Last update: 12/27/2021

Motif Discovery in *Drosophila*

Wilson Leung

Prerequisites

Annotation of Transcription Start Sites in *Drosophila*

Resources

Website	Web Address
FlyBase	https://flybase.org/
The *MEME* Suite	https://meme-suite.org/
Fly Factor Survey	https://mccb.umassmed.edu/ffs/

Files for this walkthrough

The package containing the files for this walkthrough are available through the "Motif Discovery in *Drosophila*" page on the GEP website.

Introduction

Interactions between proteins and DNA are one of the main mechanisms for regulating chromosome function and gene expression. A subset of DNA binding proteins, including the transcription factors, exhibits sequence-specific affinity. Consequently, short conserved motifs are often found near transcription start sites (TSS), corresponding to sites where transcription factors bind to the DNA. These interactions generally help regulate the expression of nearby genes, often in a tissue or developmental time point specific manner.

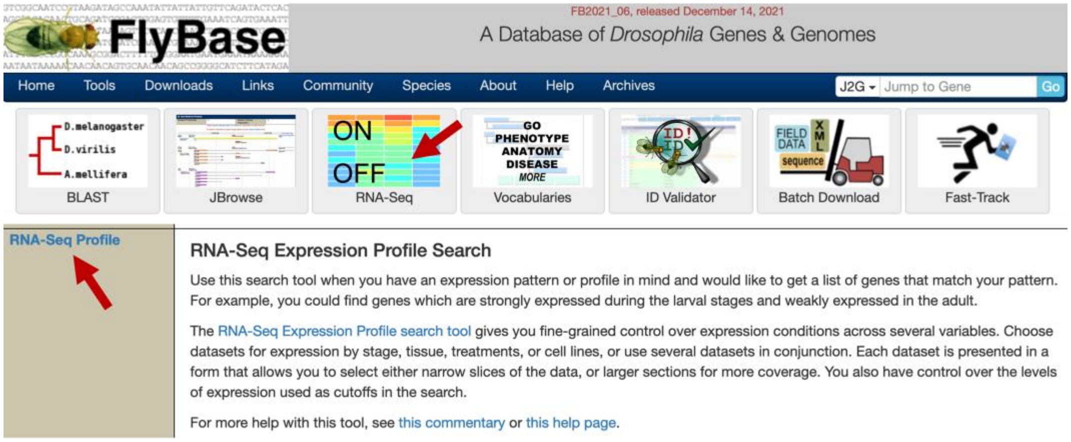
In this walkthrough, we will try to identify conserved motifs upstream of a group of the dot chromosome (also known as the Muller F element) genes in *Drosophila melanogaster* using publicly available *Drosophila* databases and genome analysis tools. Identifying such motifs will allow us to look for unique features in the regulation of dot chromosome genes. These genes function in an unfavorable (heterochromatic) environment, and so may require additional signals (multiple copies, an enhanced cluster, etc.) for activation.

Identify genes with similar expression patterns

The first step in our analysis is to identify a set of genes in *D. melanogaster* with similar gene expression patterns; these genes are more likely to be under the control of a similar set of transcription factors. The modENCODE project has previously generated a large set of high-throughput expression (RNA-Seq) data at different developmental stages, tissues, and cell lines. FlyBase has curated these RNA-Seq datasets and we can use the RNA-Seq search tools at FlyBase to identify genes with similar expression profiles.

Use FlyBase RNA-Seq tools to identify genes with similar expression patterns

Open a web browser and navigate to the FlyBase website at https://flybase.org. Click on the "**RNA-Seq**" image on the FlyBase home page and then click on the "**RNA-Seq Profile**" link (Figure 1) to access the "RNA-Seq Expression Profile Search" tool.



RNA-Seq Expression Profile Search

Use this search tool when you have an expression pattern or profile in mind and would like to get a list of genes that match your pattern. For example, you could find genes which are strongly expressed during the larval stages and weakly expressed in the adult.

The RNA-Seq Expression Profile search tool gives you fine-grained control over expression conditions across several variables. Choose datasets for expression by stage, tissue, treatments, or cell lines, or use several datasets in conjunction. Each dataset is presented in a form that allows you to select either narrow slices of the data, or larger sections for more coverage. You also have control over the levels of expression used as cutoffs in the search.

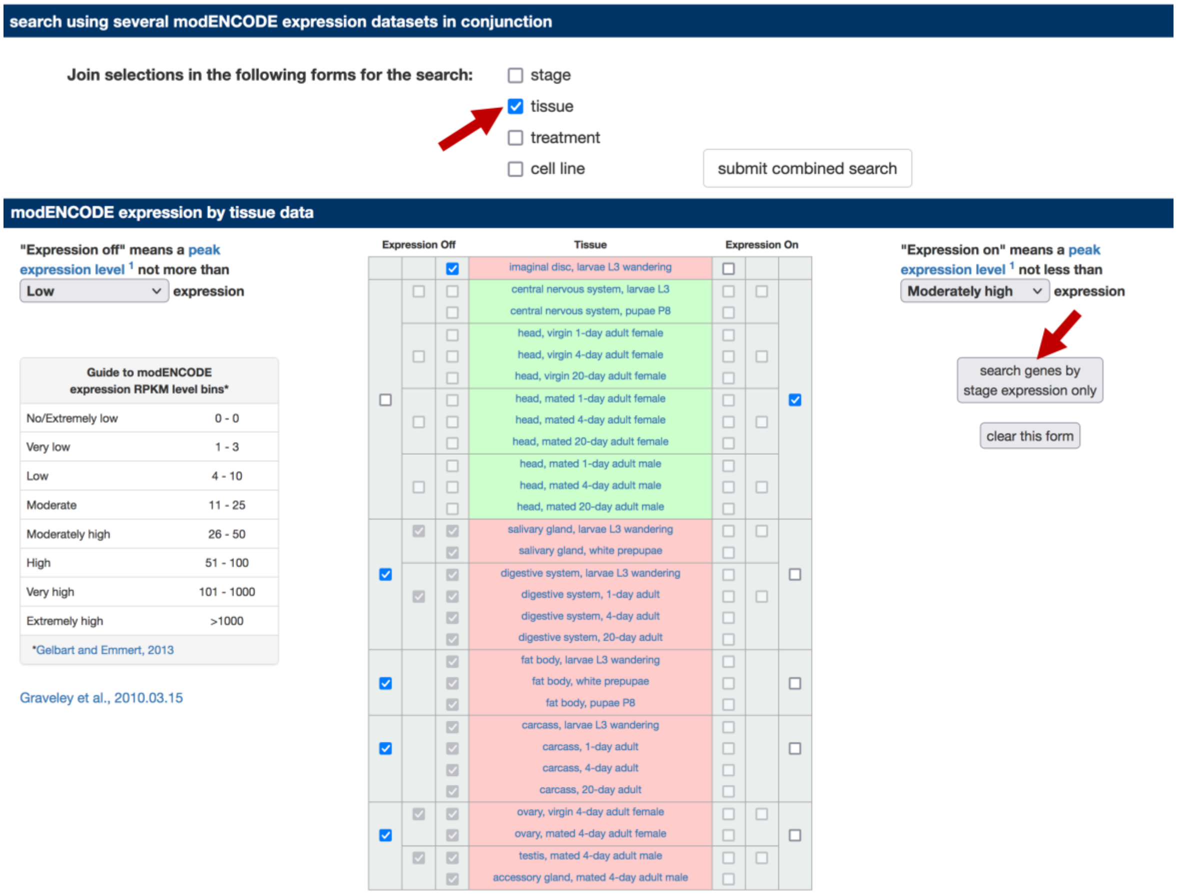
For more help with this tool, see this commentary or this help page.

Figure 1 Click on the RNA-Seq image on the FlyBase home page and then click on the "RNA-Seq Profile" link to access the "RNA-Seq Expression Profile Search" tool.

We can use this tool to identify genes that are expressed at specific developmental times, tissues, treatments, cell types, or any combinations thereof. By default, genes with **low** or no expression according to the modENCODE gene expression scale are considered to be "off" while genes with **moderately high** expression or above are considered to be "on" [see (Graveley *et al.* 2011) for details].

In this walkthrough, we will identify the subset of *D. melanogaster* genes that are expressed only in head tissues. Uncheck the "**stage"** checkbox under the "**search using several modENCODE expression datasets in conjunction**" section. Select the "**Expression On"** checkbox for all the head tissues and the central nervous system entries. Select the "**Expression Off**" checkbox for all the other tissues (Figure 2). Click on the "**search genes by stage [sic] expression only**" button to run the search.

