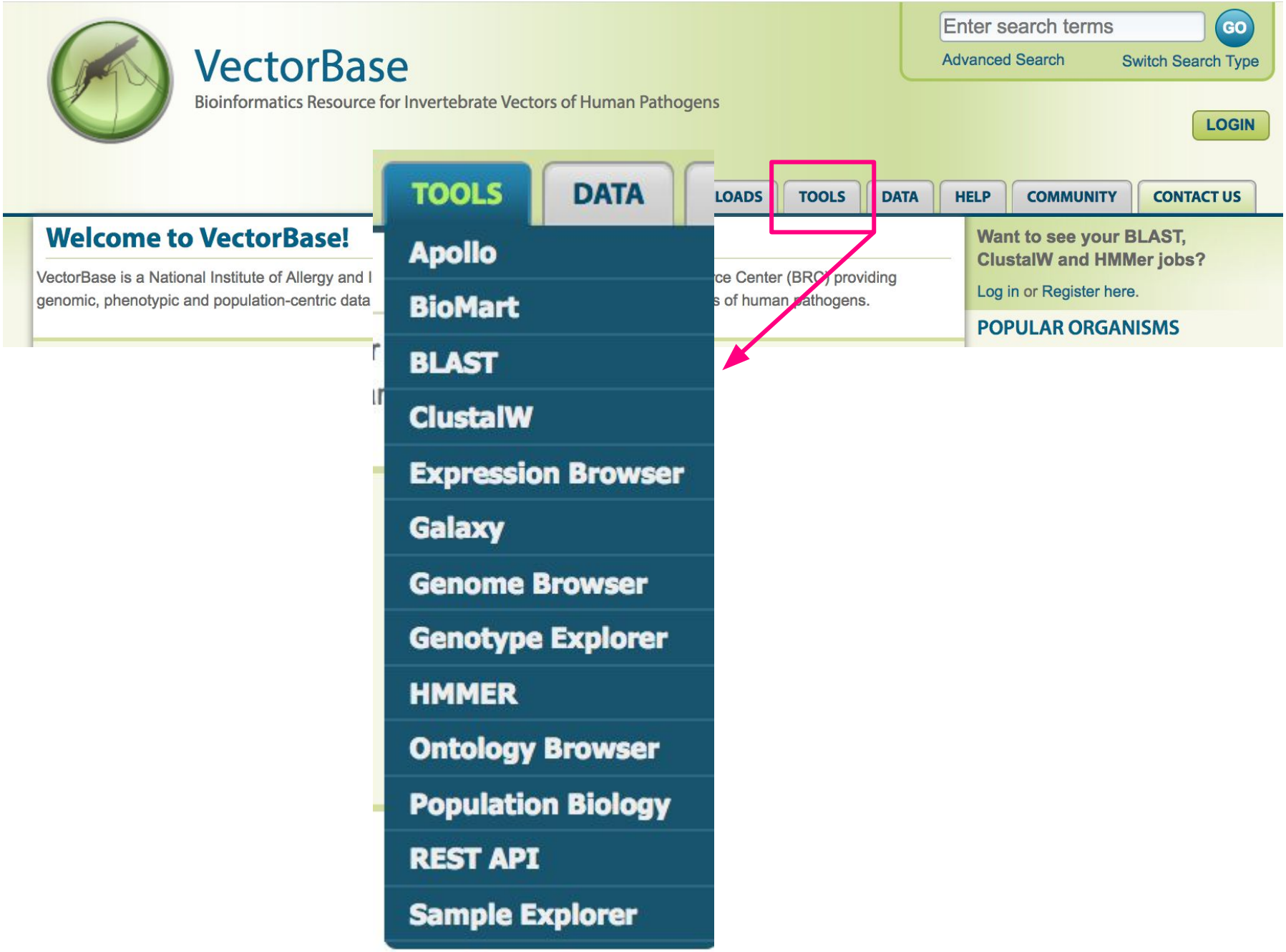
An overview of VectorBase tools and resources

Gloria I. Giraldo-Calderón
September 2017

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens


VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens
Enter search terms
GO
Advanced Search
Switch Search Type
LOGIN
TOOLS
DATA
LOADS
TOOLS
DATA
HELP
COMMUNITY
CONTACT US
Welcome to VectorBase!
VectorBase is a National Institute of Allergy and I
genomic, phenotypic and population-centric data
ce Center (BRC) providing
s of human pathogens.
Want to see your BLAST, ClustalW and HMMer jobs?
Log in or Register here.
POPULAR ORGANISMS
Apollo
BioMart
BLAST
ClustalW
Expression Browser
Galaxy
Genome Browser
Genotype Explorer
HMMER
Ontology Browser
Population Biology
REST API
Sample Explorer

Tools

Tutorials

Apollo
Is an instantaneous, collaborative, genome annotation editor. Apollo is designed to support geographically dispersed researchers.

Biomart
Use for (small and big scale) data mining queries that are not as easy or even possible to do using VectorBase Search

BLAST
Finds regions of local similarity between sequences. Available data sets include contigs, scaffolds, chromosomes ESTs, RNAseq, transcripts and peptides.

ClustalW
Is a sequence alignment tool. Can be used to generate input files for HMMER. After running a job just click on the link "Send to HMMER".

Expression Browser and Map
It has microarrays and RNAseq mostly from *An. gambiae* and *Ae. aegypti*. Data from different publications is processed through the same pipeline so that results can be compared side-by-side.

Galaxy
Galaxy is an open, web-based platform for data intensive biomedical research.

Genome Browser
Makes genomic data accessible. Data is not only the genome sequence itself, but also other features such as comparisons between species including *in silico* and experimental data.

Genotype Explorer NEW!
Explore variation data associated with biological samples in genomic , protein and multi-species contexts.

HMMER
It looks for homologous genes, but unlike BLAST it aims to be more accurate and better to detect remote homologs. Input file is a protein multiple sequence alignment (MSA) from ClustalW.

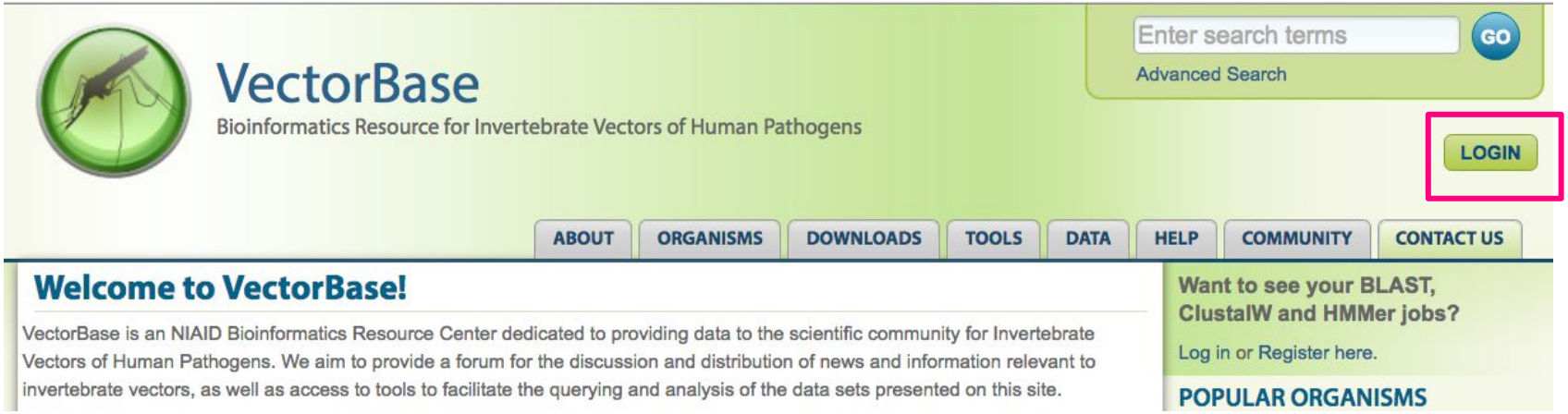
Ontology Browser
Ontologies are the structural framework for organizing information and are used in the Expression Browser and PopBio. You can also use it to annotate the metadata of your research.

