February 2019

RNAseq resources at VectorBase

Contents

Introduction

We have begun to alter the way that we handle RNAseq data. We now collect RNAseq metadata from the Short Read Archive (SRA) and combine it with information supplied to us from the community in a dedicated RNAseq database. This database allows us to curate and expose RNAseq experiments and short read data submitted to us in a unified manner across the site. All RNAseq data is now processed via a standardized short read alignment pipeline based on the HISAT sequence aligner, the core alignment engine in Tophat3 [1]. These alignments are made available via the genome browser and Web Apollo systems, allowing the community working on gene annotation to see the exact same data in both tools. We have also improved searching RNAseq metadata so that it should now be easier for users to locate studies, and link to the relevant resources. To achive these objectives we collect RNAseq metadata from the Short Read Archive (SRA) and combine it with information supplied to us from the community in a dedicated RNAseq database. This database allows us to curate and expose RNAseq experiments and short read data submitted to us in a unified manner across the site.

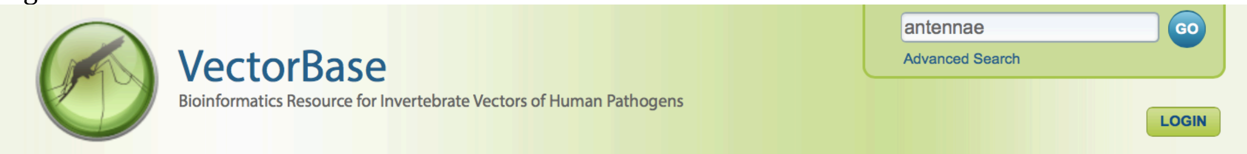
Producing a consensus set of alignment parameters for all RNAseq studies is a problematic task, as parameters will vary according to sampling depth, objective of the study, and the genetic similarity between the samples used and the reference sequences used for the alignment process. VectorBase therefore also accepts optimized RNAseq alignments direct from the community, and if supplied with the relevant alignment parameters we can also integrate these optimizations into our standard pipeline. This allows RNAseq alignments to be kept in synchronization with changes to genome assemblies and gene set models, which helps keep studies relevant to the needs of other researchers.

VectorBase is using UCSC track hubs to expose the Short read Archive (SRA) RNAseq experiments (see Figure 5). In the SRA a study contains a list of experiments, each with their own associated sequencing run and sample information. We use the information about the samples and runs to group replicates and perform sequence alignment back to the relevant reference sequences. The data from one or more sequencing lanes may be combined into a track that can be displayed within the VectorBase genome browser. When you explore the RNAseq data present in VectorBase you will see the terms tracks and track groups used – this is because some studies may have multiple experiments (such as a time course looking at expression in multiple different tissues). Where there are a large number of individual lanes of data we may group multiple individual tracks into an RNAseq track group to make it more convenient for the end user to select and manipulate the expression information within the genome browser.

Locating RNAseq data via site search

RNAseq data is now integrated into the VectorBase site search, which is present at the top of all pages. If you want to explore all of the RNAseq expression studies available at VectorBase use "rnaseq"as your search term, otherwise type into the search entry box any term such as the name of a tissue, experiment, or sequence accession (see Figure 1).