The search results page shows 301 *D. melanogaster* genes that are expressed only in head tissues (Figure 3). To view this list of genes in a more compact format, click on the "**Table**" button under the "View As" section (blue arrow in Figure 3). We can click on the links under the "Symbol" column to learn more about each gene (the blue flag next to the gene symbol indicates the record has changed recently). We can use the "**Export**" button to export the gene list and use the "**Analyze**" button to filter this list of genes.



search using several modENCODE expression datasets in conjunction

Join selections in the following forms for the search: ☐ stage ☑ tissue ☐ treatment ☐ cell line

modENCODE expression by tissue data

"Expression off" means a peak expression level [1] not more than Low expression

"Expression on" means a peak expression level [1] not less than Moderately high expression

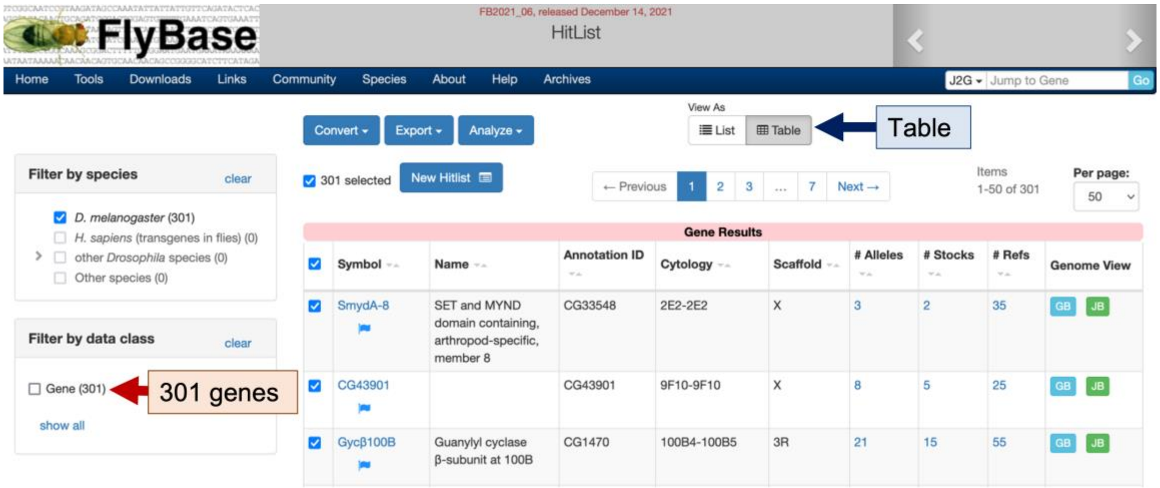
Guide to modENCODE expression RPKM level bins*	
No/Extremely low	0 - 0
Very low	1 - 3
Low	4 - 10
Moderate	11 - 25
Moderately high	26 - 50
High	51 - 100
Very high	101 - 1000
Extremely high	>1000

*Gelbart and Emmert, 2013

Graveley et al., 2010.03.15

Expression Off			Tissue	Expression On		
		☑	imaginal disc, larvae L3 wandering	☐		
☐	☐	☐	central nervous system, larvae L3	☐	☐	☑
		☐	central nervous system, pupae P8	☐		
	☐	☐	head, virgin 1-day adult female	☐	☐	
		☐	head, virgin 4-day adult female	☐		
		☐	head, virgin 20-day adult female	☐		
	☐	☐	head, mated 1-day adult female	☐	☐	
		☐	head, mated 4-day adult female	☐		
		☐	head, mated 20-day adult female	☐		
	☐	☐	head, mated 1-day adult male	☐	☐	
		☐	head, mated 4-day adult male	☐		
		☐	head, mated 20-day adult male	☐		
☑	☑	☑	salivary gland, larvae L3 wandering	☐	☐	☐
		☑	salivary gland, white prepupae	☐		
	☑	☑	digestive system, larvae L3 wandering	☐	☐	
		☑	digestive system, 1-day adult	☐		
		☑	digestive system, 4-day adult	☐		
		☑	digestive system, 20-day adult	☐		
☑		☑	fat body, larvae L3 wandering	☐		☐
		☑	fat body, white prepupae	☐		
		☑	fat body, pupae P8	☐		
☑		☑	carcass, larvae L3 wandering	☐		☐
		☑	carcass, 1-day adult	☐		
		☑	carcass, 4-day adult	☐		
		☑	carcass, 20-day adult	☐		
☑	☑	☑	ovary, virgin 4-day adult female	☐	☐	☐
		☑	ovary, mated 4-day adult female	☐		
	☑	☑	testis, mated 4-day adult male	☐	☐	
		☑	accessory gland, mated 4-day adult male	☐		

Figure 2 Use the FlyBase RNA-Seq search tool to identify *D. melanogaster* genes that are only expressed in head tissues and in the central nervous system. Click on the "search genes by stage [sic] expression only" button to search for genes that match the tissue-specific expression profiles.



Filter by species

☑ *D. melanogaster* (301)
☐ *H. sapiens* (transgenes in flies) (0)
☐ other *Drosophila* species (0)
☐ Other species (0)

Filter by data class

☐ Gene (301)

☑ 301 selected

Items 1-50 of 301

Per page: 50

Gene Results

☑	Symbol	Name	Annotation ID	Cytology	Scaffold	# Alleles	# Stocks	# Refs	Genome View
☑	SmydA-8	SET and MYND domain containing, arthropod-specific, member 8	CG33548	2E2-2E2	X	3	2	35	GB JB
☑	CG43901		CG43901	9F10-9F10	X	8	5	25	GB JB
☑	Gycβ100B	Guanylyl cyclase β-subunit at 100B	CG1470	100B4-100B5	3R	21	15	55	GB JB

Figure 3 The FlyBase RNA-Seq Expression Profile Search tool identified 301 genes (red arrow) that exhibit the defined expression profile. Click on the "Table" button (blue arrow) to view the HitList in a more compact format.

Filtering the gene list by chromosomes

In this example, we would like to investigate the set of dot chromosome genes that are only expressed in head tissues. To filter this initial list of genes by chromosome, click on the "**Analyze**" button and select "**Chromosome arm**" in the drop-down menu (Figure 4).

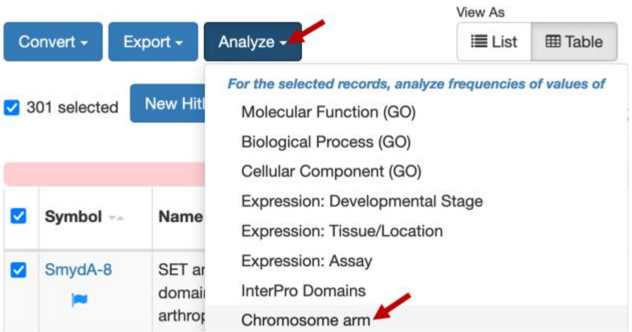
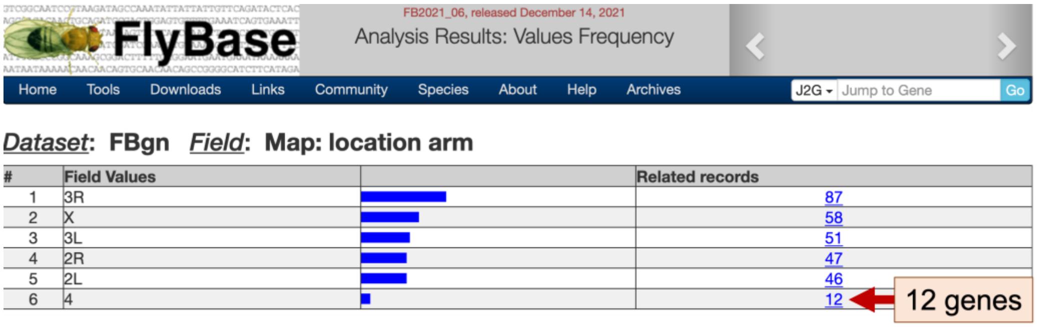


Figure 4 Use the "Analyze" button to filter the gene list by chromosome.