Population Biology (PopBio)
Is part of our ongoing efforts to integrate genomic, phenotypic (including insecticide resistance) and population data (including SNPs and microsatellites).

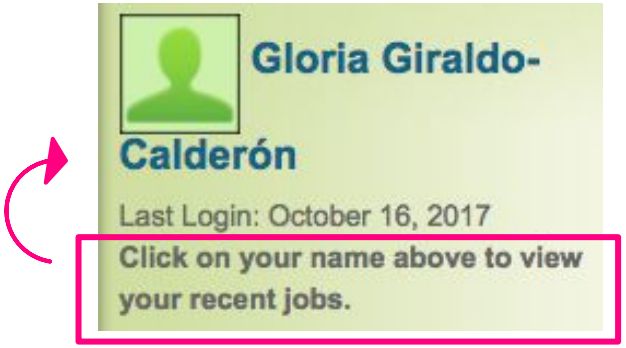
REST API NEW!
Direct programmatic access to VectorBase species data.

Sample Explorer NEW!
Search and explore metadata associated with biological samples and display them in the Genome Browser and PopBio applications.

Optional

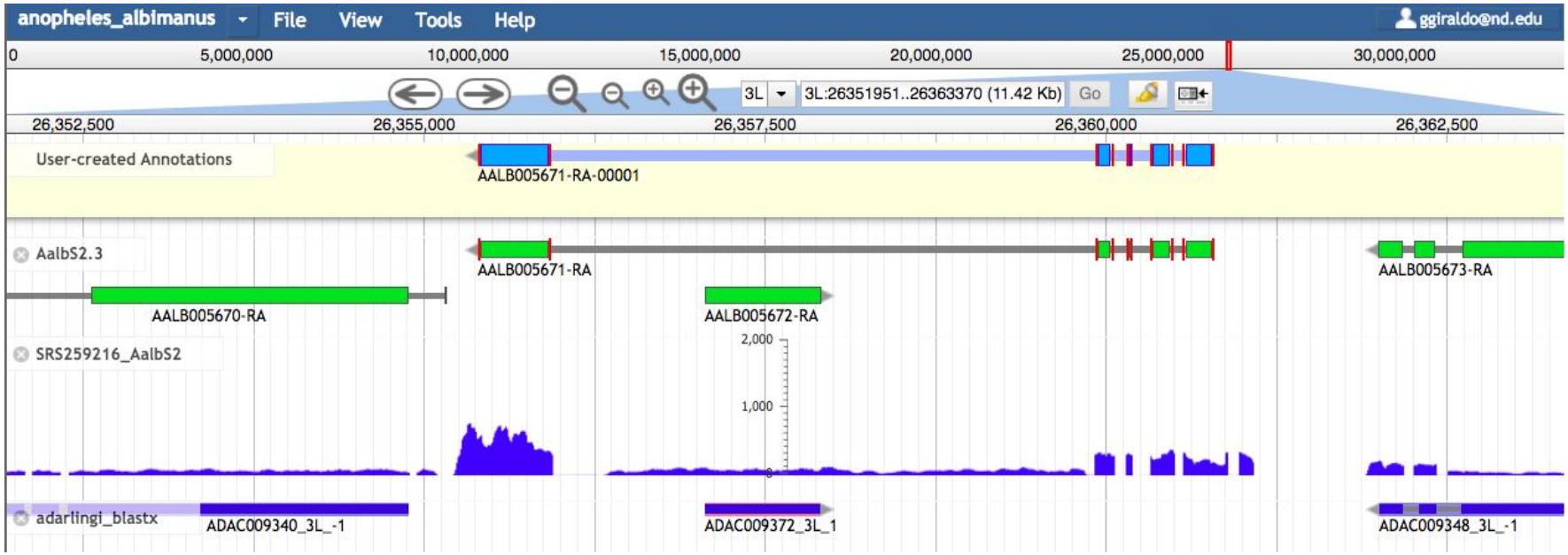


If you LOGIN, your Blast, ClustalW and Hmmer jobs will be saved and viewable on your user page.



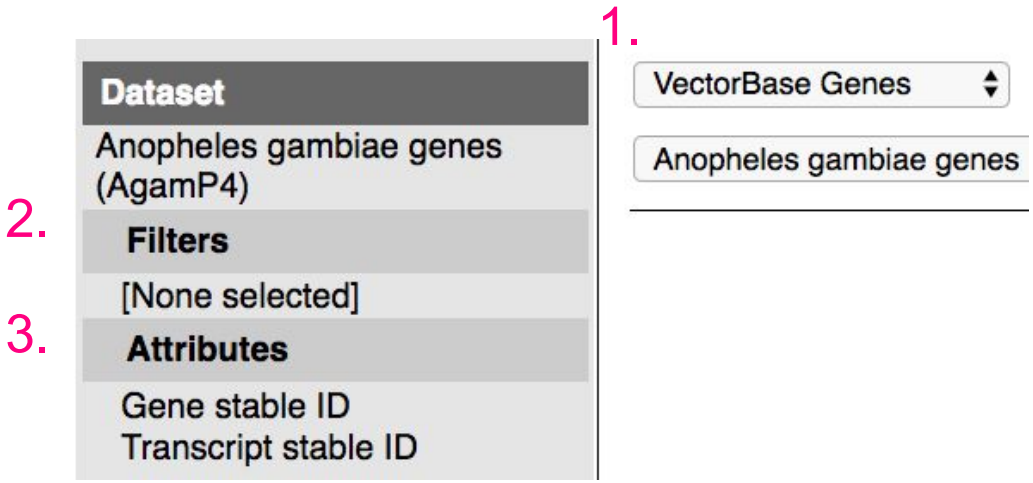
Apollo

- It is a gene editor
- It's a web applications, which allows instantaneous, collaborative work
- In real time, all users can see each others gene manual annotations (intron-exon boundaries)



BioMart

- It is a biological data mining tool for complex queries
- Could be used across organisms
- Allows users to perform specific queries
- Output: on screen (HTML), file (XLS) and others



Sample query: How many genes are on chromosome arm 2L?

1. Select the database 'genes' and species '*An. gambiae*' dataset
2. Select filters:
   Region ---> Chromosome: 2L
3. Select attributes:
   Gene ---> Gene description

BLAST

It finds regions of local similarity between sequences

Results

Job 135133 *To view results, select a link from the Database list below*

Description: AaGPRop1
Submitted: Wednesday, June 8th, 2016 13:18:38 -0400
Compute Runtime: 8 seconds

CLEAR RESULTS

Organism	Database	HSPs
Aedes albopictus	(Peptides) Foshan strain peptide sequences, AaloF1.1 geneset.	350

Checked Hits

Quick align | Pass to ClustalW | Download

☑ include query

Show Query/Hit Numbers

Hit	Gene Name	Description	Query	Aln Length	E-value	Score	Identity	Query Hit	DB Sequence Hit
AALF017696-PA		long wavelength sensitive opsin	AaGPRop2	372	0.0	2665	97.9%	>	>
AALF012989-PA			AaGPRop4	377	0.0	2650	96.6%	>	>
AALF012988-PA									