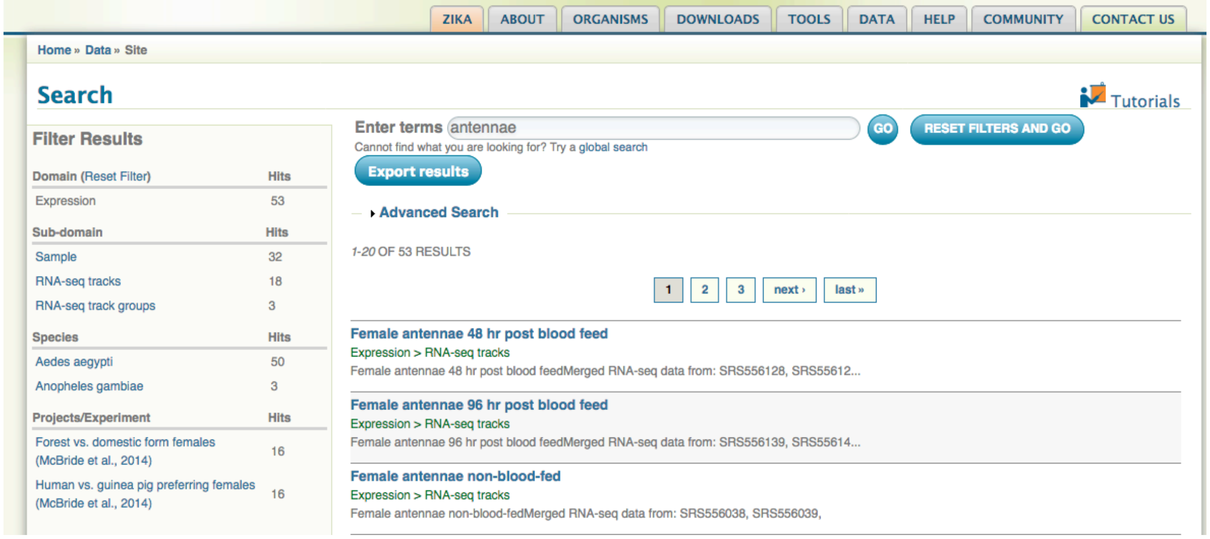
Figure 1. VectorBase site search.



VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens
antennae
GO
Advanced Search
LOGIN

If you then select the "expression" domain from the filter results section you can explore the records retrieved in more detail, and further filter them by species, projects, experiments or samples as desired.

Figure 2. Exploring site search expression data.



Search

Filter Results

Domain (Reset Filter)	Hits
Expression	53
Sub-domain	**Hits**
Sample	32
RNA-seq tracks	18
RNA-seq track groups	3
Species	**Hits**
Aedes aegypti	50
Anopheles gambiae	3
Projects/Experiment	**Hits**
Forest vs. domestic form females (McBride et al., 2014)	16
Human vs. guinea pig preferring females (McBride et al., 2014)	16

Enter terms: antennae

Cannot find what you are looking for? Try a global search

1-20 OF 53 RESULTS

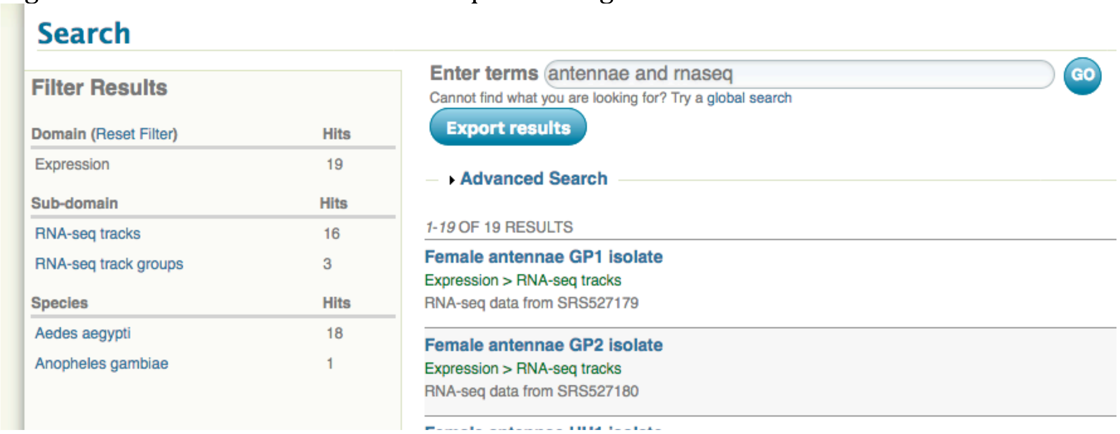
Female antennae 48 hr post blood feed
Expression > RNA-seq tracks
Female antennae 48 hr post blood feedMerged RNA-seq data from: SRS556128, SRS55612...