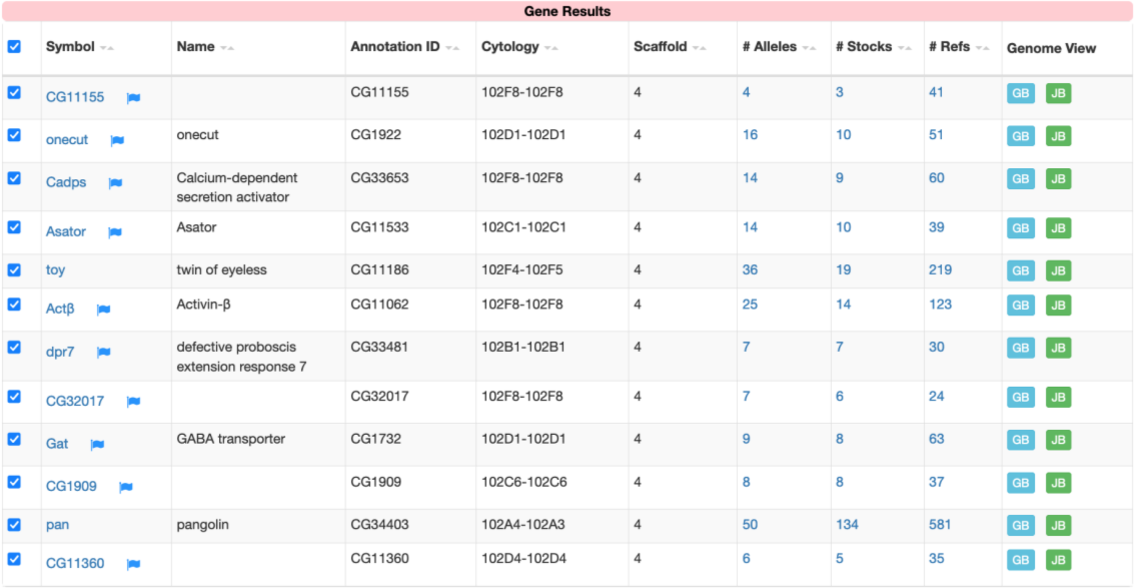
The "Values Frequency" page shows the distribution of the genes in the gene list among the different chromosomes. Under the "**Related records**" column, we see that 12 of the 301 genes that are expressed only in head tissues are found on the dot chromosome (Figure 5). Click on this link (with the label 12) to filter the original list of genes so that only dot chromosome genes are in the new gene list (Figure 6). We could also apply other filters to the gene list (e.g., using controlled vocabularies to filter the list by molecular functions or biological processes) to manipulate the list of gene candidates that we would like to use to discover common motifs.



Dataset: FBgn *Field*: Map: location arm

#	Field Values		Related records
1	3R		87
2	X		58
3	3L		51
4	2R		47
5	2L		46
6	4		12

Figure 5 Of the 301 genes that are expressed only in head tissues, 12 of them are located on the dot chromosome.



Gene Results

☑	Symbol	Name	Annotation ID	Cytology	Scaffold	# Alleles	# Stocks	# Refs	Genome View
☑	CG11155		CG11155	102F8-102F8	4	4	3	41	GB JB
☑	onecut	onecut	CG1922	102D1-102D1	4	16	10	51	GB JB
☑	Cadps	Calcium-dependent secretion activator	CG33653	102F8-102F8	4	14	9	60	GB JB
☑	Asator	Asator	CG11533	102C1-102C1	4	14	10	39	GB JB
☑	toy	twin of eyeless	CG11186	102F4-102F5	4	36	19	219	GB JB
☑	Actβ	Activin-β	CG11062	102F8-102F8	4	25	14	123	GB JB
☑	dpr7	defective proboscis extension response 7	CG33481	102B1-102B1	4	7	7	30	GB JB
☑	CG32017		CG32017	102F8-102F8	4	7	6	24	GB JB
☑	Gat	GABA transporter	CG1732	102D1-102D1	4	9	8	63	GB JB
☑	CG1909		CG1909	102C6-102C6	4	8	8	37	GB JB
☑	pan	pangolin	CG34403	102A4-102A3	4	50	134	581	GB JB
☑	CG11360		CG11360	102D4-102D4	4	6	5	35	GB JB

Figure 6 Table view of the 12 dot chromosome genes after filtering the original gene list by chromosome.

Use sequence comparisons to identify conserved motifs

In this walkthrough, we will analyze this list of 12 dot chromosome genes that are expressed only in head tissues to ascertain if there are motifs upstream of the transcription start site (TSS) that appear at high frequencies.

Identify the regions upstream of the transcription start sites

The first step of this analysis is to retrieve the genomic sequences upstream of the TSS for each gene in the gene list. Specifically, we will extract the region from the start of the TSS of the gene of interest to the beginning of the adjacent gene. (Depending on your research goal, you can adjust the size of the search region. For example, many regulatory elements are found within 2kb upstream of the TSS.) The procedure used to retrieve the upstream sequences is described below. For teaching purposes, the file with the upstream sequences (**dot_genes_head_expressed.fasta**) is available in the package for this walkthrough.

Open a new tab and navigate to FlyBase. Enter the name of a gene in the gene list (e.g., *Actbeta*, which appears as "Actβ") into "**Jump to Gene**" (J2G) search box on the top right corner of the main navigation bar and then click "**Go**" (Figure 7).

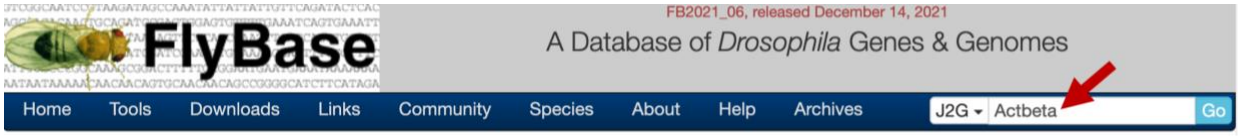
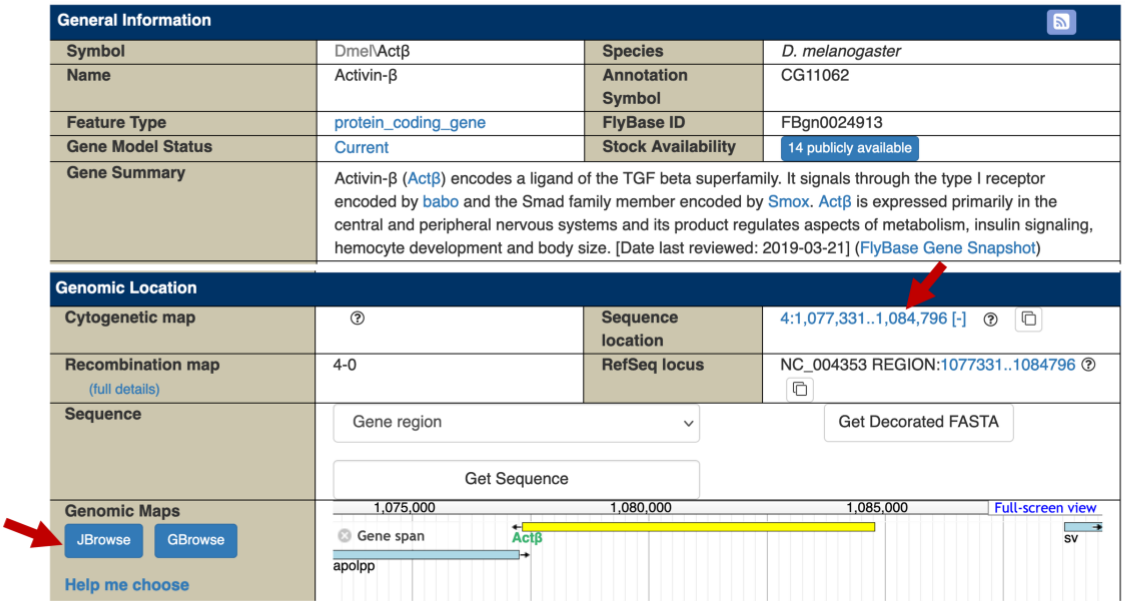

FB2021_06, released December 14, 2021
FlyBase
A Database of Drosophila Genes & Genomes
Home
Tools
Downloads
Links
Community
Species
About
Help
Archives
J2G
Actbeta
Go

Figure 7 Search for the *Actbeta* gene record using the "Jump to Gene" (J2G) search box.

The *JBrowse* image under the "Genomic Location" section shows that the *Actbeta* gene is on the minus strand and the "**Sequence location**" field shows that this gene spans from 1,077,331-1,084,796 (Figure 8). Because this gene is on the minus strand, the first nucleotide before the TSS of *Actbeta* is located at 1,084,797.