BLAST HSP Details

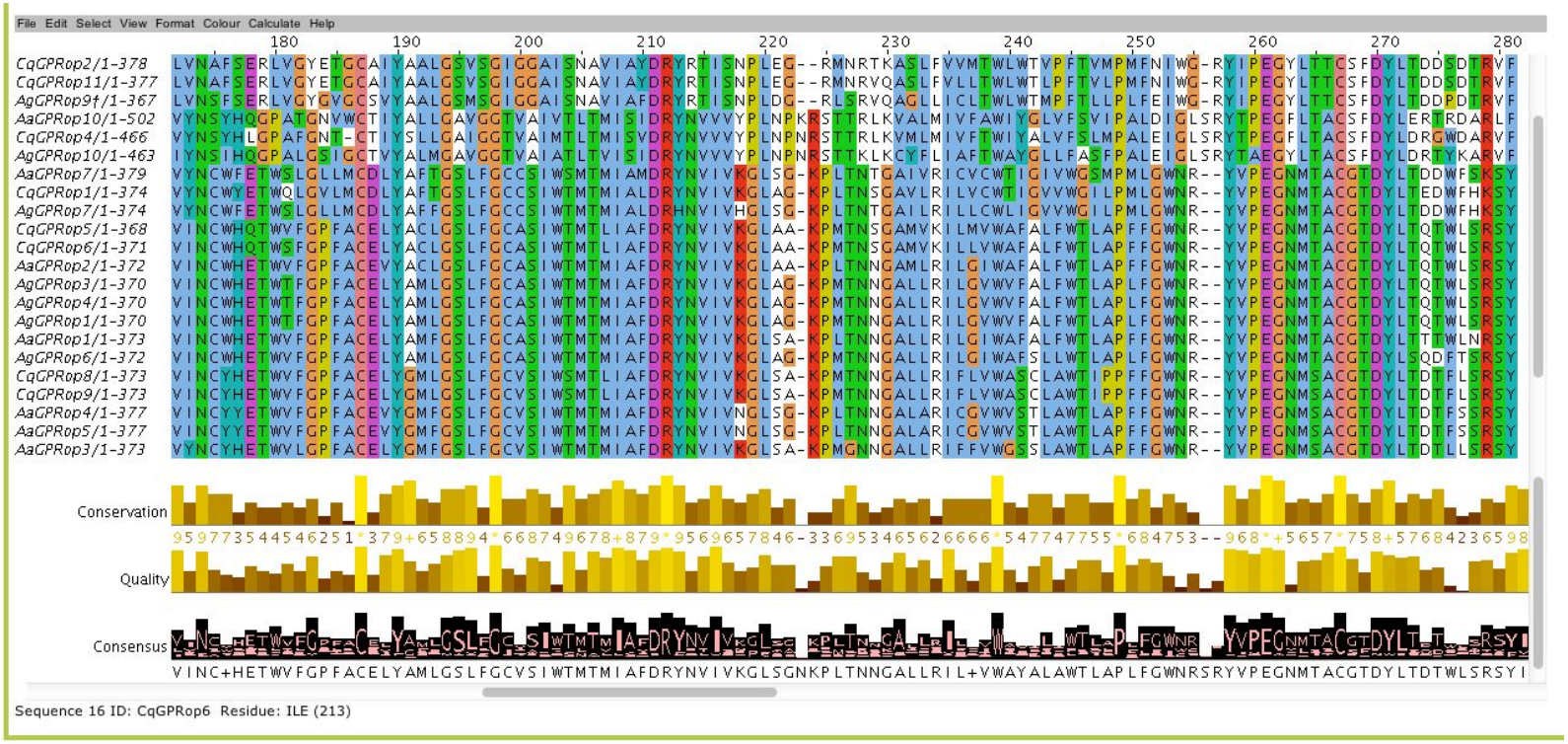
Browse Genome

```
Query: AaGPRop5
> AALF012988-PA
|protein_coding|JXUM01S000315:207503:208627:1|gene:AALF012988
Length = 374

Score = 2645
Expect = 0.0
Identity = 96.5517%
Strand = Plus/Plus
Aln Length = 377

Frame: 1 / 1


isTagged: , hitName: AALF012988-PA

Query 1    MASYGAWMAAQSAGHAVASNLTVVDRVPADMLHMVDAHWYQFPPMNPLWHSLLGFAIAVL 60
           MASYGAWMAAQSAG AVA+NLTVVDRVPADMLHMVDAHWYQFPPMNPLWHSLLGFAIAVL
Sbjct 1    MASYGAWMAAQSAG-AVATNLTVVDRVPADMLHMVDAHWYQFPPMNPLWHSLLGFAIAVL 59
```

Notice how for some genes the ‘Gene Name’ and ‘Description’ fields has no data

ClustalW

It is a multiple sequence alignment program. It can be used to generate the input for Hmmer.



hmmer

It looks for homolog genes. See previous slide for input

```
CLUSTAL 2.1 multiple sequence alignment


AaGPRop8            ----------------------------------------MPFEEHLSDNFTAVLRPEA
CqGPRop3            ----------------------------------------MPFLEHLSDNYTAVLRPEA
AgGPRop8            ---------------------------------------------MGLVQLDNQTAYRPEA
AaGPRop9_1          ----------------------------------------MFLLNETD---AAIFPMA
AaGPRop9_2†       ----------------------------------------MFLLNETD---AAIFPMA
CqGPRop2            ----------------------------------------MFLLNETD---AVLLPAA
CqGPRop11           ----------------------------------------MFLVNETD---AVLLPAA
AgGPRop9†         ----------------------------------------MFLGNESISEGAMLMPMA
AaGPRop10           MKLILFFSFHFTPPIIVRHSTATKTLIPKIDTRKLFANSQLCLSSCKRERLQSAVVLFKN
```

Job Control

Load results

Job ID

Add Description

Description

SUBMIT

RESET

Upload HMMer Input File

Browse... No file selected.

Want to save your options? Login or register here.

