type' has no limit in the number of gene IDs. But if you paste the genes directly in the Search box you can only query with a maximum of 10 simultaneously.

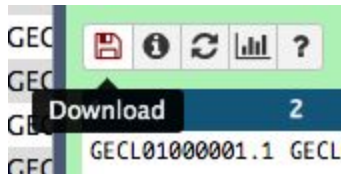
11. Is it possible to classify the genes in functional subcategories using hmmer output?  
(yes or no)

No

12. Go back to Galaxy. Is your BLAST job done? Click in 'view data'



13. Click on the file name to obtain more options. Download the output file

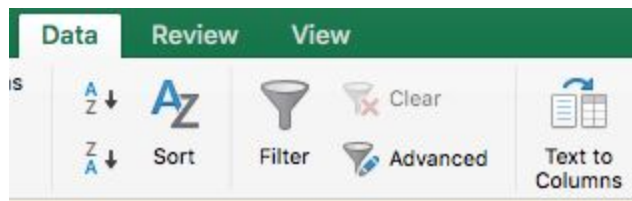


14. Change the format to \*csv and open in Excel.

15. Describe the file format and the columns, do they look as they were shown in Galaxy?  
Explain your answer

The data look different and was not organized in columns

16. Select column A > Data > Text to columns > Next (x2) > Finish



17. Describe the data format now

Data is organized as shown in galaxy, with 25 data columns

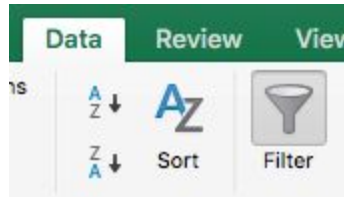
18. Label the following columns using the 1st row

- *Triatoma* transcript IDs
- *Rhodnius* transcript IDs
- % of identical matches
- alignment length
- e-value
- bit score

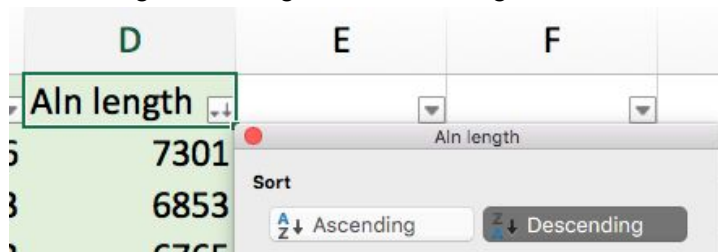
19. Are the top three the best hits?

GECL01000001.1	RPRC003346-RA	85.12	168	5.00E-53	190
GECL01000002.1	RPRC011317-RA	85.54	242	2.00E-79	279
GECL01000002.1	RPRC011318-RA	83.26	227	1.00E-67	239

20. Select all the cells and click on Data > Filter



21. Alignment length > Descending



22. Are the new top three the best hits?

GECL01003684.1	RPRC006053-RA	89.36	7301	0	9696
GECL01003461.1	RPRC017363-RA	89.23	6853	0	9019
GECL01003703.1	RPRC010799-RA	83.92	6765	0	7359

## Critical thinking: BLAST vs Hmmer vs Comparative Genomics

23. What do you think are the advantages and disadvantages of each approach to identify homologous genes?

	Advantages	Disadvantages
BLAST		
HMMER		
Comparative Genomics		

24. Let's examine these genes using VectorBase **comparative genomics** tools. Use one gene ID from the sample data set as keyword in Search: AAEL006498. In the results page filter with Comparative > Gene tree

### Apollo

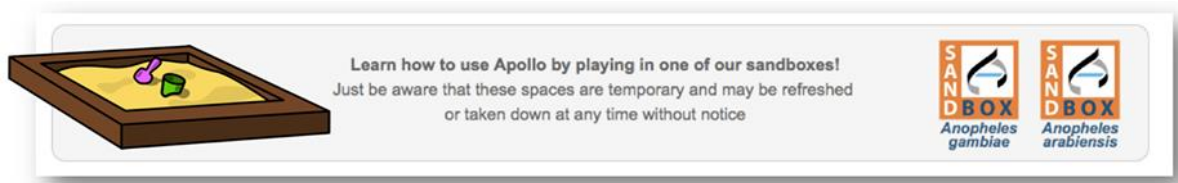
25. Examine with attention the alignments on the right-hand side of the tree. Identify all the genes that seem different from their taxonomic group or tree clade. Those are the genes that need to be fixed, how many you found?

