Comparative Genomics

Answer key

Comparative Genomics Basics

VectorBase has protein and DNA comparisons:

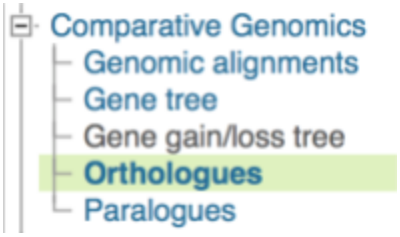
- Protein comparisons are done for all genes in all VectorBase genomes. These alignments are the input to generate the gene trees. With the gene trees we can infer homologous relationships, paralogues and orthologues genes.
- DNA comparisons are done for the closely-related (selected genome) species, these are pairwise alignments for both coding and non-coding regions, not ideal for fragmented genomes.

Statement	**True**	**False**
The power and resolution of comparative genomics pre-calculated analyses is more with a very fragmented genome (with many supercontigs)		x
VectorBase has a pipeline to annotate genomes using automatic computational approaches and similarity based evidence.	X	
Similarity data (for genome annotation), comes from both previously sequenced genomes and experimental evidence for the species under study such as RNAseq.	X	
VectorBase gene annotation pipeline is 100% accurate to find genes: - if there are mistakes or gaps in the assembly, - if the genes in question are rapidly evolving and very different from the ones of previous species, - or if they have no experimental evidence available.		X
Is not necessary to work in the improvement of genome assemblies and gene sets, they do not improve comparative genomics computed data		X

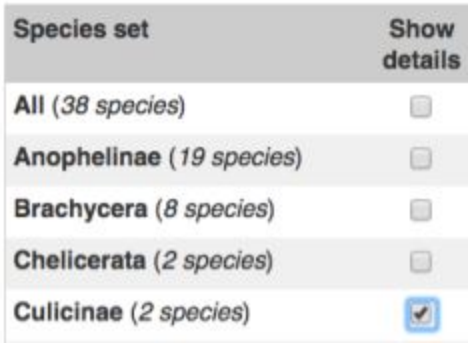
Part I: Gene trees, orthology and paralogy

1. Finding gene Orthologues

Use AAEL006498 (*Ae. aegypti* opsin 1, GPRop1) in a VectorBase Search. Click on the top hit. Once in the genome browser, click on the "orthologues" link of the gene-based display menu



See counts of orthologues in various clades. Select Culicinae for the display

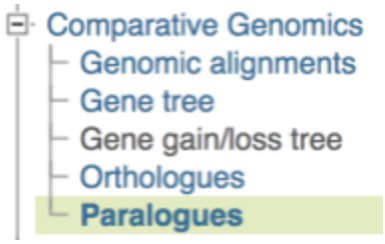


How many *Aedes albopictus* orthologues are predicted for AAEL006498? Give its gene ID.

One, AALF009534

2. Finding gene Paralogues

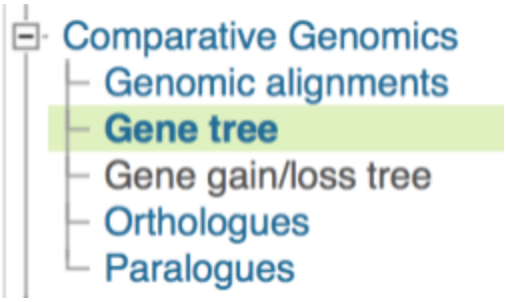
Click on Paralogues.



Based on the paralogues table, are GPRop4 and GPRop5 in the same supercontig?

	True	False
Yes, and the supercontigs are AAEL005625 and AAEL005621		x
Yes, and the supercontig is 1.165	X	
No, these genes are not in the same supercontig		X

6. Click on Gene tree to have a visual representation of AAEL006498 homolog genes



7. Click on the collapsed node called “Stegomyia: 2 homologs”. In the pop out select “Image expand this sub-tree”. What are the two species and their gene IDs and/or gene symbol?

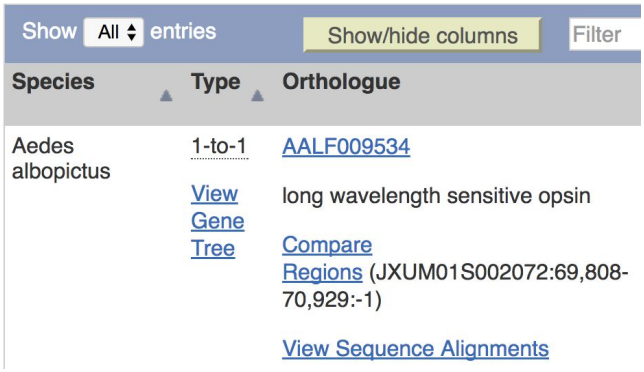
GPROP2, Aedes aegypti
AALF017696, Aedes albopictus

Part II: Genomic alignments

8. Go back again to the Orthologues display of AAEL006498 and again select Culicinae

In the “Orthologue” column, click on *Ae. albopictus* “Compare Regions” link

Selected orthologues



Which Genome Browser page (or tab) is the display showing?

Location

9. Go back to the Orthologues display in the Gene page or tab

Click on 'View Sequence Alignments'. What are the target and query percentage of identity between the two proteins?