General Information

Symbol	Dmel\Actβ	Species	D. melanogaster
Name	Activin-β	Annotation Symbol	CG11062
Feature Type	protein_coding_gene	FlyBase ID	FBgn0024913
Gene Model Status	Current	Stock Availability	14 publicly available
Gene Summary	Activin-β (Actβ) encodes a ligand of the TGF beta superfamily. It signals through the type I receptor encoded by babo and the Smad family member encoded by Smox. Actβ is expressed primarily in the central and peripheral nervous systems and its product regulates aspects of metabolism, insulin signaling, hemocyte development and body size. [Date last reviewed: 2019-03-21] (FlyBase Gene Snapshot)		

Genomic Location

Cytogenetic map	ⓘ	Sequence location	4:1,077,331..1,084,796 [-] ⓘ
Recombination map (full details)	4-0	RefSeq locus	NC_004353 REGION:1077331..1084796 ⓘ
Sequence	Gene region Get Sequence	Get Decorated FASTA	
Genomic Maps			

Figure 8 The "Sequence location" field of the FlyBase Gene Report shows *Actbeta* spans from 1,077,331-1,084,796 on the dot chromosome. Click on the "*JBrowse*" button to view the genomic region surrounding the *Actbeta* gene.

The *JBrowse* image in the "Genomic Maps" section shows that the first gene upstream of the TSS of *Actbeta* is *sv*. To determine the precise coordinates of the *sv* gene, click on the "***JBrowse***" button (under "Genomic Maps") and then click on the large "Zoom out" icon in the toolbar (Figure 9).

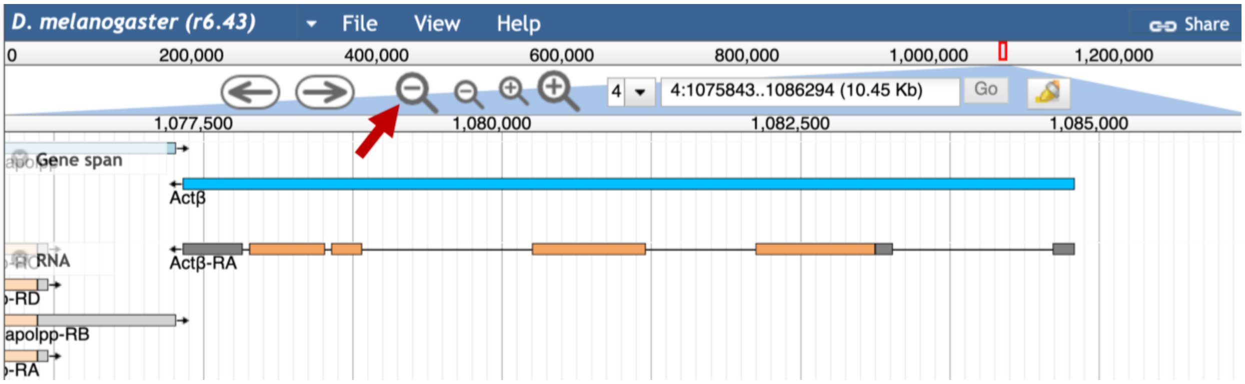
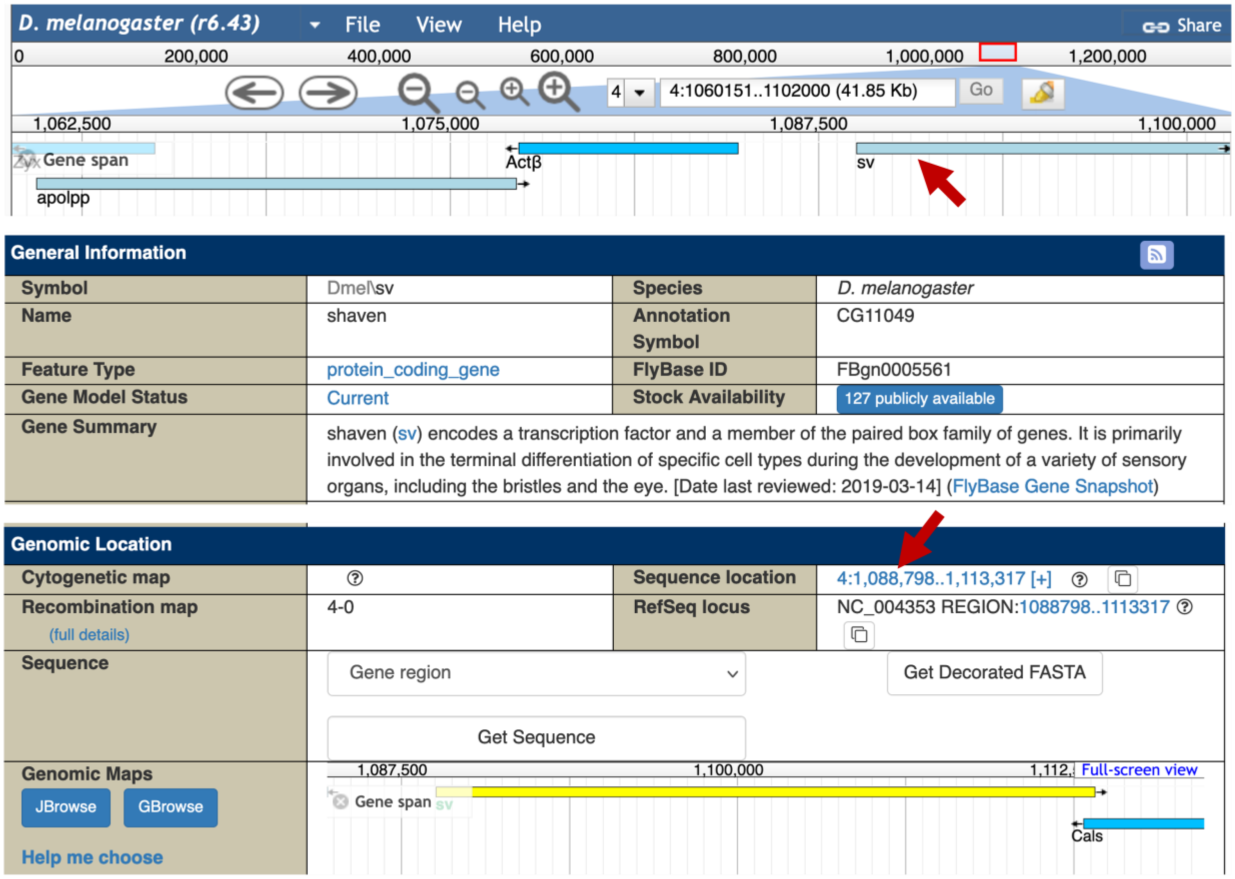


Figure 9 Click on the large "zoom out" icon (red arrow) to show the genomic region surrounding the *Actbeta* gene.

Click on the *sv* feature in the "Gene span" track. The FlyBase Gene Report shows that *sv* is on the plus strand and it begins at 1,088,798 (Figure 10). Consequently, the genomic coordinates for the region between the TSS of *Actbeta* and the beginning of *sv* is 4:1084797-1088797.



General Information

Symbol	Dmel\sv	Species	*D. melanogaster*
Name	shaven	Annotation Symbol	CG11049
Feature Type	protein_coding_gene	FlyBase ID	FBgn0005561
Gene Model Status	Current	Stock Availability	127 publicly available
Gene Summary	shaven (sv) encodes a transcription factor and a member of the paired box family of genes. It is primarily involved in the terminal differentiation of specific cell types during the development of a variety of sensory organs, including the bristles and the eye. [Date last reviewed: 2019-03-14] (FlyBase Gene Snapshot)		

Genomic Location

Cytogenetic map	ⓘ	Sequence location	4:1,088,798..1,113,317 [+] ⓘ
Recombination map (full details)	4-0	RefSeq locus	NC_004353 REGION:1088798..1113317 ⓘ
Sequence	Gene region Get Sequence		Get Decorated FASTA
Genomic Maps JBrowse GBrowse Help me choose			

Figure 10 FlyBase Gene Report shows that the *sv* gene begins at 1,088,798 on the dot chromosome.