~50 genes

26. Intron exon boundaries are fixed with Tools > **Apollo**. For examples of what can be achieved take a look at the documentation found here: Help > Tutorials > Community Annotation > Apollo, <https://www.vectorbase.org/tutorials/community-annotation-tutorials/apollo>. Identify three kind of fixes for your list of genes above

split or merge genes  
delete or create exons  
correct intron-exon boundaries or gene translation start and stop

27. Email [info@vectorbase.org](mailto:info@vectorbase.org) to request an Apollo sandbox account. Download the amino acid sequence of the genes to fix and aligned them with <http://multalin.toulouse.inra.fr/multalin/> and a gene from *A. gambiae* and/or *D. melanogaster*, to get a more detailed view of the changes to be done. Learn to use Apollo with the three videos found here: <https://www.vectorbase.org/tutorials/community-annotation-tutorials/apollo>, locate the selected genes (from question 25) and fix them in either of the VectorBase Apollo sandboxes.



### Critical thinking: Gene manual annotation submissions

28. Think again about metadata submission, and add to it manual annotation submissions. Research group 'a' assigns metadata and manually fixes ~200 genes from an insect vector. The data is published in a paper. In the same year, and without you knowing about the previous paper, you come to VectorBase and rename these same genes, edit them in Apollo and submit the data to VectorBase and publish a new paper. Authors from the original paper, now submit their gene manual annotations and metadata to VectorBase.

What would VectorBase do about this? What nomenclature is other people going to use?

## Case study #3: Transcript Expression

1. Go back to case study #2 and look at the table on step 7. How many *Aedes aegypti* long opsins it has?

6

2. Let's explore experiments in which the paralog genes may be differentially expressed

### Advanced Search > Expression

3. In the **Search** box click on **Advanced Search** > Expression > Experiment > *Aedes aegypti*. How many experiments are available?

44

4. Select three experiments that compare between males and females.

- Female vs. male (Harker et al., 2007)
- Male vs. female (Dissanayake et al., 2010)
- Male vs female *Aedes aegypti* pupal heads (Tomchaney et al., 2014)
- *Aedes aegypti* male vs female (Hall et al., 2015)

5. Compare the previous strategy with this one to select experiments: Expand the Advanced Search form and add Experiments > Experimental Factors > select Sex > GO

Advanced Search

Domain/Sub-Domain

-Experiments Add field

Experimental factors ?

DevelopmentalStage  
Sex  
Compound  
Dose

Select all  
Authors/Contacts  
Description  
Experimental factors  
PubMed

6. How many experiments are shown in the results?

4

7. Think about the concept of "Experimental Factor". Describe the search you just performed

	The correct answer is:
Finds the experiment in which the expression of genes was compared between samples taken from males and females	X

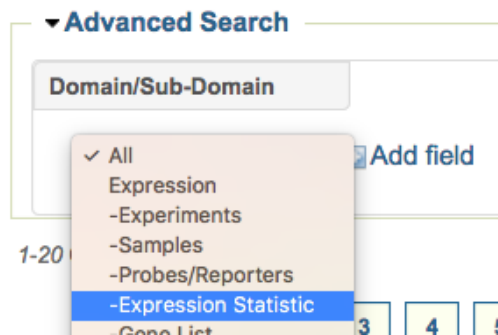


Finds the experiments in which samples were collected only from one sex	
Find the sex in which genes were upregulated in expression experiments	

8. Construct the following query:

- Go to the mosquito opsin sample file again, <https://tinyurl.com/yd2kgjen>. Copy the gene IDs for *A. aegypti* only. Paste the sequences in the Search box. Click GO.
- Open the Advanced Search form

Enter terms



- Expression statistic > Add field
- Experimental factors > Sex > GO
- Export > Export Download

9. Open the file in Excel. Select the file and expand all the columns to see all the content with a double click in the line between columns A and B.