Female antennae 96 hr post blood feed
Expression > RNA-seq tracks
Female antennae 96 hr post blood feedMerged RNA-seq data from: SRS556139, SRS55614...

Female antennae non-blood-fed
Expression > RNA-seq tracks
Female antennae non-blood-fedMerged RNA-seq data from: SRS556038, SRS556039,

You can also perform more advanced searches using simple Boolean terms such as “and” / “or” as described in the search tutorial (Figure 3).

Figure 3. Advanced search for RNAseq data using Boolean terms.



Search

Filter Results

Domain (Reset Filter)	Hits
Expression	19
Sub-domain	**Hits**
RNA-seq tracks	16
RNA-seq track groups	3
Species	**Hits**
Aedes aegypti	18
Anopheles gambiae	1

Enter terms: antennae and rnaseq

Cannot find what you are looking for? Try a global search

1-19 OF 19 RESULTS

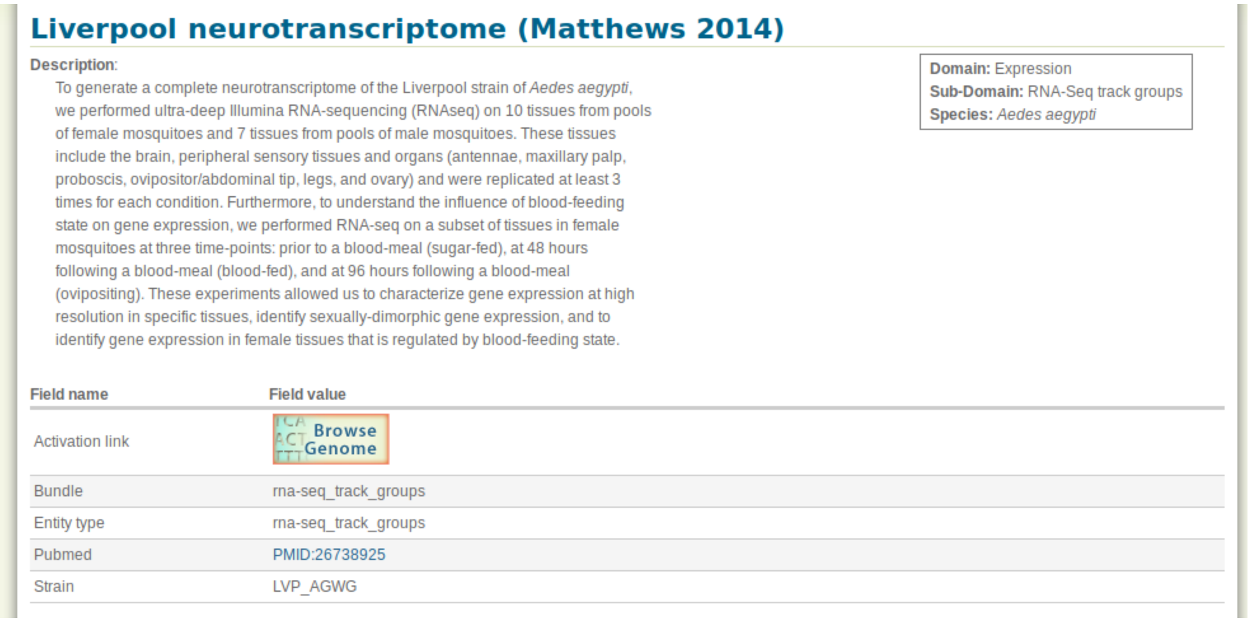
Female antennae GP1 isolate
Expression > RNA-seq tracks
RNA-seq data from SRS527179

Female antennae GP2 isolate
Expression > RNA-seq tracks
RNA-seq data from SRS527180

Once you have isolated a list of interesting RNAseq studies you can gain further information by clicking on the relevant result which will display a short description of the study and its tracks (Figures 4 and 5), and a link allowing you to activate the group of tracks as a track hub for this study in the genome browser (Figure 6).