We can apply this protocol to the other genes in our gene list to determine the coordinates for all the upstream regions. The coordinates of all the upstream regions are listed below:

Gene	Upstream region
Actbeta	4:1084797-1088797
Asator	4:488995-501809
Cadps	4:1230714-1230836
CG1909	4:580123-583107
CG11155	4:1125326-1126601
CG11360	4:669885-671141
CG32017	4:1162159-1164881
dpr7	4:253330-254996
Gat	4:626032-629690
onecut	4:607650-610683
pan	4:67208-69325
toy	4:978490-989724

Retrieve the upstream sequences

Once we have generated the list of coordinates for the upstream regions, we can use the FlyBase *Sequence Downloader* tool to retrieve the genomic sequences for these upstream regions. You can access this tool through the FlyBase main navigation bar (under "Tools" ➔ "Query by symbols/IDs" ➔ "*Sequence Downloader*", Figure 11).

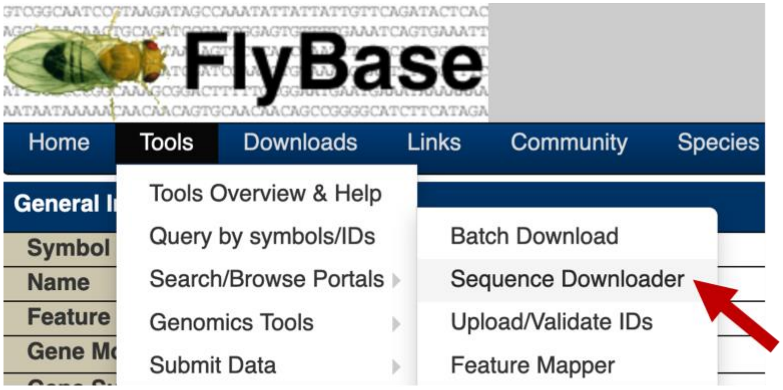
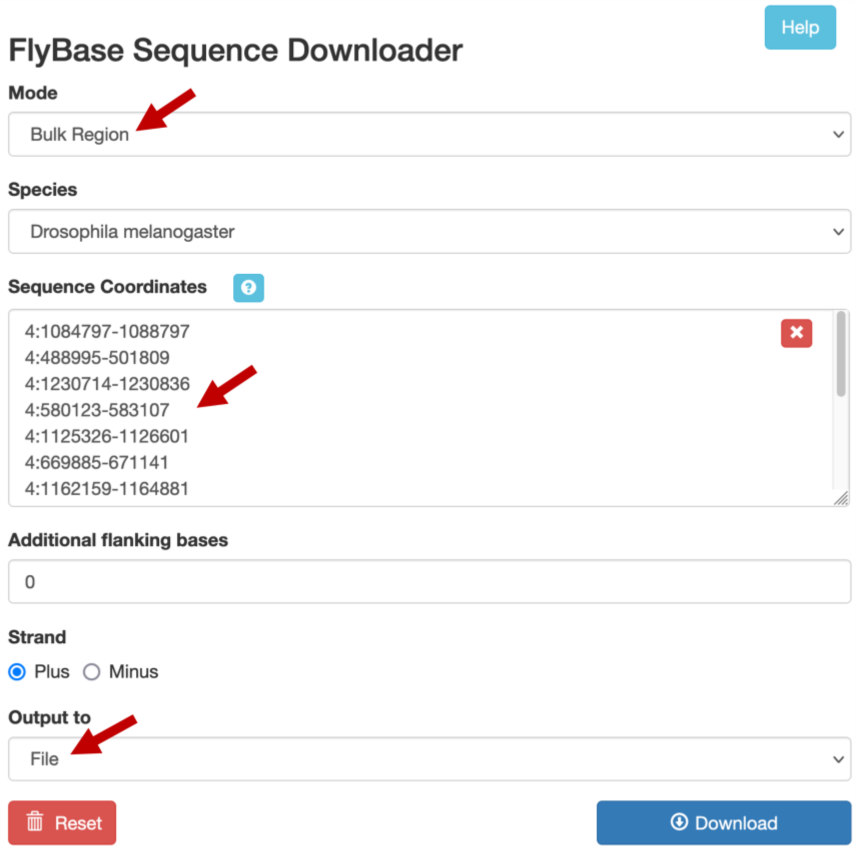


Figure 11 Use the main navigation bar to access the FlyBase *Sequence Downloader* tool.

Select the "**Bulk Region**" option under the "Mode" field. Verify that "***Drosophila melanogaster***" is selected under the "Species" field. Copy and paste the list of upstream sequence coordinates into the "Sequence Coordinates" text box (Figure 12). Verify that the "Additional flanking bases" field is set to "**0**", and that the "Strand" field is set to "**Plus**". Change the "Output to" field to "**File**". Click on the "**Download**" button to run the program. Save the output file as **dot_genes_head_expressed.fasta**.

If a blank page appears after you click on the “Download” button, you might need to change the configuration of your web browser to allow pop-up windows from FlyBase. Alternatively, you can use the **dot_genes_head_expressed.fasta** file in the package for this walkthrough.



FlyBase Sequence Downloader

Mode: Bulk Region

Species: Drosophila melanogaster

Sequence Coordinates

```
4:1084797-1088797
4:488995-501809
4:1230714-1230836
4:580123-583107
4:1125326-1126601
4:669885-671141
4:1162159-1164881
```

Additional flanking bases: 0

Strand: ● Plus ○ Minus

Output to: File

Figure 12 Change the *Sequence Downloader* to the “Bulk Region” mode to download the sequences upstream of the 12 dot chromosome genes.

To facilitate interpretation of the results during the motif discovery analysis, we will modify the definition lines (i.e. the lines that begin with the **>** symbol) in the FASTA file. Open the FASTA file in a text editor (e.g., WordPad on MS Windows, TextEdit on macOS). Change each definition line so that it begins with the gene name, followed by a space, followed by “loc=” and the coordinates of the extracted region.

For example, for the upstream region of *Actbeta* (i.e. the feature at 1084797-1088797), we will change the definition line to the following:

```
>Actbeta loc=4:1084797-1088797
```

Once you have modified the definition lines, save and close the sequence file.

Identify enriched motifs in the collection of upstream sequences

Now that we have generated the collection of upstream sequences using the FlyBase *Sequence Downloader* tool, we can search for motifs that are enriched in this collection of sequences. Among the plethora of motif discovery tools that are publicly available, the *MEME* suite is one of the most popular solutions (Bailey *et al.* 2009). In this walkthrough, we will use the web-based *MEME* interface to analyze our collection of sequences. Navigate to the *MEME* website at https://meme-suite.org/, and click on the "*MEME*" icon (Figure 13).