10. At the bottom of the p-value column add significance levels to help you better interpret the data: 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.05. Color them.

11. Select the file. Use Data > Filter. Unselect the file and organize the data from the lowest to the highest p-value (A -> Z Ascending). Compare this with the column called 'Description'.

12. Is there differential expression between males and females?

a) Provide the correct answer Yes X No \_\_\_\_\_

b) For which genes?

AAEL003035  
AAEL005621  
AAEL006259  
AAEL009615  
AAEL005625  
AAEL007389



c) What is the name of the Experiment(s) and the paper(s) author and year?

Male vs. female (Dissanayake et al. 2010)  
Male vs. female *Aedes aegypti* pupal heads (Tomachaney et al., 2014)

d) Are these: RNAseq \_\_\_\_\_ or Microarrays \_\_X\_\_

**Hint:** Go back to VectorBase Search and click on any two hits

e) Is higher in males or females? Or is different for each gene?

In the all six genes is higher in males

13. Go to **Advanced Search** and provide the number of experiments for:

<i>Anopheles albimanus</i>	1
<i>Anopheles darling</i>	0
<i>Aedes aegypti</i>	44

14. Select the following *Anopheles* experiments and answer if the statements are true or false:

	True	False
<i>A. albimanus</i> experiment is about the evolutionary aspects of sex-biased genes in carcass and reproductive tissues	X	
<i>A. coluzzii</i> does not have microbiota experiments		X

## Expression Browser

15. Go to the Tools menu > **Expression Browser**.

16. This tool shows the all the 133 experiments you found previously using Search

17. We have an internal pipeline that compares all experiments among all, for each species.

18. Query for the gene AGAP013149, using this tool 'Gene or gene symbol' box not VectorBase main Search box.

### Go directly to:

Gene or gene symbol:

e.g.: AGAP001111

19. Answer the following:

	Answer
How many experiments have differential expression for this gene?	12
Are there experiments showing differential expression between males and females? If yes, which one(s)?	Sex-biased expression in gambiae (Papa et al. 2017)

### Critical thinking: *in vivo* differential expression data

20. VectorBase expression data can be used to report new results on previously published data, propose new hypothesis or conduct metanalyses. You are interested in the interaction vector with parasite or virus, and will like to write a proposal to fund your research. How can you use VectorBase to look for experiments, in all vector species, to identify potential candidate genes? Briefly list the steps and your results

Advanced Search > Experiments > Experimental Factors > Disease State

Aedes aegypti	16 experiments (~20,000 genes and ~34,000 transcripts)
Anopheles gambiae	12
Glossina morsitans	2
Glossina palpalis	1
Ixodes scapularis	1

**Examples:** plasmodium over time in different tissues (midgut, fat body), hemolymph response to bacteria, mosquito infected with microsporidia, different mosquito strains against dengue, different virus in the same mosquito strain (dengue, west nile, yellow fever, tsetse Trypanosome infection in different organs/tissues, time series of infected blood meals (24 h, 96 h).

### Case study #4: Insecticide resistance from Ethiopia

#### Population Biology > Insecticide Resistance

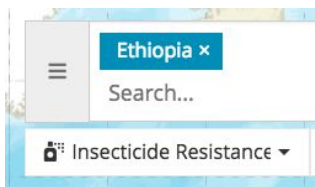
While you interact with the PopBio map interface please send your feedback to [info@vectorbase.org](mailto:info@vectorbase.org) (e.g., I like ..., I do not like ..., It was easy for me to ..., I could not figure out how to ...), VectorBase developers are actively working on this tool, thank you!

1. Go to Tools > **Population Biology** or PopBio. In the map select the insecticide view

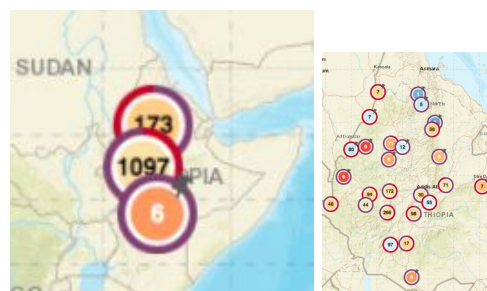


Option 1	Option 2

2. Use Ethiopia as a keyword. Notice how autocomplete suggest this word for the geography (ontology) category



3. Do not zoom. How many markers (circles) do you see? Draw them with their numbers in this image of the country map. Notice there is one with a pin, make sure to draw the pin too.



answer key: 173, 1.097, 6

4. Slowly zoom in Ethiopia.

- What happen to the markers? Do you see more or less?