Parameters

Basic

Sequence Type

- Protein

Program

- phmmer
- hmmsearch

Cut-Offs

- E-value
- Bit Score

Cut-off values

Significance

Sequence	0.01
Hit	0.03

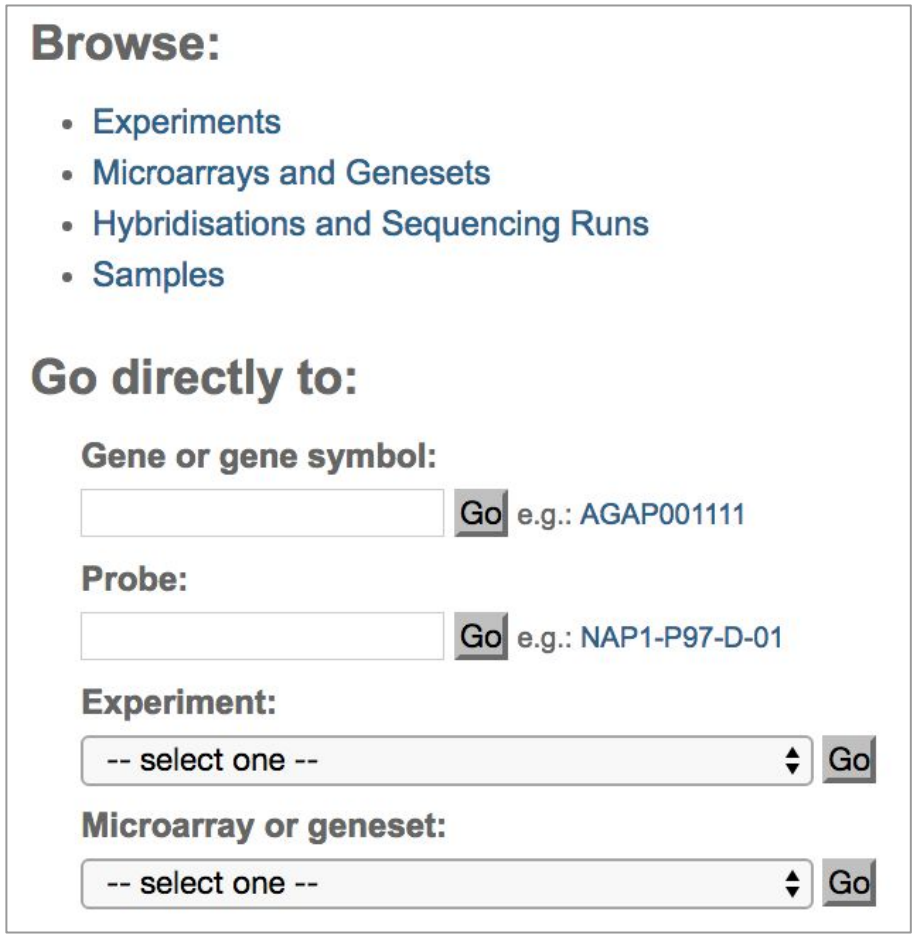
Datasets

- [ ] Select All Datasets
- [ ] **Peptides** Aedes aegypti, Liverpool strain, AaegL3.3 geneset.
- [x] **Peptides** Aedes albopictus, Foshan strain, AaloF1.1 geneset.
- [ ] **Peptides** Anopheles albimanus, STECLA strain, AalbS1.3 geneset.

```
Query:       135163.query  [M=399]
Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence       Description
    ------- ------ -----    ------- ------ -----   ---- --  --------       -----------
   5.5e-173  575.2  14.8   6.1e-173  575.1  10.3    1.0  1  AALF017696-PA long wavelength
     1e-172  574.3  18.1   1.1e-172  574.2  12.5    1.0  1  AALF009534-PA long wavelength
   2.4e-170  566.5  18.1   2.6e-170  566.4  12.6    1.0  1  AALF009531-PA long wavelength
```

Expression Browser

It hosts microarray and RNAseq data from multiple experiments that have been compared side by side with the same pipeline.



Expression data can also be queried using 'Search' and 'Advanced Search'

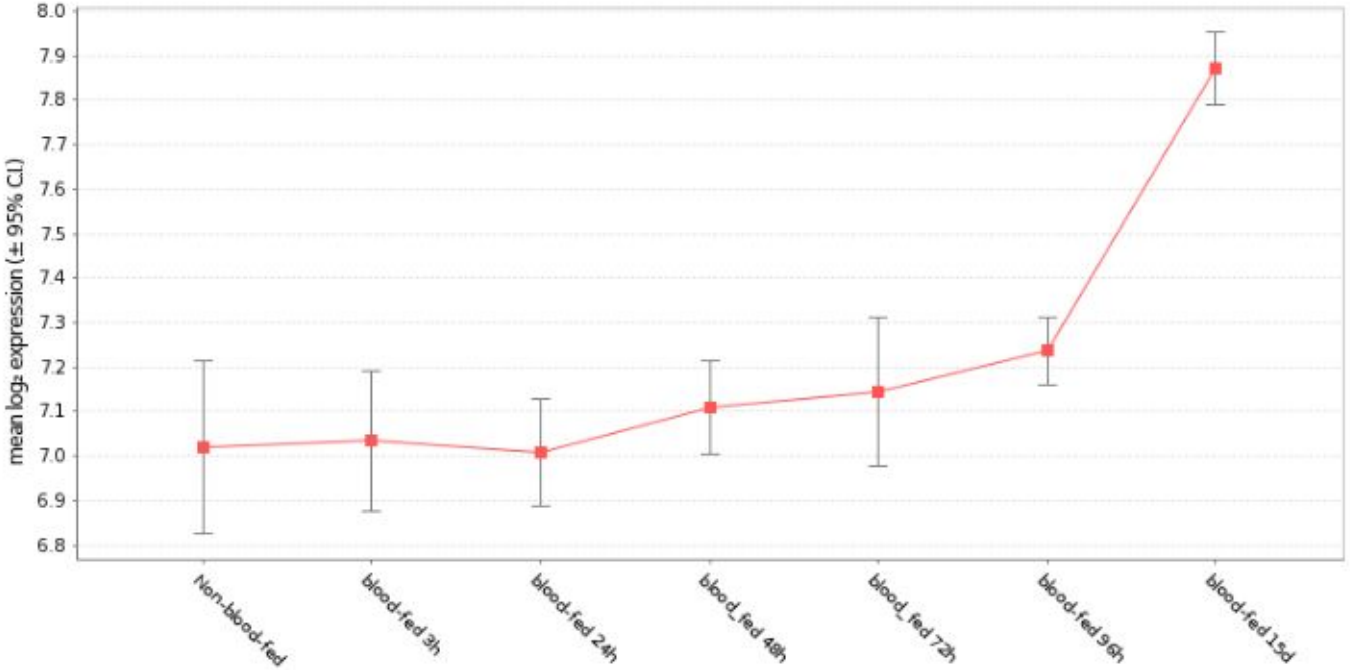
For each gene, under each experimental condition, there is a expression summary table, plots and other raw and process data.

Expression summary:

Experiment	P-value	Test	Experimental factor	Summary
Blood meal time series (Marinotti et al., 2006) Microarray experiment info	1.7e-10	ANOVA	Growth condition	Significant **differential expression** ↑ blood-fed 15d ↓ blood-fed 24h

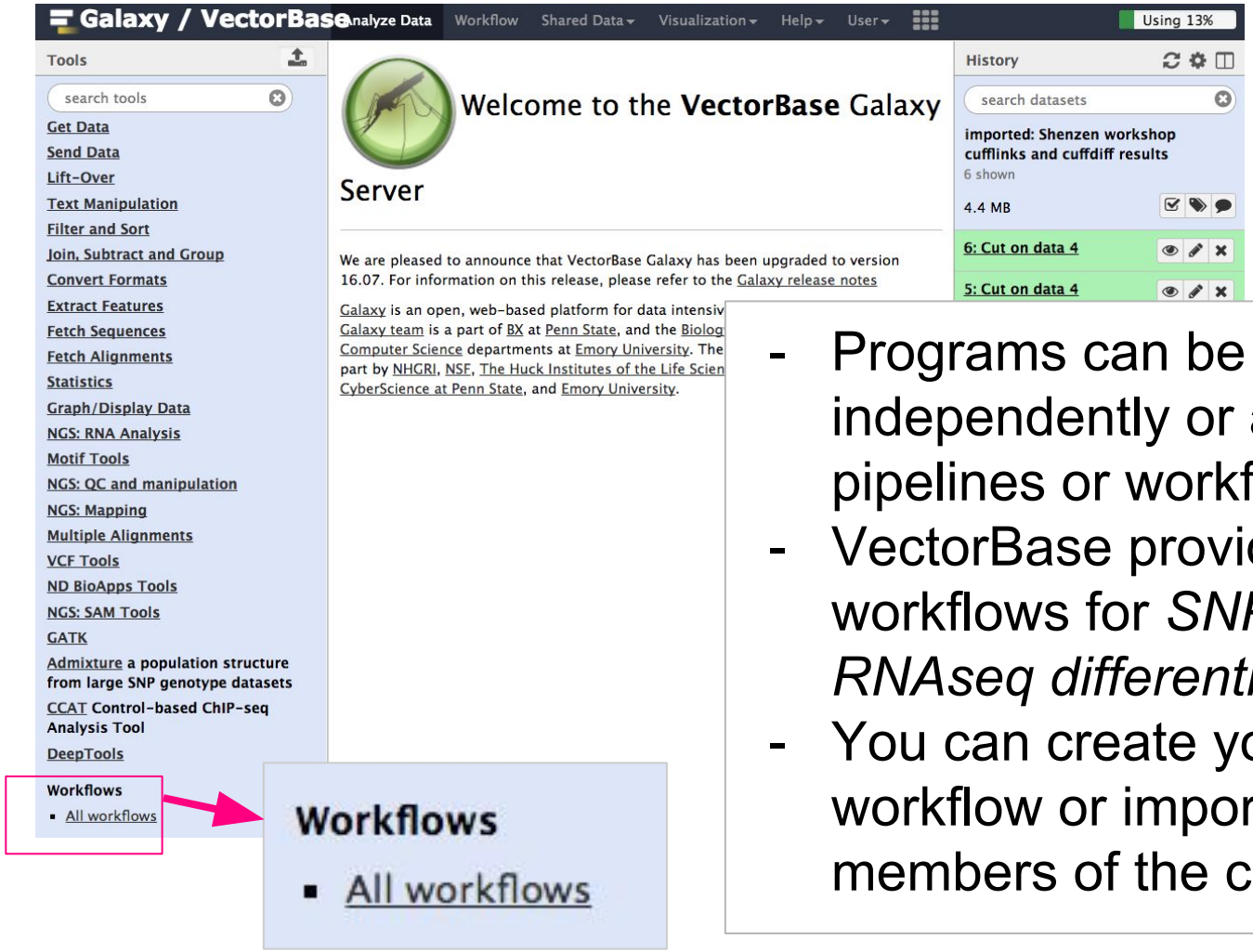
Click here for more results from Blood meal time series (Marinotti et al., 2006)