For each track of the study, the page also provides a detailed list of the SRA accessions used, activation and download links for the bam and bigwig files, as well as the list of commands used to generate those files, and a link to the reads coverage of every gene generated with HTSeq-count (Figure 5).

Figure 4. Summary of an RNAseq study showing the activation URL for this study in the genome browser.



Liverpool neurotranscriptome (Matthews 2014)

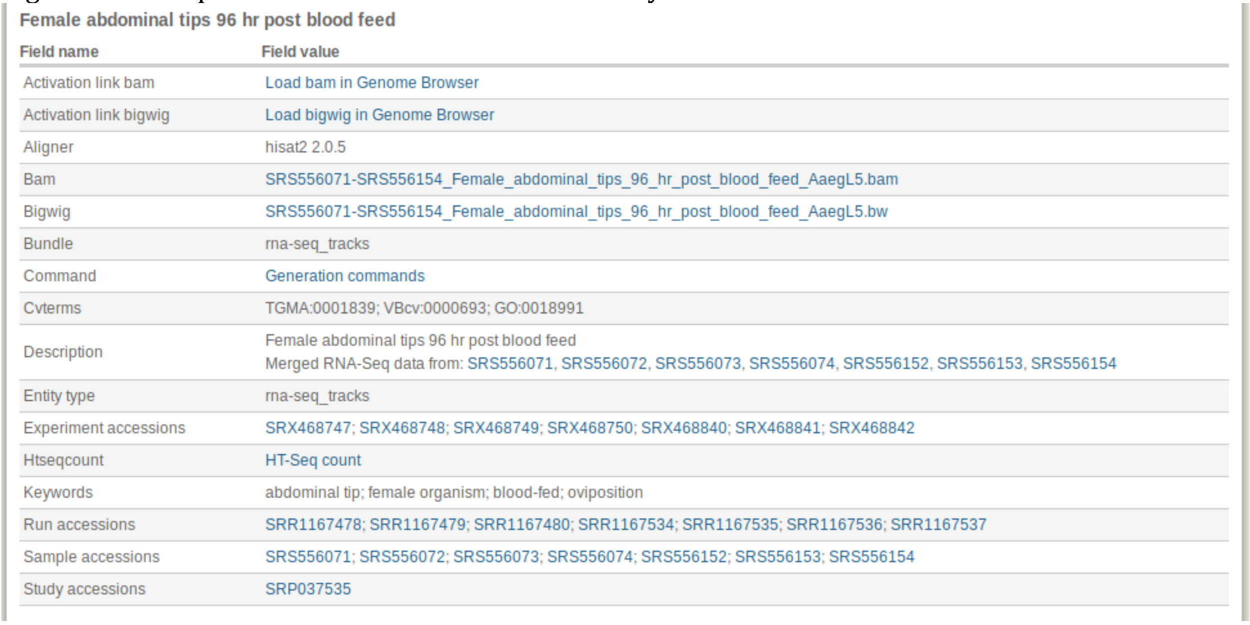
Description:

To generate a complete neurotranscriptome of the Liverpool strain of *Aedes aegypti*, we performed ultra-deep Illumina RNA-sequencing (RNAseq) on 10 tissues from pools of female mosquitoes and 7 tissues from pools of male mosquitoes. These tissues include the brain, peripheral sensory tissues and organs (antennae, maxillary palp, proboscis, ovipositor/abdominal tip, legs, and ovary) and were replicated at least 3 times for each condition. Furthermore, to understand the influence of blood-feeding state on gene expression, we performed RNA-seq on a subset of tissues in female mosquitoes at three time-points: prior to a blood-meal (sugar-fed), at 48 hours following a blood-meal (blood-fed), and at 96 hours following a blood-meal (ovipositing). These experiments allowed us to characterize gene expression at high resolution in specific tissues, identify sexually-dimorphic gene expression, and to identify gene expression in female tissues that is regulated by blood-feeding state.