The markers divide, there are more markers

- Do you still recognize any of the original markers? Go back to question 3 and draw a square around it.

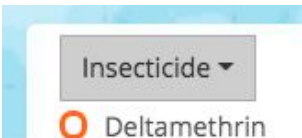
5. Let's take a closer look at the markers. Look at the right hand side filters. Which species are in the area?



Anopheles arabiensis and A. gambiae sensu lato

6. Using the left-hand side menu:

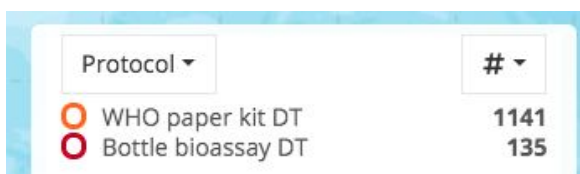
	<p><u>Summary statistics.</u> Select two markers, one at the time. List the number of species that you see.</p> <p>1. A marker with one color for the outer ring:  <input type="text" value="One species"/></p> <p>1. A marker with two colors for the outer ring:  <input type="text" value="Two species"/></p>
	<p><u>List details- Details for selected samples.</u> Look for any two markers with a number inside less than 10. Based on your observations what the numbers inside the markers mean? Also look for a white bar with a text in the bottom of your screen</p> <p><input type="text" value="The number of assays summarized by species"/></p>

7. Let's visualize the data with different filters

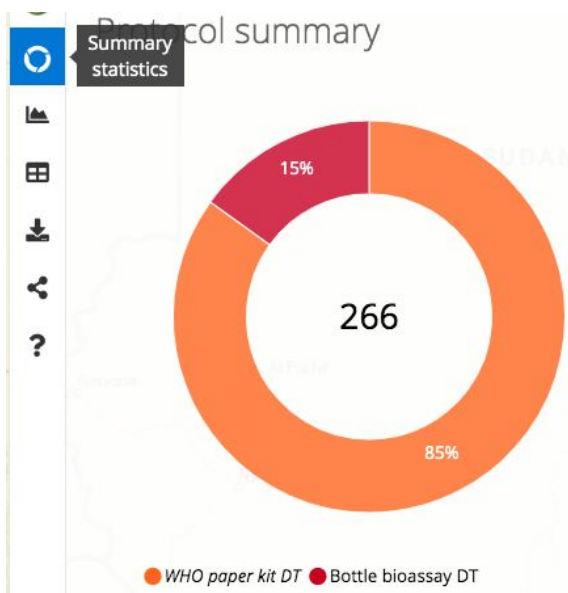
	Answer
<p>Use the right hand side menu to filter by Insecticide type. How many insecticide types are there for our country query?</p> 	<p>13</p>

<p>Notice the colors are not the best, notice how the last three in the list are gray. Click on 'Optimize colors'. What happened?</p> 	<p>The markers are better color identified now</p>
<p>Click on a marker with a number lower than 10. Click again in list details. What the number inside the marker means now?</p> 	<p>The number of assays summarized by insecticide</p>

8. In the right hand side menu filter by protocol.



9. Select a marker and click on summary statistics



10. Briefly explain the results obtained, counts and percentages, based on the images above.

Overall, the complete country query (Ethiopia), has 1141 insecticide assays (89%) done with WHO paper kit and 135 (11%) assays with bottle bioassay. For the selected marker, 266 assays have been performed, 85% using WHO paper kit and 15% with the bottle assay.

11. Filter for assays with a desired level of insecticide susceptibility

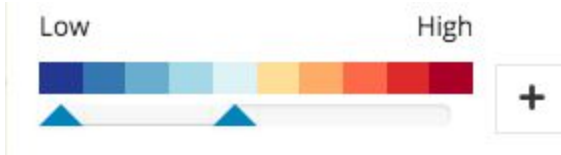
- List the colors observed inside the markers

different tones of blue and red

- Click on Search menu (the three horizontal lines)



- Move the arrow heads to select only the blue, susceptible results. Click on the plus.