94.64% for both

What is the length of the proteins? **Hint**: Click on 'View sequence alignment'

Both have 373 amino acids

10. In the Orthologue alignment display, there is a table and below its alignment.

Orthologue alignment

Download homology

Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Aedes aegypti	AAEL006498	AAEL006498-PA	373 aa	94 %	99 %	supercont1.208:1715269-1716390
Aedes albopictus	AALF009534	AALF009534-PA	373 aa	94 %	99 %	JXUM01S002072:69808-70929

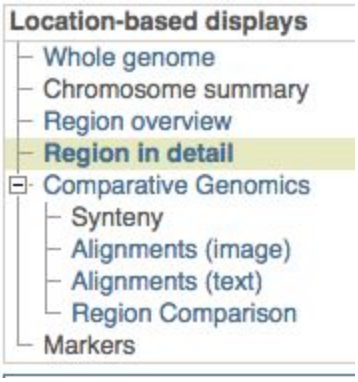
```
CLUSTAL W(1.81) multiple sequence alignment

AAEL006498-PA/1-373     MAAFVAPHFDAW-QSSGNMTVVDKVPPEMLHMVHPHWNQFPPMNPLWHSILGFAIFVLGV
AALF009534-PA/1-373     MAAFVEPHFDAWQQTTSNMTVVDKVPPEMLHLVHPHWNQFPPMNPLWHSILGFAIFVLGV
                        ***** ****** *:::.***************:*****************************
```

click on the link of the "Genomic location" column. You are now in the genome browser of which species?

Ae. albopictus

In this new page select "Alignments (image)". Which species is available to do the alignments?

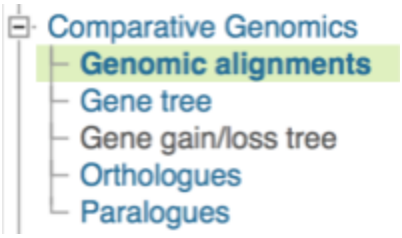


Ae. aegypti

Select "Alignments (text)". Select the available species for the alignment and click 'GO'. Click on "Download alignment". List three of the available formats to download the alignment.

FASTA, Nexus and Phylip

11. Close the pop out window. Go back to *Aedes aegypti* AAEL006498 gene page or tab and click on "Genomic alignments".

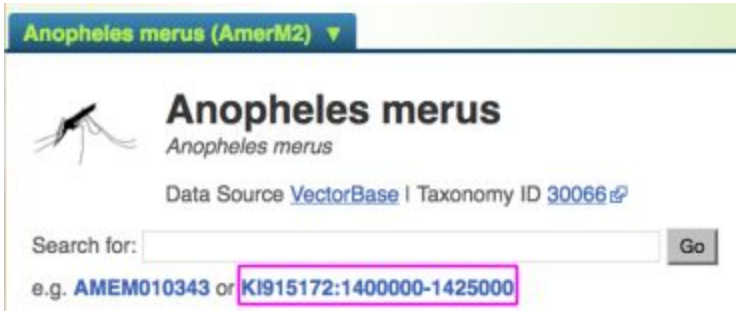


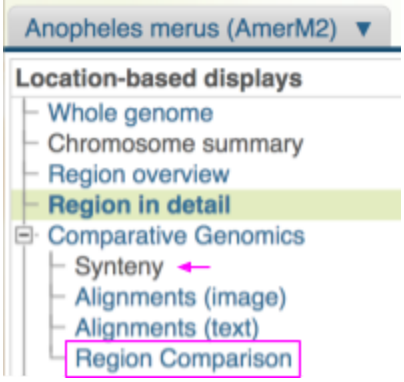
Unlike the previous ones, these are not proteins alignments, these are DNA alignments. Which species is/are available to do the alignments?

Aedes albopictus, Anopheles darlingi, Anopheles gambiae, Culex quinquefasciatus, Ixodes scapualris and *Pediculus humanus*

12. Let's go to *Anopheles merus* from the top left hand side navigation bar, clicking on the arrow head

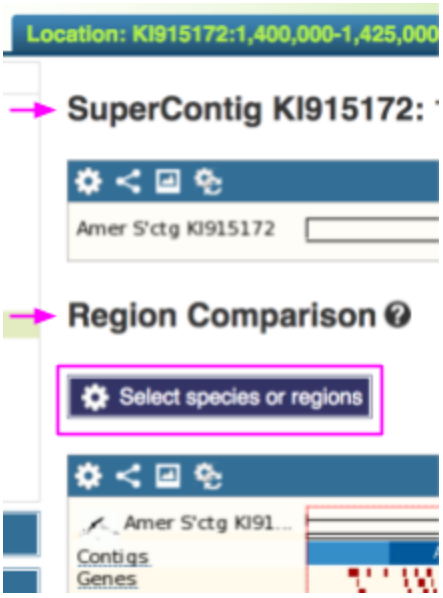
Select the example under the Search box with the supercontig and base pair range





You arrive to the Location page. Although the synteny link is not functional (is greyed out), you can still visualize some syntenic regions using the three different "Region" links. Click on "Region Comparison".

Click on "select species or regions". Select all available species. Save and close. Explore the three panels available for the region: SuperContig (top panel), Region comparison (middle) and a much detailed view at the bottom.



Using the "Configure this image" gear icon,

you can add select the "Comparative features" -> "Join genes" option - which draws lines between orthologs.

How this might be useful for locating "missing genes"? (i.e., those that have not yet been annotated by VectorBase or the community, but are expected to be present in the genome).

Statement	**True**	**False**
A strategy to find 'missing genes' is with VectorBase pre aligned DNA data/results.		

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.