Domain: Expression
Sub-Domain: RNA-Seq track groups
Species: *Aedes aegypti*

Field name	Field value
Activation link	Browse Genome
Bundle	rna-seq_track_groups
Entity type	rna-seq_track_groups
Pubmed	PMID:26738925
Strain	LVP_AGWG

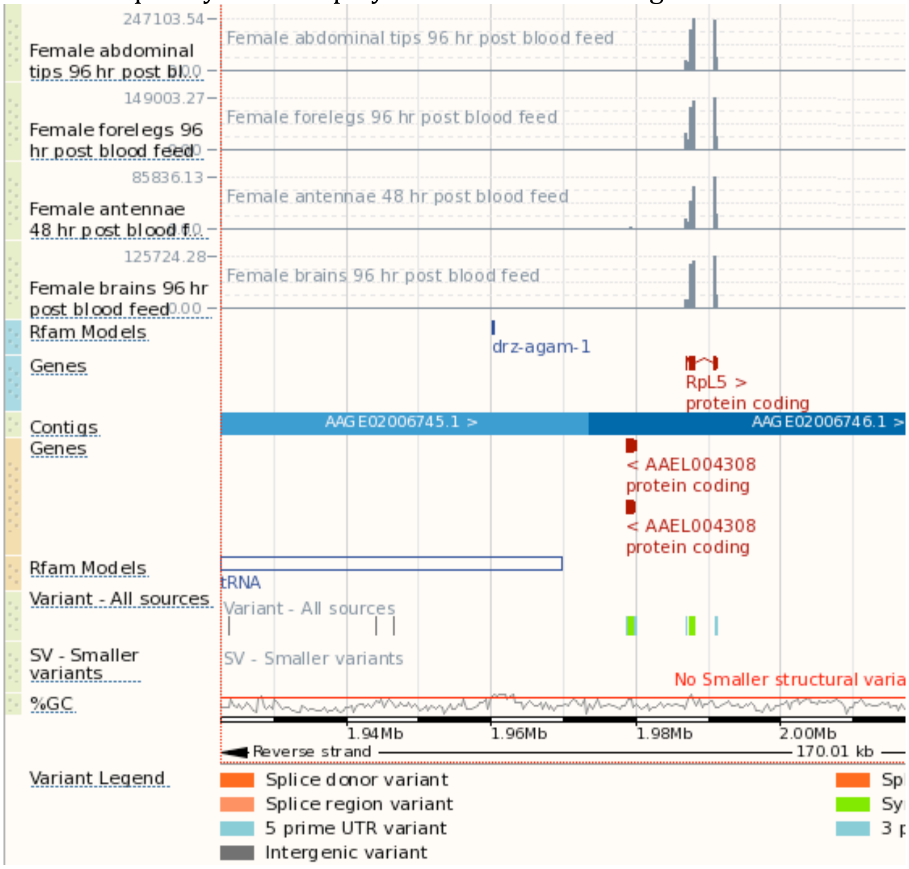
Figure 5. Description of one of the tracks of the study.