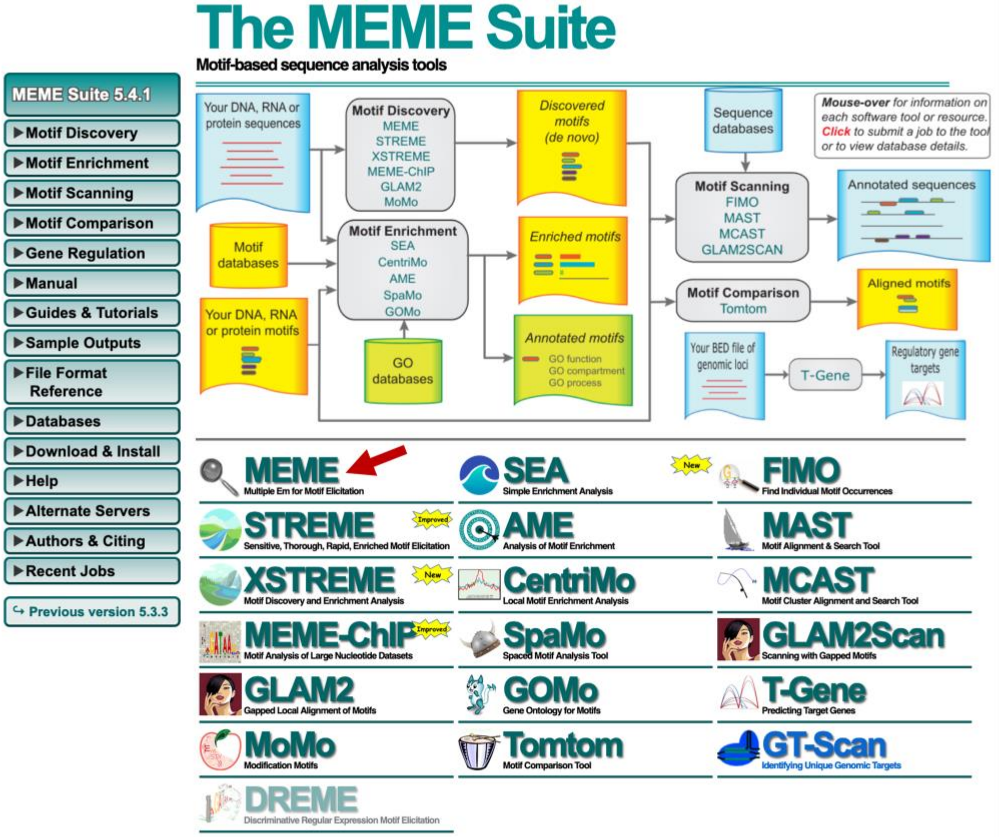
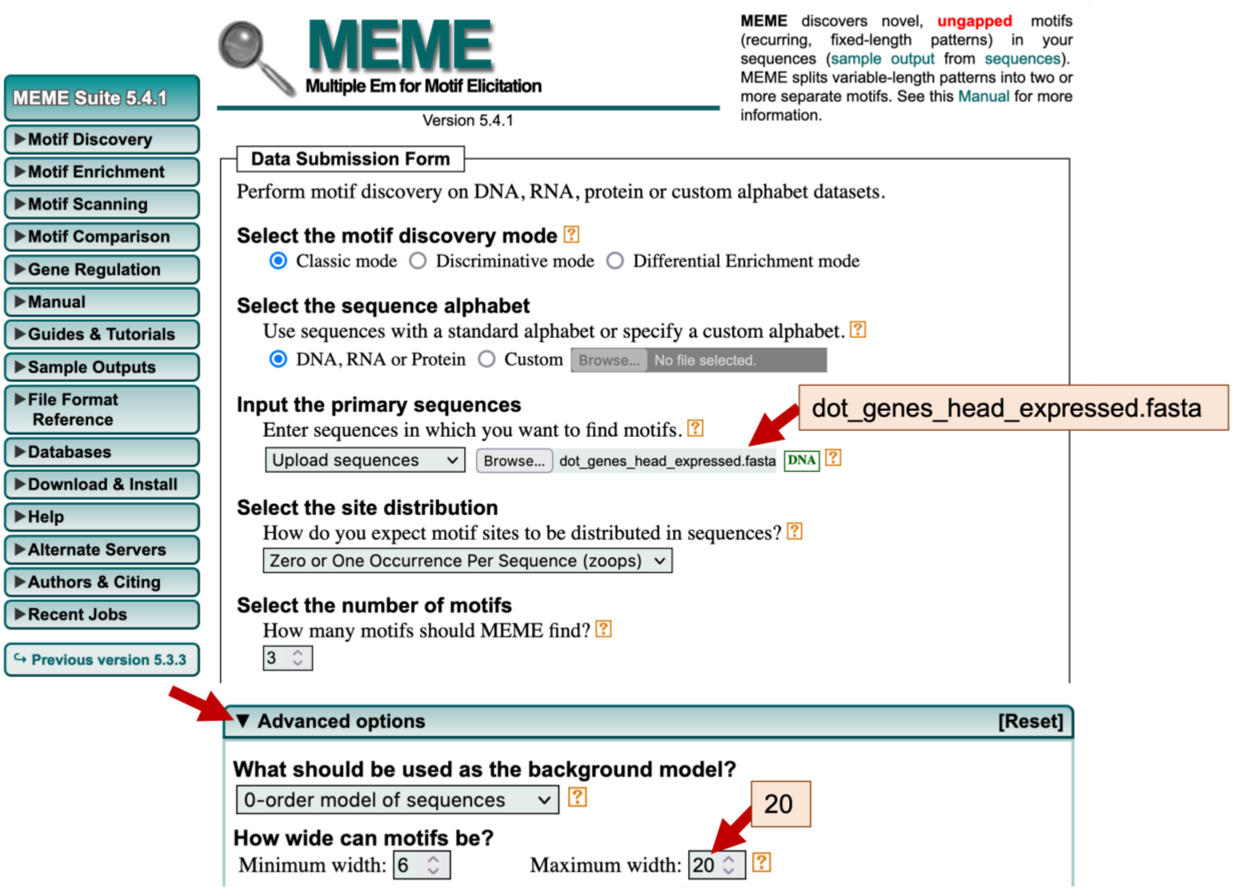


Figure 13 Click on the *MEME* logo on the *MEME* Suite website to access the *MEME* web interface.

We will perform the initial *MEME* search using mostly default parameters. Verify that the "Select the motif discovery mode" field is set to "**Classic mode.**" Under the "Input the primary sequences" field, click on the "**Browse**" or the "**Choose File**" button and select the file with the upstream sequences (i.e. **dot_genes_head_expressed.fasta**). Because our collection of upstream sequences may contain different motifs, we should verify that the "Select the site distribution" field is set to "**Zero or One Occurrence Per Sequence (zoops)**." We should also verify that that the "Select the number of motifs" field is set to **3**.

Because most conserved motifs are short [e.g., the average length of transcription factor binding sites is 10 nucleotides (Stewart *et al.* 2012)], we should also change the maximum width of the sequence motifs that are identified by *MEME*. Click on the "Advanced options" header to expand this section. Under the "How wide can motifs be?" section, change the "Maximum width" to **20** (Figure 14). (If a "*MEME* SUITE NEWS" message box appears at the bottom of the page, close the message box to access the "Advanced options" header.)



MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in your sequences (sample output from sequences). MEME splits variable-length patterns into two or more separate motifs. See this Manual for more information.

Data Submission Form

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

Select the motif discovery mode

● Classic mode ○ Discriminative mode ○ Differential Enrichment mode

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet.

● DNA, RNA or Protein ○ Custom Browse... No file selected.

Input the primary sequences

Enter sequences in which you want to find motifs.

Upload sequences Browse... dot_genes_head_expressed.fasta DNA

Select the site distribution

How do you expect motif sites to be distributed in sequences?

Zero or One Occurrence Per Sequence (zoops)

Select the number of motifs

How many motifs should MEME find?

3

▼ Advanced options [Reset]

What should be used as the background model?

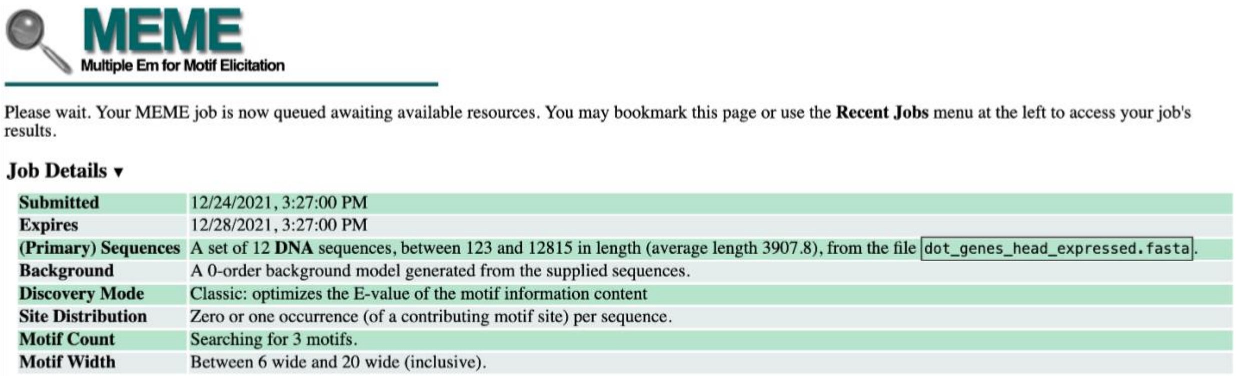
0-order model of sequences

How wide can motifs be?

Minimum width: 6 Maximum width: 20

Figure 14 Configure *MEME* to search the collection of upstream sequences for over-represented motifs.

Click on the “Start Search” button at the bottom of the page to run the *MEME* program on our collection of upstream sequences. Depending on how busy the server is, this search could take a while to complete (Figure 15).



Please wait. Your MEME job is now queued awaiting available resources. You may bookmark this page or use the **Recent Jobs** menu at the left to access your job's results.