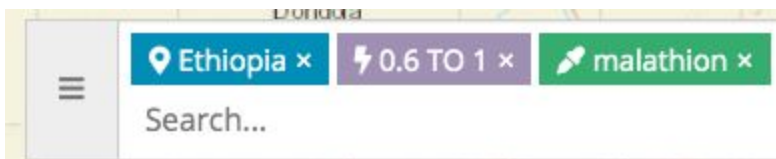
- Again, list the colors observed inside the markers

only different tones of blue

- To observe only Malathion data, click on it



12. Reset the query with a click on the Malation X



- You should have again all 11 insecticides.
- Exclude for the query the insecticides with less than 100 assays. Simultaneously click Control (Widows) or Command (Mac) and on the insecticide.

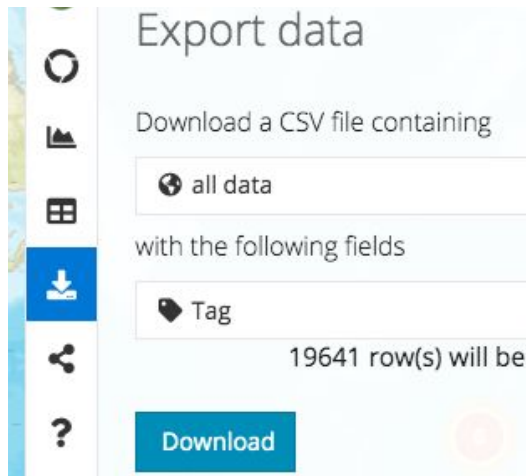


13. Select a marker, again click on List details (left hand menu).

- For a brief description of the protocol followed, click on the square and arrow next to Assay VBAxxxxxxx, a new window tab will open. Notice that information such as the protocol diagnostic dose is provided.
- Which information is provided if you click on the information icon (i) next to VBPxxxxxxx?

Publication details, time since data in VB and tags in available

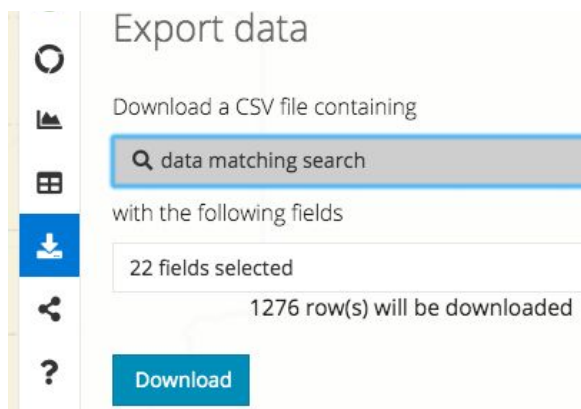
14. Download: all data > tags



- which tag is available for the insecticide resistance view? (different tags are available in each map view)

WIN : The Worldwide Insecticide resistance Network

15. Download: data matching search > 22 fields selected



16. Explain in your own words what the 'Data reuse policy' means and the proper citation format for VectorBase and its data.

It is important for data publishers and for the reputation of data users that data is cited correctly. Citation information for each record can be found in the downloaded file.

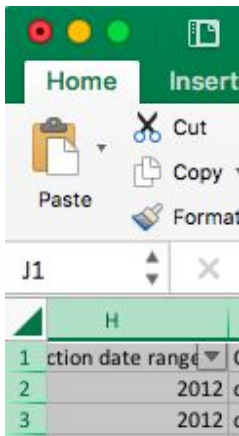
We also encourage users to support VectorBase's data collation and open data efforts. Please see [here](#) for information on how to acknowledge VectorBase.

Ask students to explain and have a quick classroom discussion.

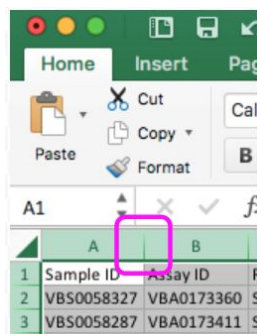
Make sure they click on the red [here](#) link.