Female abdominal tips 96 hr post blood feed

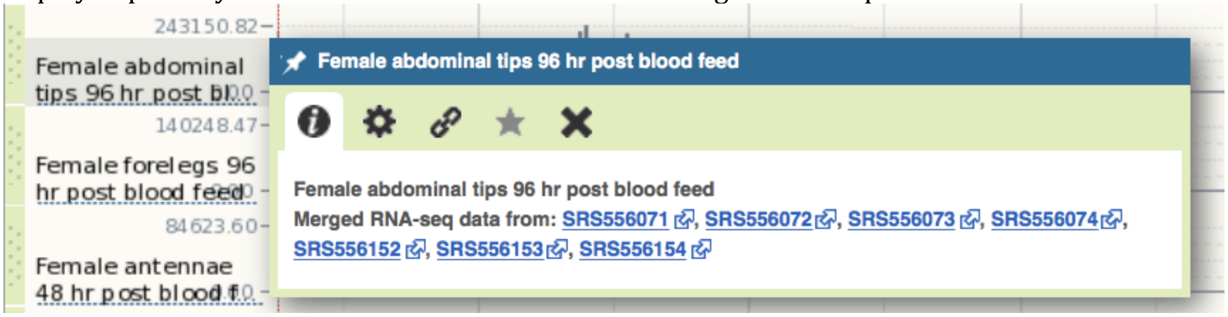
Field name	Field value
Activation link bam	Load bam in Genome Browser
Activation link bigwig	Load bigwig in Genome Browser
Aligner	hisat2 2.0.5
Bam	SRS556071-SRS556154_Female_abdominal_tips_96_hr_post_blood_feed_AaegL5.bam
Bigwig	SRS556071-SRS556154_Female_abdominal_tips_96_hr_post_blood_feed_AaegL5.bw
Bundle	rna-seq_tracks
Command	Generation commands
Cvterms	TGMA:0001839; VBcv:0000693; GO:0018991
Description	Female abdominal tips 96 hr post blood feed Merged RNA-Seq data from: SRS556071, SRS556072, SRS556073, SRS556074, SRS556152, SRS556153, SRS556154
Entity type	rna-seq_tracks
Experiment accessions	SRX468747; SRX468748; SRX468749; SRX468750; SRX468840; SRX468841; SRX468842
Htseqcount	HT-Seq count
Keywords	abdominal tip; female organism; blood-fed; oviposition
Run accessions	SRR1167478; SRR1167479; SRR1167480; SRR1167534; SRR1167535; SRR1167536; SRR1167537
Sample accessions	SRS556071; SRS556072; SRS556073; SRS556074; SRS556152; SRS556153; SRS556154
Study accessions	SRP037535

Figure 6. RNAseq study track displayed in the VectorBase genome browser.



The RNAseq tracks can be manipulated just like any other track in the browser. Each track can be turned on/off, can be linked to via an URL, and displays information about the sequencing lanes used to construct it (just place your cursor over the lane description and a dialog box showing this information should appear as shown in Figure 7).

Figure 7. Display of primary SRA lane lane data used in constructing an RNAseq track.

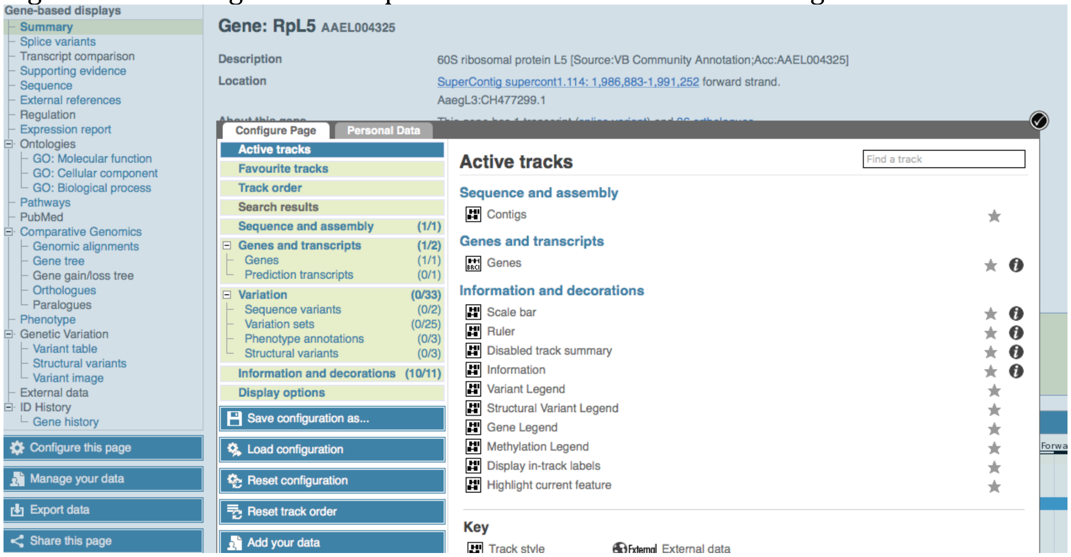

243150.82
Female abdominal tips 96 hr post bl
0.0
140248.47
Female forelegs 96 hr post blood feed
0.0
84623.60
Female antennae 48 hr post blood f
0.0
Female abdominal tips 96 hr post blood feed

Female abdominal tips 96 hr post blood feed
Merged RNA-seq data from: SRS556071, SRS556072, SRS556073, SRS556074, SRS556152, SRS556153, SRS556154

Locating RNAseq data from within the genome browser.