Plot for experimental factor 'GrowthCondition':



Galaxy

Is an open, web-based platform for data intensive biomedical research.



- Programs can be run independently or as part of pipelines or workflows
- VectorBase provides for you workflows for *SNP calling* and *RNAseq differential expression*
- You can create your own workflow or import one from other members of the community

Genome Browser

It makes the genomic data accessible

Aedes aegypti (AaegL3) ▼ | Location: supercont1.406:125,974-148,039 | Gene: TOLL11 | Transcript: TOLL11

Gene-based displays
- Summary
- Splice variants
- Transcript comparison
- Supporting evidence
- Sequence
- External references
- Regulation
- Expression report
- Ontologies
  - GO: Molecular function
  - GO: Cellular component
  - GO: Biological process
- Pathways
- PubMed
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Phenotype
- Genetic Variation
  - Variant table
  - Structural variants
  - Variant image
- External data
- ID History
  - Gene history

Configure this page

Add your data

Export data

Gene: TOLL11 AAEL009551

Description	Toll-like receptor
Location	SuperContig sup AaegL3:CH4775
About this gene	This gene has 1 t
Transcripts	Show transcript

Summary

Name	TOLL11 (VB Con
Gene type	Protein coding

erging Ensembl and TIGR prediction sets. The Ens
onerate model from a known protein or a set of alig
Exonerate models are further combined with availab
see V.Curwen et al., Genome Res. 2004 14:942-50
merHMM (M. Pertea), Genie (D. Kulp), and Twinsca
T genome alignments.

options (e.g. zooming)

annotation)

42.07 kb

120kb 130kb 140kb

The are four tabs at the top:
- Organism
- Location
- Gene (display here)
- Transcript (includes protein summary)

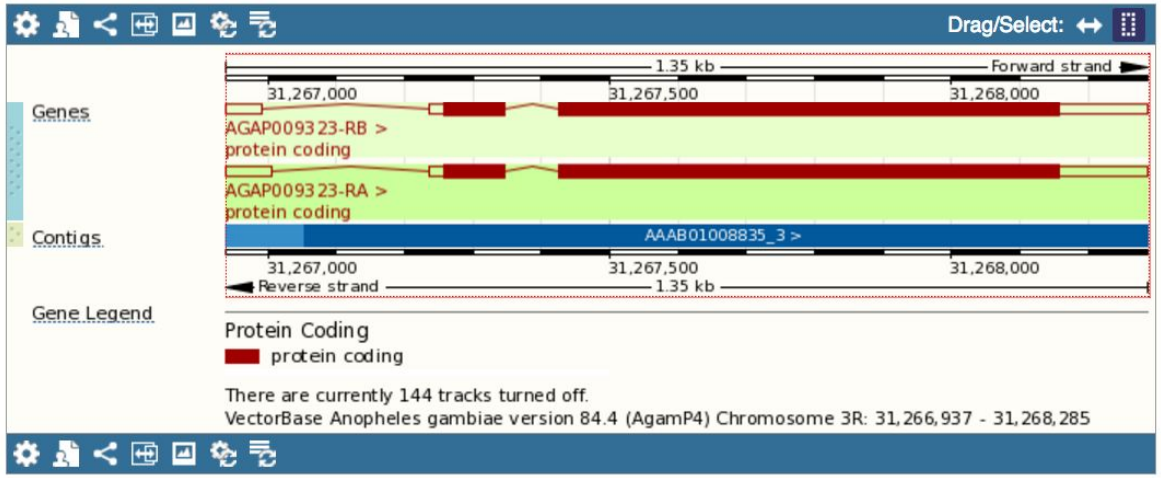
Different menu/links on the left hand side for each of the four tabs for data mining. These change the display in the center section (here in gray).

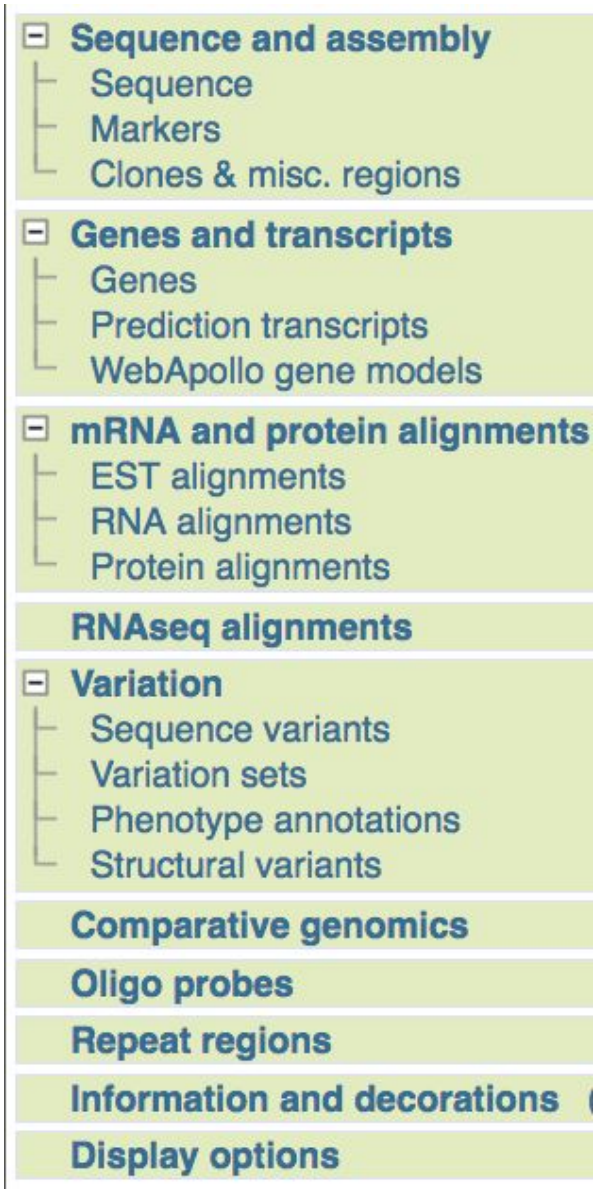
For SNPs click on ‘Variant table’ ---> ‘Variant ID’, to open a fifth top tab

Genome Browser

The image below shows the genes as display in the location tab. To activate the available features or tracks for a species, click on ‘configure this page’,

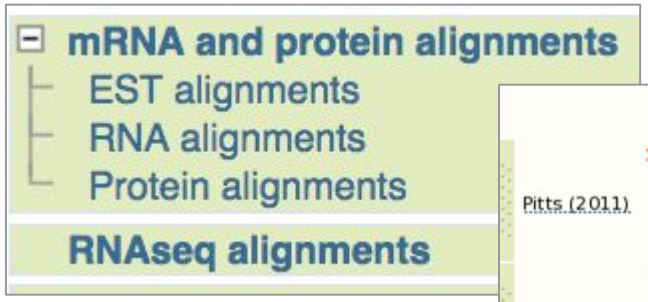
and select from a menu like the one on the right.