Job Details ▾

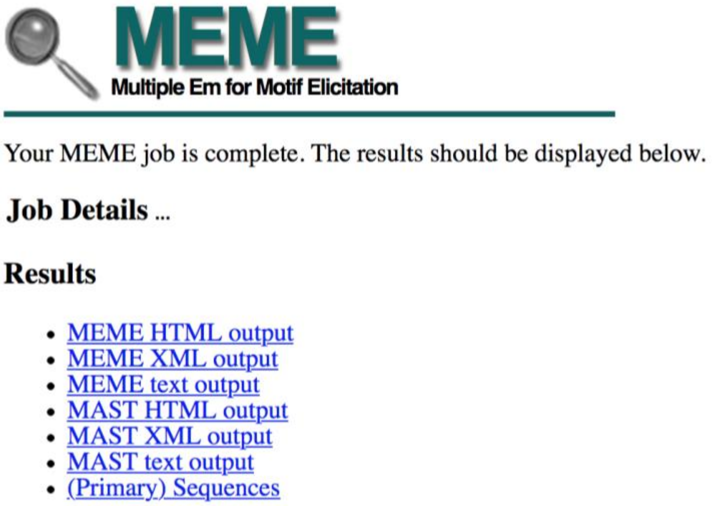
Submitted	12/24/2021, 3:27:00 PM
Expires	12/28/2021, 3:27:00 PM
(Primary) Sequences	A set of 12 **DNA** sequences, between 123 and 12815 in length (average length 3907.8), from the file dot_genes_head_expressed.fasta.
Background	A 0-order background model generated from the supplied sequences.
Discovery Mode	Classic: optimizes the E-value of the motif information content
Site Distribution	Zero or one occurrence (of a contributing motif site) per sequence.
Motif Count	Searching for 3 motifs.
Motif Width	Between 6 wide and 20 wide (inclusive).

Figure 15 Message from the *MEME* server once the *MEME* job has been queued.

This page will refresh automatically when the *MEME* search is complete. For teaching purposes, the *MEME* search results (**dot_head_expressed_MEME.html**) are available inside the MEME_results folder in the package for this walkthrough.

Interpreting the *MEME* and *MAST* results

Once the *MEME* search is complete, the search results page will show links to the *MEME* and *MAST* output in three different formats (HTML, XML, and text, Figure 16). The *MEME* program attempts to find over-represented motifs in our collection of sequences. Once these motifs have been identified, the distributions of these motifs in each upstream sequence are determined by *MAST*.



Your MEME job is complete. The results should be displayed below.

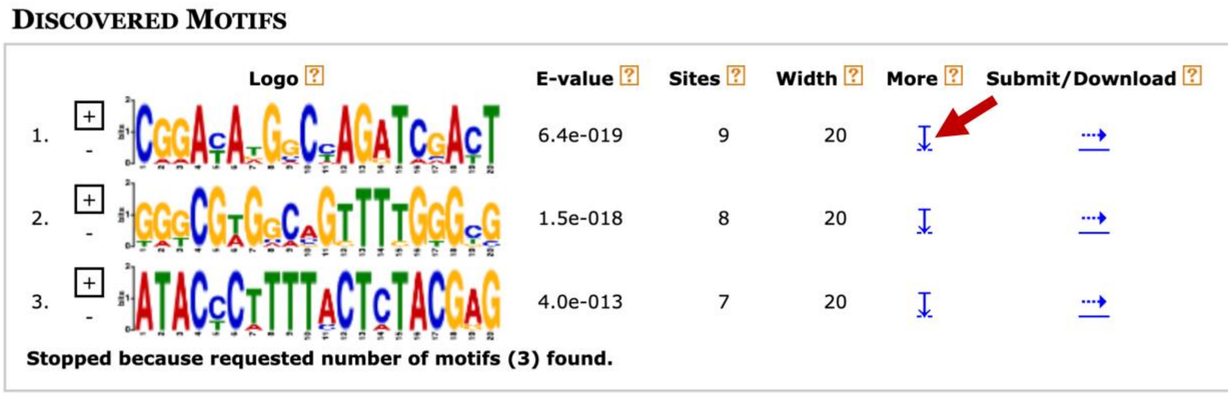
Job Details ...

Results

- MEME HTML output
- MEME XML output
- MEME text output
- MAST HTML output
- MAST XML output
- MAST text output
- (Primary) Sequences

Figure 16 Links to the *MEME* and *MAST* search results in HTML, XML, and text formats.

Right click (control-click on macOS) on the "***MEME* HTML output**" link and open the *MEME* search results page in a new tab. The "**Discovered Motifs**" section shows the three most significant motifs that were found by *MEME* (Figure 17). The "**Logo**" column contains the sequence logo for each motif, where the height of each base corresponds to the level of sequence conservation. The "+" and "-" buttons next to each sequence logo allow us to view the motif in the forward ("+") and the reverse complement ("-") orientation, respectively. The value in the "**E-value**" column corresponds to the statistical significance of the motif and the value in the "**Sites**" column corresponds to the number of sites within the primary sequences that were used to construct each motif.



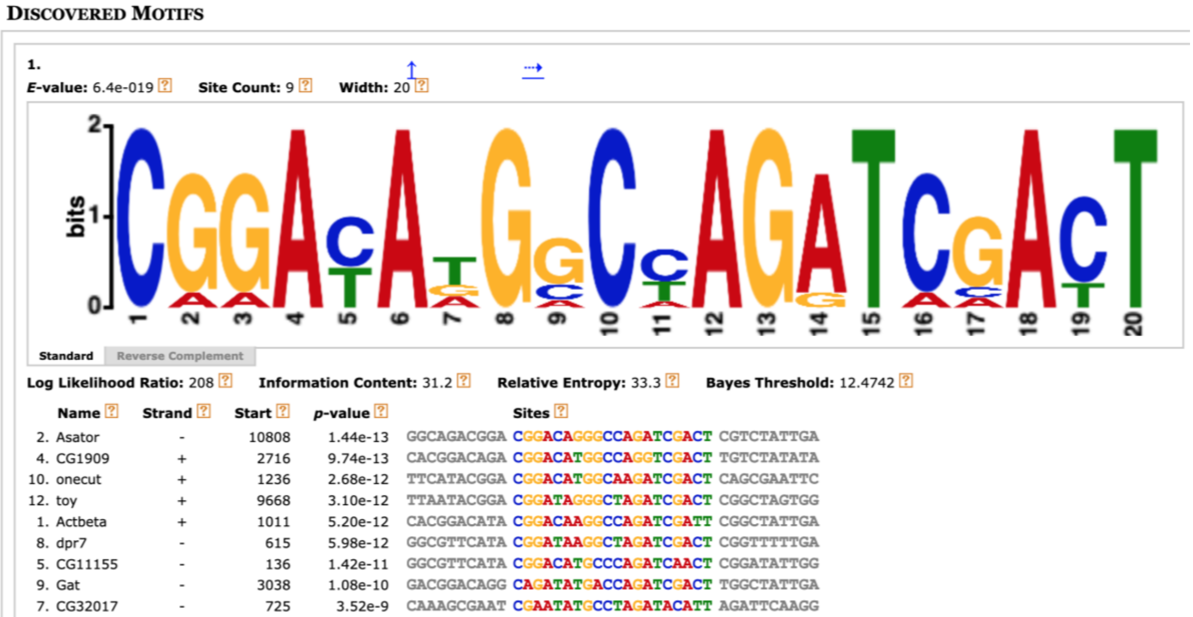
DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.	⊞ -	6.4e-019	9	20	↧	⇢
2.	⊞ -	1.5e-018	8	20	↧	⇢
3.	⊞ -	4.0e-013	7	20	↧	⇢

Stopped because requested number of motifs (3) found.

Figure 17 The three most significant motifs in the collection of dot chromosome upstream sequences. Click on the down arrow under the "More" column to learn more about each motif.

To learn more about the first motif, click on the blue down arrow link under the "More" column (Figure 17). The multiple sequence alignment beneath the sequence logo corresponds to the nine sites that were used to construct the first motif (Figure 18).