17. Open the downloaded file with Excel or another spreadsheet program. *Well-structured data is easy to understand.*

18. Look through the downloaded Ethiopian dataset and try to understand it. Select all the data, with a click on the corner mid triangle.



Double click on the space between A and B, this will make the columns wide enough to read all the column labels and read their cell content.



Use the raw data for your own analyses



19. How to include a reference of the PopBio data in your publications?, here is an example:

448

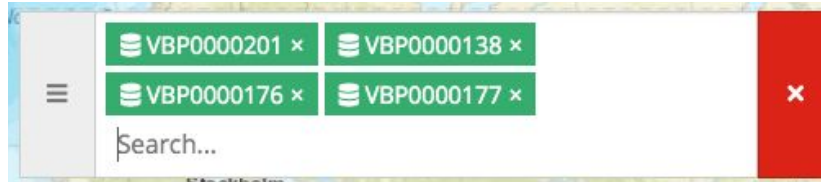
GLORIA-SORIA

**Data availability.** Microsatellite and SNP genotypes were deposited in VectorBase<sup>48,49</sup> PopBio projects: VBP0000201 (new data), and VBP0000138, VBP0000176-177 (previously published data). Sequencing data were deposited in NCBI under accession numbers MF371160–MF371174 and MG241351–MG241354.

<https://www.ncbi.nlm.nih.gov/pubmed/29260658>

Gloria-Soria, A., Lima, A., Lovin, D. D., Cunningham, J. M., Severson, D. W., & Powell, J. R. (2018). Origin of a High-Latitude Population of *Aedes aegypti* in Washington, DC. *The American journal of tropical medicine and hygiene*, 98(2), 445–452. doi:10.4269/ajtmh.17-0676

- This is how a reader from your paper could query for the data:



- Reset search and clear results with the (red) right hand X

## Critical thinking: open access data

20. What are the advantages and disadvantages of having your data in the PopBio map vs only published in a scientific paper?

The paper is a static view of the data.

The PopBio map:

- gets updated every two months
- allows data visualization with different filters
- can be analysed at different geographic levels
- data from both scientific research and control/monitor agencies can be compared
- your papers gets more exposure and users can find it both via (PubMed, Google Scholar) literature search or via the PopBio map
- if a reader from your papers wishes to have your raw data, they do not need to contact you, it can be download from VectorBase

Case study #5: Variation data of the voltage-gated sodium channel gene

- 2:11 min video: How does insecticide resistance happen?  
<https://www.youtube.com/watch?v=D1aU7HNh4jM>

## Genome Browser > Variation tab

1. In VectorBase Search box type this query 2L:2422652-2422652, filter with Variation. What are the alleles for this position?

Reference: <b>A</b>	Alternative: <b>T</b>
---------------------	-----------------------

2. Select (Genomic context >) Genes and regulation in the left hand side menu. What is the complete codon for both the alternative and reference alleles? Underline the mutated position in each

Reference: <b>TT<u>A</u></b>	Alternative: <b>TT<u>T</u></b>
------------------------------	--------------------------------

3. True or False, is this a missense variant?

<b>True</b>
-------------

4. Where is the variant located?

chromosome: <b>2L</b>	Base pair coordinates: <b>2422652-2422652</b>
-----------------------	---

5. This variant is a:

CNV:	Inversion:	SNP: <b>X</b>
------	------------	---------------

6. What happens if you hover with the mouse in the items with dotted lines?

	Answer
VectorBase collapses and shouts down!	
Nothing happens	
The terms definitions appear	<b>X</b>

7. Select the 'Population genetics' icon. Which populations have 100% of the reference allele for *Anopheles gambiae*?

	Answer
a. Burkina Faso	
b. Guinea	
c. Kenya	<b>X</b>
d. Uganda	<b>X</b>

8. In the left hand side menu go to 'Phenotype data'. In which paper is the kdr phenotype reported? Only provide last author last name and year of publication

Ranson 2011

9. What is this phenotype and its mutation name?

resistance to treatment with the insecticide permethrin, kdr

10. What is the gene for this SNP? Provide the VectorBase gene ID

AGAP004707

a. Click on 'Genes and regulation'. This shows:

- nothing \_\_\_\_
- the gene transcripts that are associated with this SNP X
- always 13 transcripts, for all genes \_\_\_\_

b. To find other variants. In the gene tab click on Genetic Variation > Variant table.