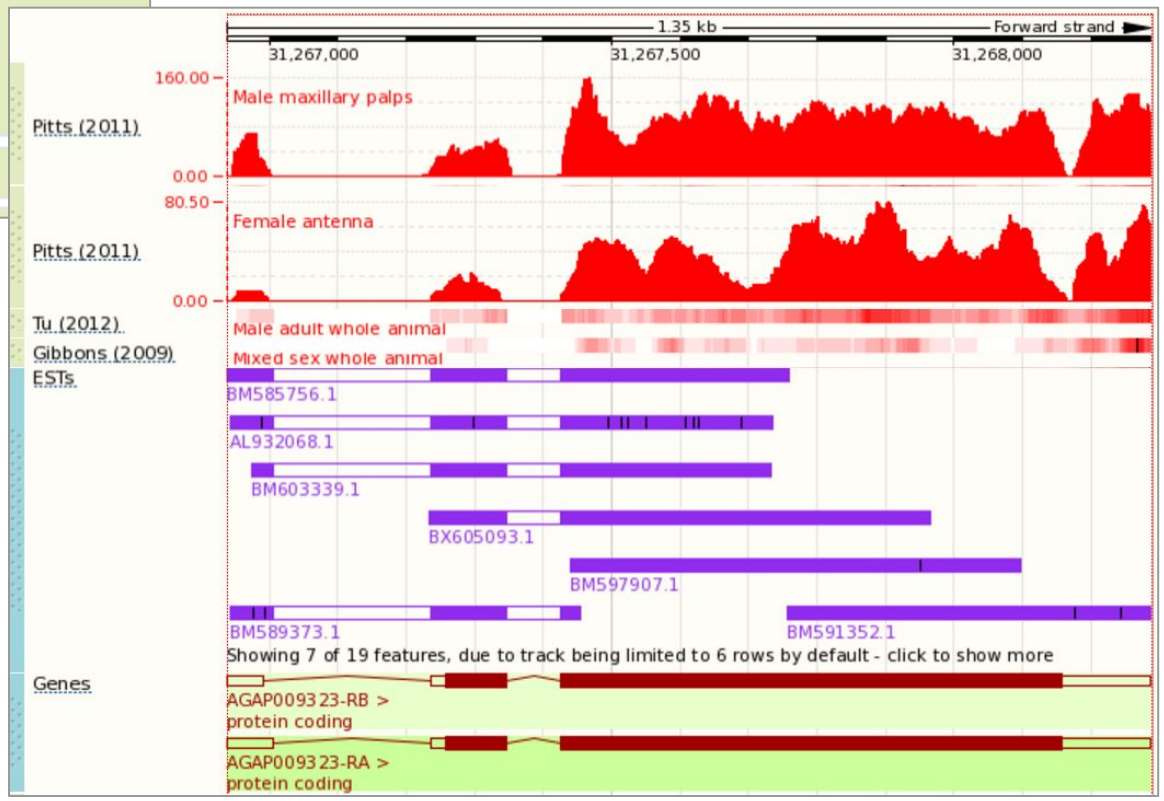




Genome Browser

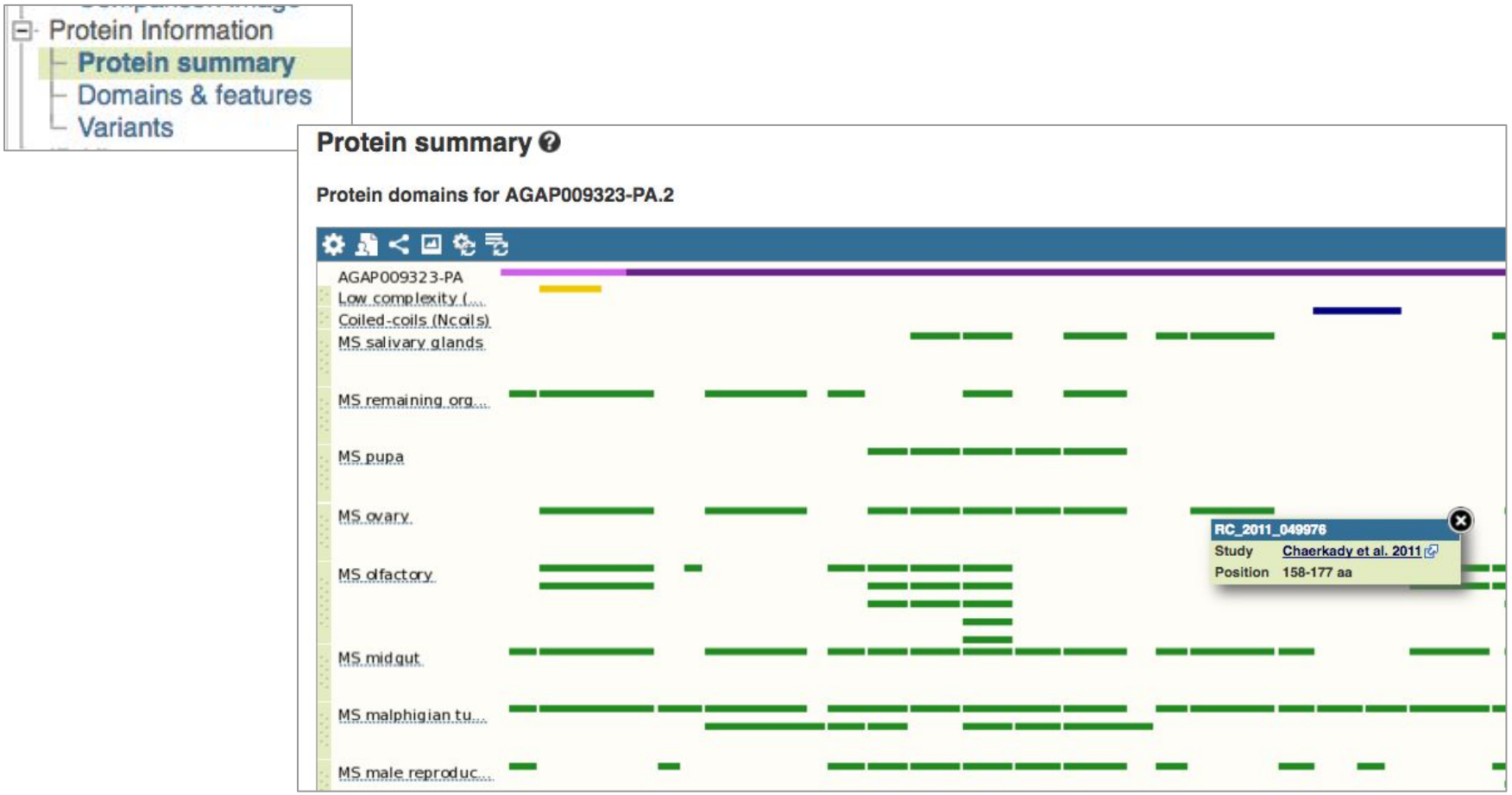
Use case: Experimental evidence, *e.g.*, Transcripts. ESTs and RNAseq reads can be activated as tracks in the location tab, with the “configure this page” link