It is also possible to search for RNAseq studies from within the genome browser making use of the trackhub registry (trackhubregistry.org) to search across all UCSC track hubs for a species. To search for and activate RNAseq tracks in the genome browser on any region view simply click on the “Configure this page” icon in the region view (Figure 8).

Figure 8. Searching for RNAseq data from within the VectorBase genome browser.



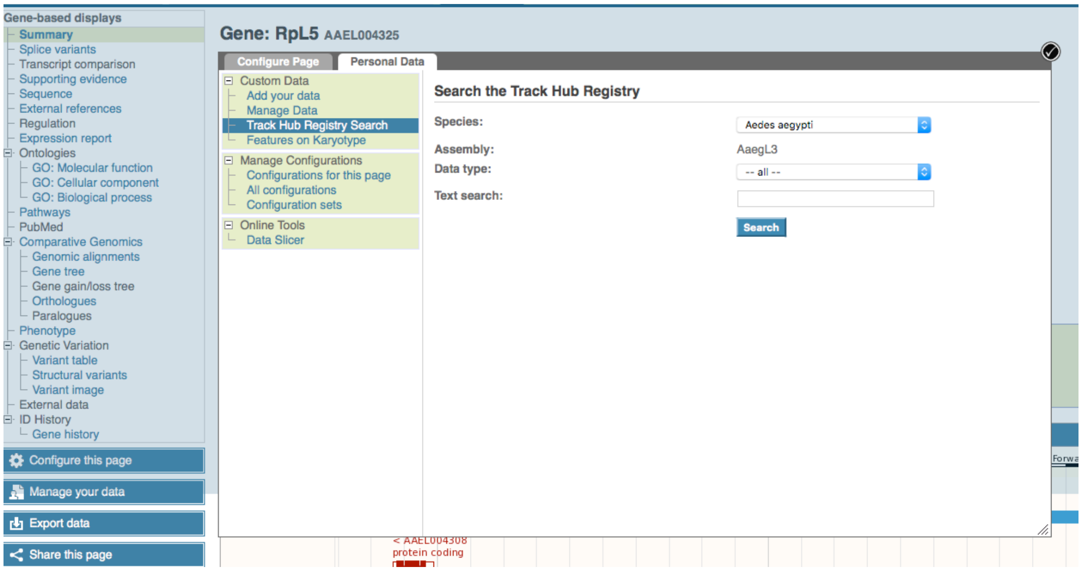
Gene: RpL5 AAEL004325

Description 60S ribosomal protein L5 [Source:VB Community Annotation;Acc:AAEL004325]

Location SuperContig supercont1.114: 1,986,883-1,991,252 forward strand.
AaegL3:CH477299.1

This brings up a dialog box showing available tracks which by default shows the existing VectorBase tracks for a species – to search the track hub registry for new data click on the tab labelled “Personal data” and select the “Track Hub Registry Search” option (Figure 9).

Figure 9. Searching the track hub registry.