- Scroll to the table below
- There are different filters available.
- For example, SIFT: calculates preservation of the amino acid sequence and domain in relationship to its orthologues. Values go from deleterious (0, red) to tolerated (1, green). Filter for SIFT 0 to 0.5 & Consequences: missense variant
- How many positions in this gene have variants? Hint: column headers can be sorted and table can be download

Two

11. How to locate SNPs (and INDELs) using Search?

- Advanced Search > Variation > short variations<sup>5</sup>
- What are the species with the higher and lower number of short variants?

Higher: *A. gambiae*

Lower: *C. quinquefasciatus*

- Selecting a hit will take you to the variant details as displayed in which tool?

genome browser

12. Search > AGAP004707 > Genome Browser > gene tab

13. Are there more SNPs in this gene? Follow these steps to find out:

- In the left hand menu select Sequence
- Scroll to the bottom of the page. You should see the gene sequence

---

<sup>5</sup> Note: the variant IDs are temporal, tmp\*, do not use them in your publication. Instead, reference them based on position, chromosome or superocnting and base pair range (position start and end).

- Go up again and click on 'Configure this page'
- In the pop out window select 'Show variants > yes and show links'. Save & Close (check mark icon or outside the popup window).
- Result: the position of the variants relative to the gene

14. Are the variants in the introns or only in the exons?

Both in the introns and exons

15. The variants are color coded based on ontology terms. Match the columns of the term and its definition. **Hint:** this is the page for the Sequence Ontology, <http://www.sequenceontology.org/>



a. Downstream	( <b>b</b> ) A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved
b. Missense	( <b>d</b> ) A sequence variant where there is no resulting change to the encoded amino acid.
c. Splice region	( <b>a</b> ) A sequence variant located 3' of a gene
d. Synonymous	( <b>e</b> ) A sequence variant located 5' of a gene.
e. Upstream	( <b>c</b> ) A region surrounding a cis_splice site, either within 1-3 bases of the exon or 3-8 bases of the intron

16. The variants are labeled using the IUPAC nomenclature.

- mouse over the variants to see the reference and alternative allele for each one
- click on each variant for a window with more details

c. What is the source of the first variant?

Source: **VBP0000163**

d. What are the nucleotides for the reference and the alternative alleles in the first variant?

Reference **G** / Alternative **C**

17. Visualize the SNPs in the gene and its splice variants.

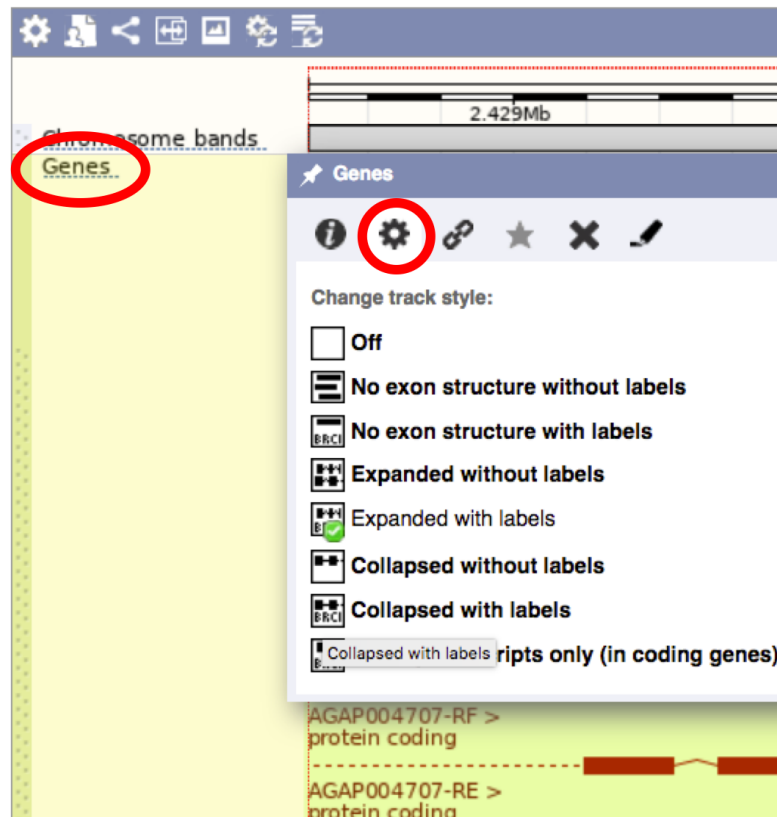
e. How many splice variants has AGAP004707?