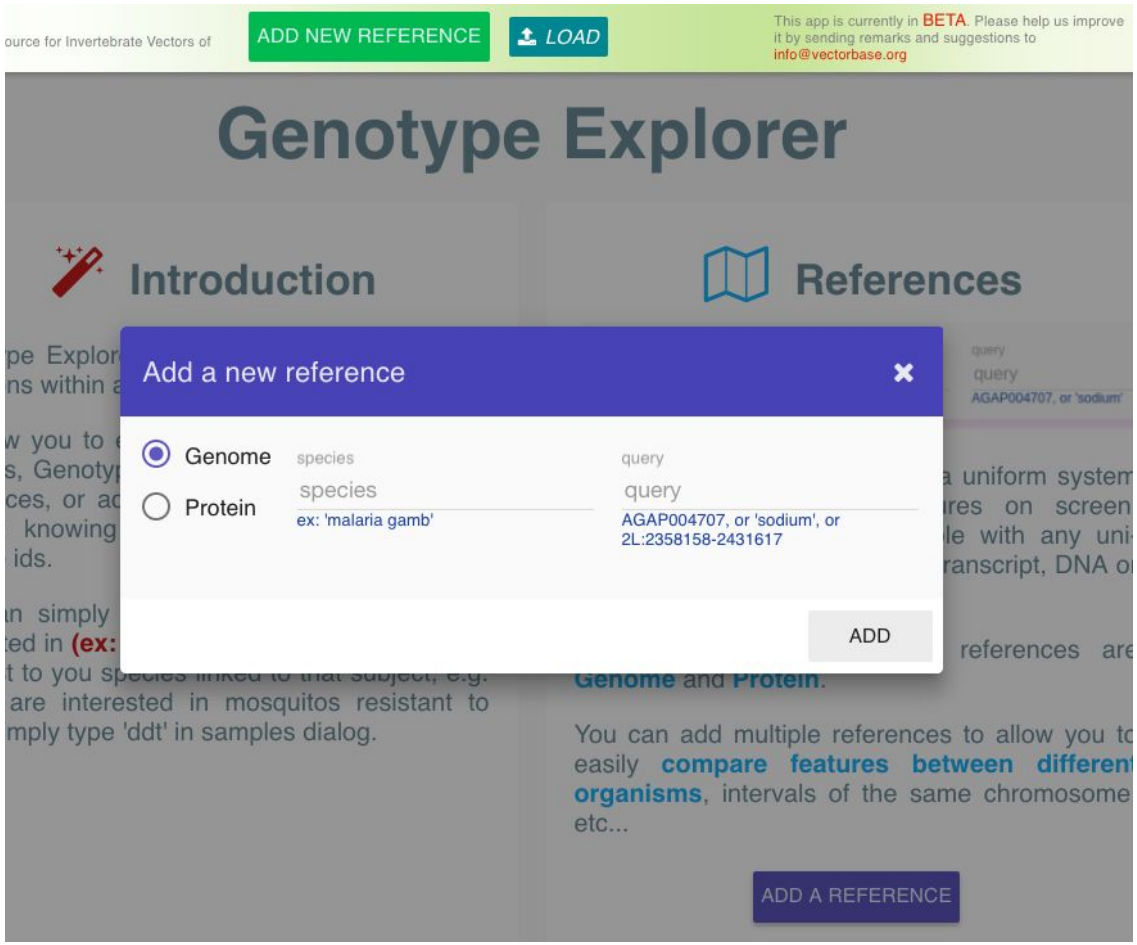
Genome Browser

Use case: Experimental evidence, e.g., Peptides. All are preloaded in the transcript tab, in the “protein summary” link. Click on each track for more details in the form of small pop out windows with links to the paper or study.



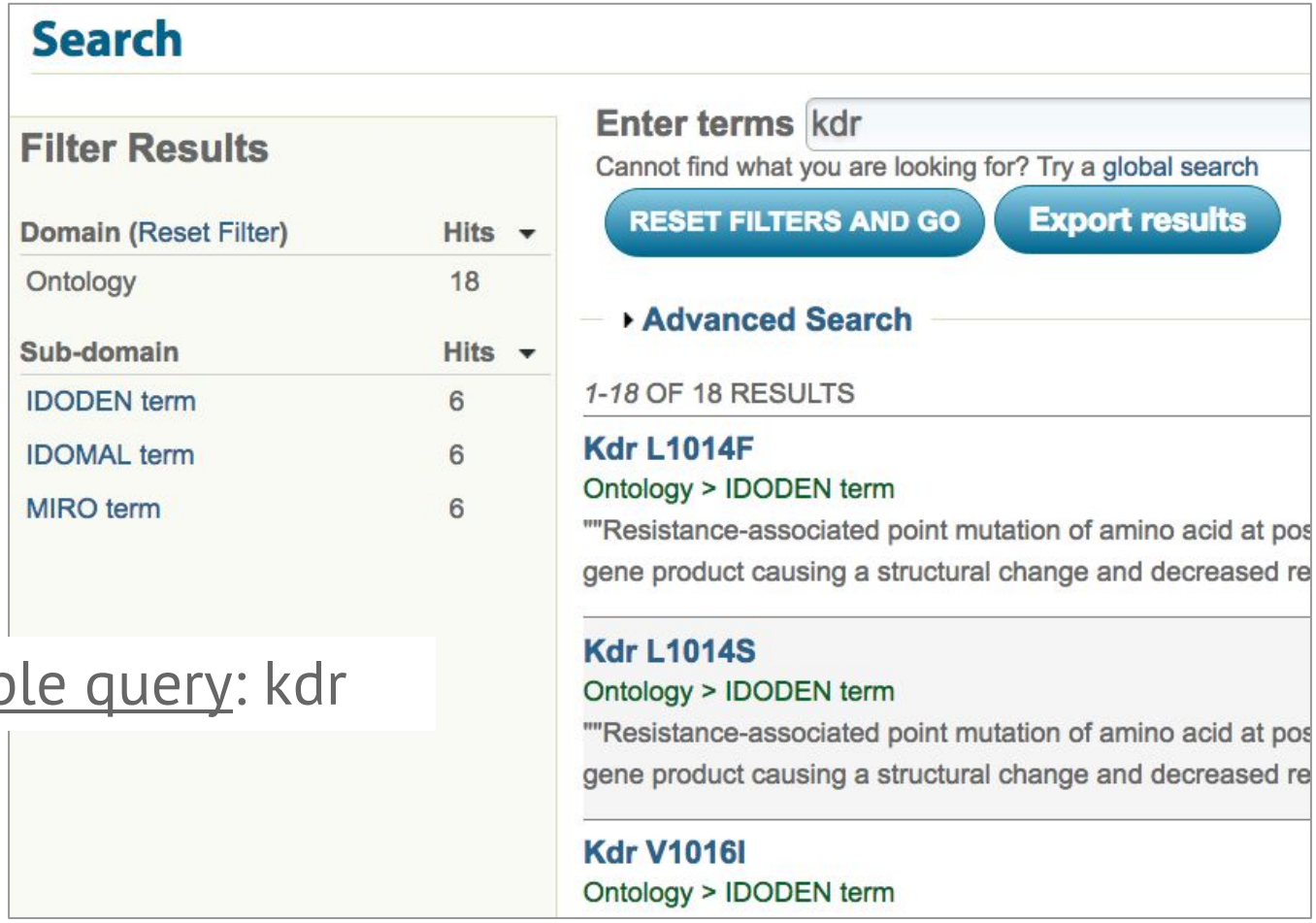
Genotype Explorer

Explore *variation* data associated with biological samples in genomic, protein and multi-species contexts.