Gene: RpL5 AAEL004325

Search the Track Hub Registry

Species: Aedes aegypti

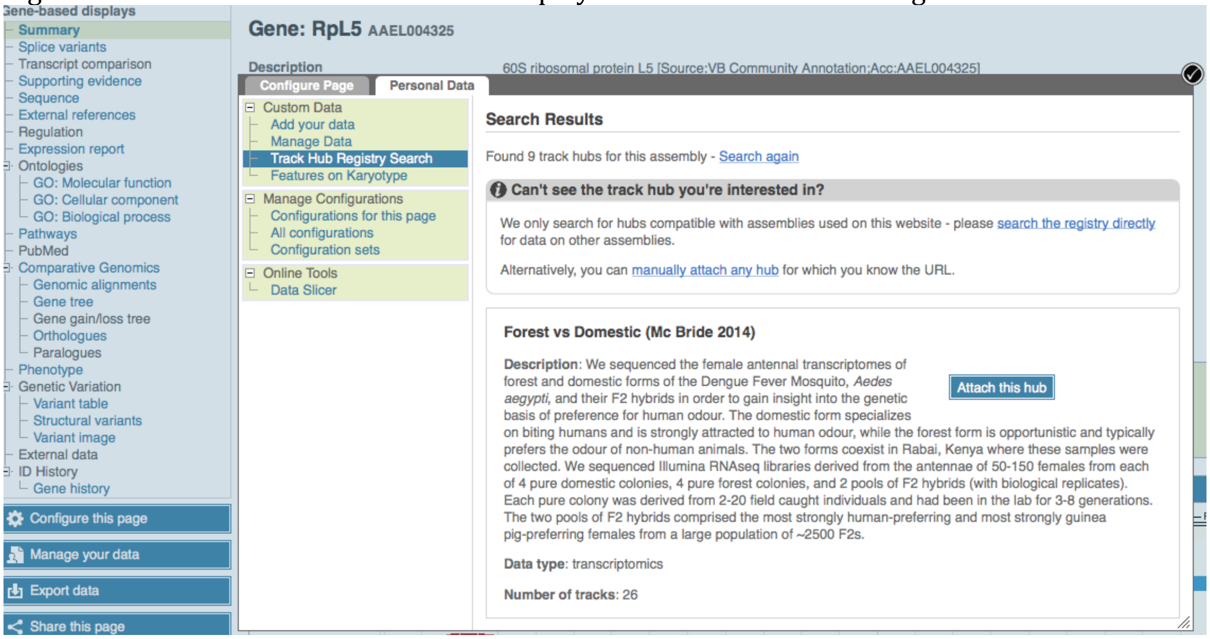
Assembly: AaegL3

Data type: -- all --

Text search:

By default the track hub registry species will be automatically set to the species you were viewing in the genome browser, and the text search field will be blank. If you wish to view all available track hubs for the species just leave the search field blank and click on the "search" button, otherwise type in a search term and press "search". The search results will be presented in the same window , and to select one for display in the genome browser just select the "Attach this hub" option (Figure 10).

Figure 10. Results of a track hub search displayed within the VectorBase genome browser.



Gene: RpL5 AAEL004325

Description 60S ribosomal protein L5 [Source:VB Community Annotation;Acc:AAEL004325]

Search Results

Found 9 track hubs for this assembly - Search again

Can't see the track hub you're interested in?

We only search for hubs compatible with assemblies used on this website - please search the registry directly for data on other assemblies.

Alternatively, you can manually attach any hub for which you know the URL.

Forest vs Domestic (Mc Bride 2014)

Description: We sequenced the female antennal transcriptomes of forest and domestic forms of the Dengue Fever Mosquito, *Aedes aegypti*, and their F2 hybrids in order to gain insight into the genetic basis of preference for human odour. The domestic form specializes on biting humans and is strongly attracted to human odour, while the forest form is opportunistic and typically prefers the odour of non-human animals. The two forms coexist in Rabai, Kenya where these samples were collected. We sequenced Illumina RNAseq libraries derived from the antennae of 50-150 females from each of 4 pure domestic colonies, 4 pure forest colonies, and 2 pools of F2 hybrids (with biological replicates). Each pure colony was derived from 2-20 field caught individuals and had been in the lab for 3-8 generations. The two pools of F2 hybrids comprised the most strongly human-preferring and most strongly guinea pig-preferring females from a large population of ~2500 F2s.

Data type: transcriptomics

Number of tracks: 26

Your selected RNAseq data should be automatically displayed in the browser as a set of tracks in the region view display (see Figure 6).

Feedback and help.

We are always interested in receiving feedback or providing help. Please contact us either via the website contact page

https://www.vectorbase.org/contact

or email is directly to

info@vectorbase.org

Cited Literature

1. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357-360.