**13**

- f. Go to the location tab. You arrive to the 'Region in detail' page (look at the left hand menu). The gene is shown in three levels of magnification: chromosome, region in detail and the gene alone.
- g. Click on 'Configure this page'. Select the Variation section. Add these tracks: variant - VBP0000004 and VBP0000163, track style 'collapsed'. Click VBP0000163 information and provide its description. Save and close.

VBP0000163: **Anopheles 1000 genomes consortium**

- h. zoom in the 3' region of the gene, it shown both missense, synonymous and splice region variants until you see approximately a window of 2 kb
- i. Change the style of the 'Genes' track from 'Expanded with labels' to 'Collapsed with labels'.  
**Hint:** click on the Genes' track dotted line > gear icon



- j. What do these two styles show?

- Collapsed shows the gene with all of its exons.
- Expanded shown each one of the gene splice variants, with their corresponding exons.

18. Also in the lower panel:

- k. With the mouse select a region and 'mark region' or 'jump to region'.
- l. Change the 'variant - VBP0000163' track from collapsed to expanded style, or vice versa.

19. When visualising SPNs, details may include a phenotype.

- a. Go to this location

**Location:** 2L:2422617-2422687

- b. Select the variant with only a 'W' and click on 'phenotype data'

Variant: tmp\_2L\_2422652\_A\_T

[more about tmp\\_2L\\_2422652\\_A\\_T](#)

Class	SNP
Location	2L:2422652
Alleles	A/T
Ambiguity code	W
Consequence	missense variant
Sources	VBP0000004, VBP0000163, Pubmed

[Population genetics](#)

[Phenotype data](#)

- m. Additionally, all the phenotypes associated with a gene, above you are only looking at a single position in the gene, can be found in the gene tab > phenotypes

Location: 2L:2,422,617-2,422,687 **Gene:**

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
- Sequence
  - Secondary Structure
- Literature
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Ontologies
  - GO: Cellular component
  - GO: Biological process
  - GO: Molecular function
- Phenotypes**

20. Go back to a specific variant. Locate yourself in the variant tab

**Anopheles gambiae** (AgamP4) ▼

Location: 2L:2,422,617-2,422,687 **Variant: tmp\_2L\_2422652\_A\_T**

**Variant displays**

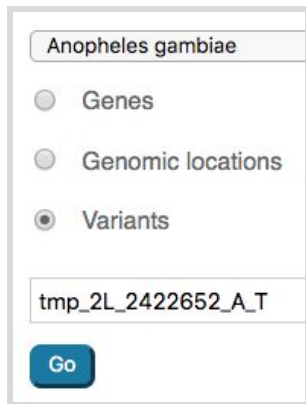
- Explore this variant**
- Genomic context

tmp\_2L\_2422652\_A\_T :

## Find data display > Variations

21. How to find more types of data display? Tools menu > Find a data display

22. Select *Anopheles gambiae* > variations > (default query) > Go



23. How many views are available total?

29

24. How many views are available for each category? Select one and describe it

	Available per total views	Describe one view per category
Locations	4/5	
Genes		
Transcripts		
Proteins		
Phenotypes		
Populations & Individuals		

## Critical Thinking: VectorBase data tools and resources

25. Can you use VectorBase figures, tables, images, screenshots raw data in your scientific papers, posters thesis, talks or any other publication type?

Help> How to cite VectorBase

<https://www.vectorbase.org/faqs/how-cite-vectorbase>

Download > Images

<https://www.vectorbase.org/image-gallery>

Tools > Find a data display

<https://www.vectorbase.org/info/website/gallery.html>