Ontology Browser

It is a structural framework to organize information



Sample query: kdr

Ontology Browser

Please select an ontology:

Mosquito Insecticide Resistance

Search:

- object_aggregate
- object
  - biological material
  - chemical compound
    - insecticidal substance
    - protein
      - crystal protein
      - voltage gated sodium channel
        - Kdr L1014F
        - Kdr L1014S
        - Kdr I1011V
        - Kdr I1011M
        - Kdr V1016I
        - Kdr V1016G
      - acetylcholinesterase

Term information

ID	MIRO:00000125
Name	Kdr L1014S
Definition	Resistance-associated point mu position 1014 (L-S) in the insecti channel gene product causing a decreased response to the insec
Source	PMID:15242706
Comment	
Relationship(s)	*is_a* voltage gated sodium chann channel structural change

Domain

Ontology

Population Biology

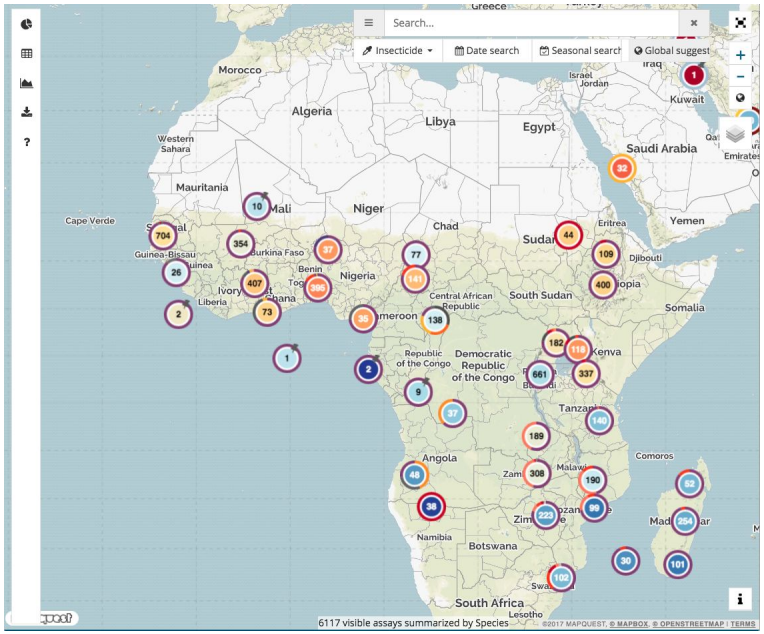
It integrates genomic, phenotypic and population data

Collection sites map

Insecticide resistance map

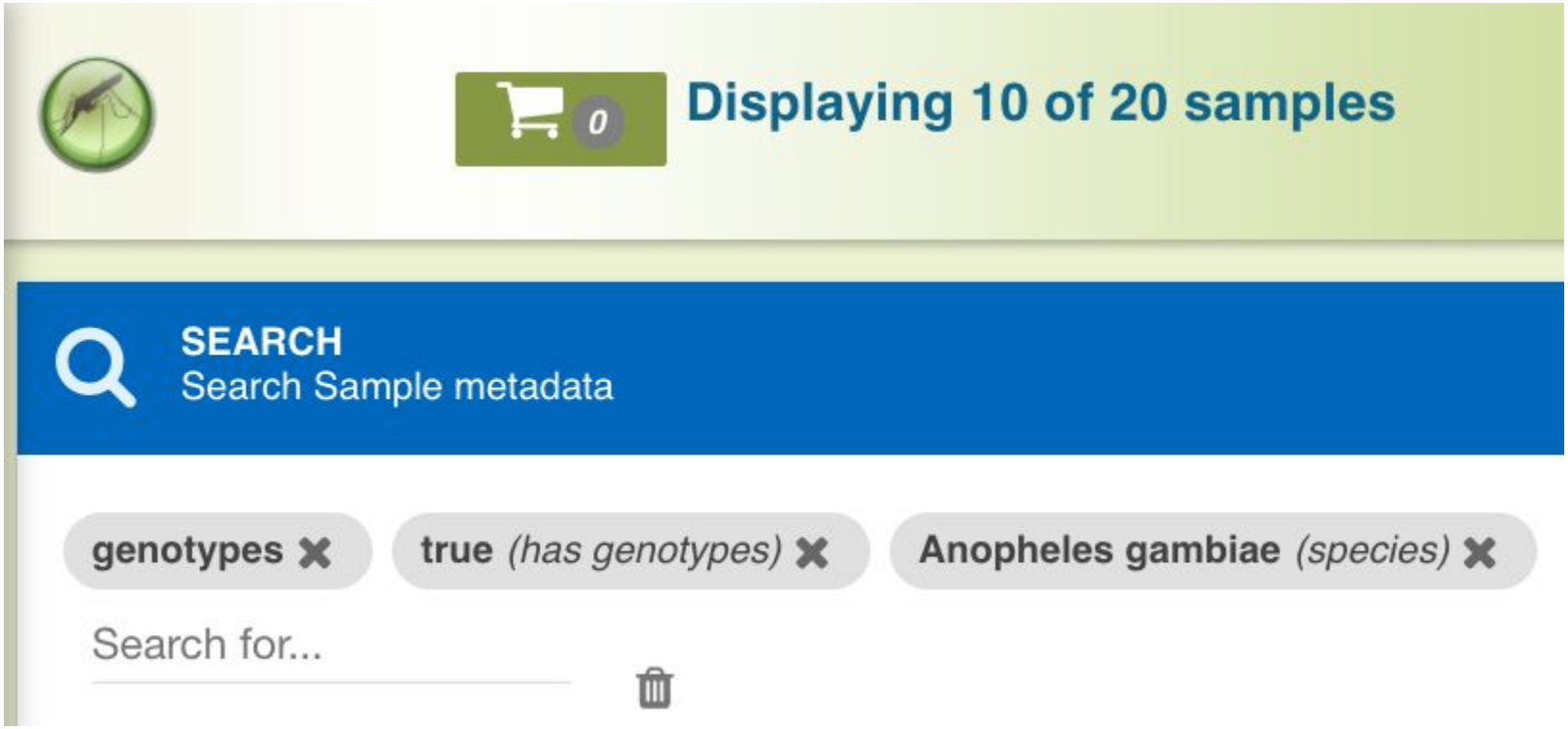
Search

Sample Explorer
NEW!



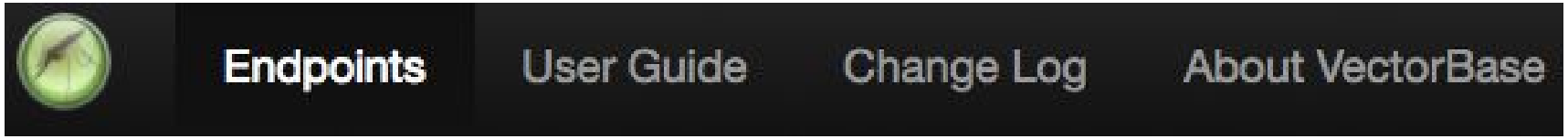
Sample Explorer

Search and explore *metadata* associated with biological samples and *display* them in the Genome Browser and PopBio (previous slide) applications



REST API

Direct programmatic access to VectorBase species data



- There is a front page for the REST API with a link to a user guide on it
- Keep in mind that many of the links go to Ensembl

How to search for more information or help?

E-mail us at

info@vectorbase.org