DISCOVERED MOTIFS

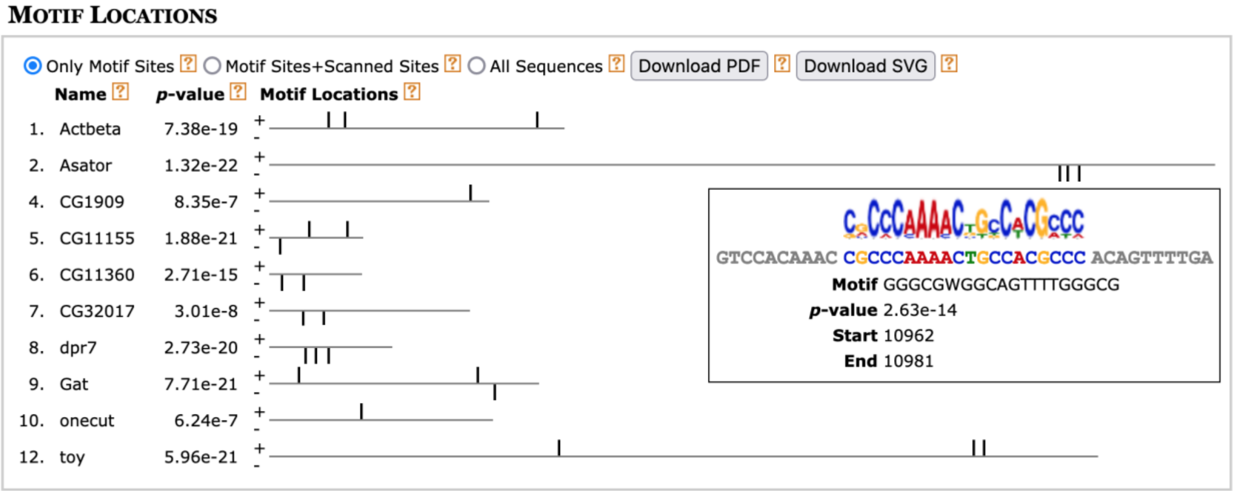
1.
E-value: 6.4e-019 ? Site Count: 9 ? Width: 20 ?

Log Likelihood Ratio: 208 ? Information Content: 31.2 ? Relative Entropy: 33.3 ? Bayes Threshold: 12.4742 ?

Name ?	Strand ?	Start ?	*p*-value ?	Sites ?
2. Asator	-	10808	1.44e-13	GGCAGACGGA CGGACAGGGCCAGATCGACT CGTCTATTGA
4. CG1909	+	2716	9.74e-13	CACGGACAGA CGGACATGGCCAGGTCGACT TGTCTATATA
10. onecut	+	1236	2.68e-12	TTCATACGGA CGGACATGGCAAGATCGACT CAGCGAATTC
12. toy	+	9668	3.10e-12	TTAATACGGA CGGATAGGGCTAGATCGACT CGGCTAGTGG
1. Actbeta	+	1011	5.20e-12	CACGGACATA CGGACAAGGCCAGATCGATT CGGCTATTGA
8. dpr7	-	615	5.98e-12	GGCGTTCATA CGGATAAGGCTAGATCGACT CGGTTTTTGA
5. CG11155	-	136	1.42e-11	GGCGTTCATA CGGACATGCCCAGATCAACT CGGATATTGG
9. Gat	-	3038	1.08e-10	GACGGACAGG CAGATATGACCAGATCGACT TGGCTATTGA
7. CG32017	-	725	3.52e-9	CAAAGCGAAT CGAATATGCCTAGATACATT AGATTCAAGG

Figure 18 The nine sequences used by *MEME* to construct the first motif.

Because transcription factors often work in concert with each other, we often would like to know the distribution of all the motifs identified by *MEME*. A diagram of the collection of non-overlapping motif instances are shown in the "**Motif Locations**" section along with their statistical significance (height of the block) (Figure 19). A tooltip will appear with additional details on each motif site when you hover the mouse on top of each block.



MOTIF LOCATIONS

● Only Motif Sites ○ Motif Sites+Scanned Sites ○ All Sequences

Name ?	*p*-value ?	Motif Locations ?
1. Actbeta	7.38e-19	
2. Asator	1.32e-22	
4. CG1909	8.35e-7	
5. CG11155	1.88e-21	
6. CG11360	2.71e-15	
7. CG32017	3.01e-8	
8. dpr7	2.73e-20	
9. Gat	7.71e-21	
10. onecut	6.24e-7	
12. toy	5.96e-21	

GTCCACAAAC CGCCCAAAACTGCCACGCCC ACAGTTTTGA
Motif GGGCGWGGCAGTTTTGGGCG
***p*-value** 2.63e-14
Start 10962
End 10981

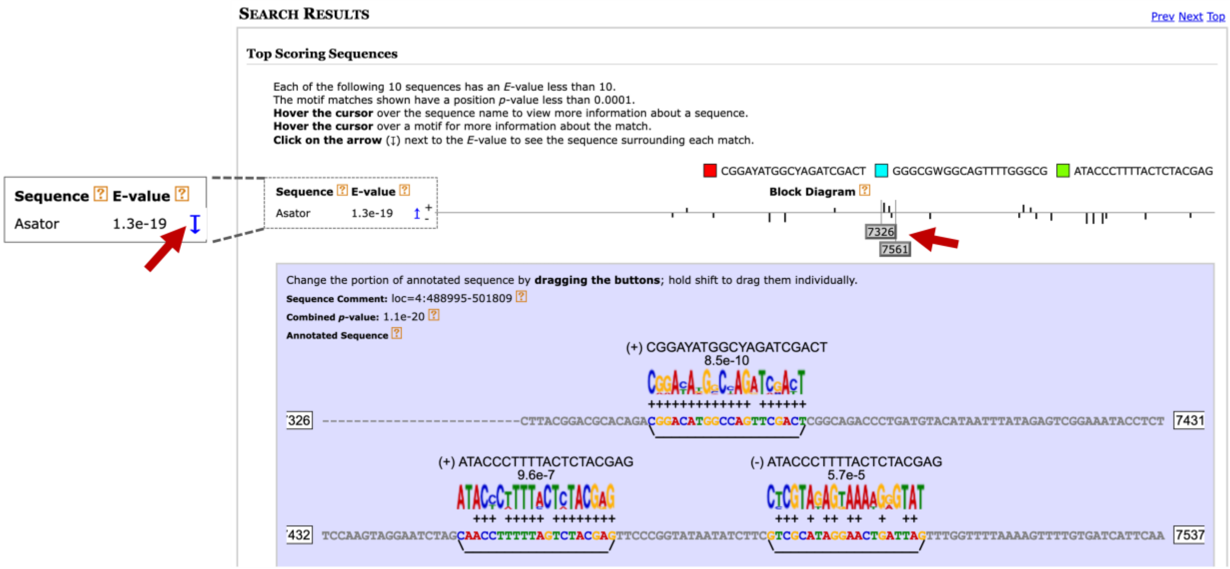
Figure 19 Block diagram that shows the distribution of all three motifs in the collection of upstream sequences. The height of each block corresponds to the p-value of the motif instance.

By default, the "**Only Motif Sites**" option (located underneath the "**Motif Locations**" header) is selected and the diagram will only show the motif instances that were used to construct each motif. Because we have previously specified in our search parameters that we expect to find zero or one instance of each motif within each sequence, each sequence in the block diagram will contain at most one motif instance.

If you select the "**Motif Sites+Scanned Sites**" option, *MEME* will search each motif against the entire collection of upstream sequences and there could be multiple significant matches to each motif. Selecting the "**All Sequences**" option will show all 12 sequences that are in our input sequence file irrespective of whether they contain any of the three motifs.

We can examine the set of significant non-overlapping motif instances in more detail using the "**Top Scoring Sequences**" section of the *MAST* HTML output. To view the *MAST* results, go back to the tab with the links to the *MEME* and *MAST* output (Figure 16). Right click (or control-click on macOS) on the "***MAST* HTML output**" link and open the *MAST* search results page in a new tab.

Scroll down to the "Search Results" section. Click on the blue arrow next to the E-value of the sequence you want to investigate (e.g., *Asator*) to expand the section. Then drag the gray buttons under the **Block Diagram** section to navigate to the specific motif matches you want to examine (Figure 20). Hold the shift key before dragging the gray buttons in the block diagram to adjust the range of the expanded section.



SEARCH RESULTS

Top Scoring Sequences

Each of the following 10 sequences has an *E*-value less than 10.
The motif matches shown have a position *p*-value less than 0.0001.
Hover the cursor over the sequence name to view more information about a sequence.
Hover the cursor over a motif for more information about the match.
Click on the arrow (↕) next to the *E*-value to see the sequence surrounding each match.

🟥 CGGAYATGGCYAGATCGACT 🟦 GGGCGWGGCAGTTTTGGGCG 🟩 ATACCCTTTTACTCTACGAG

Sequence	E-value	Block Diagram
Asator	1.3e-19	

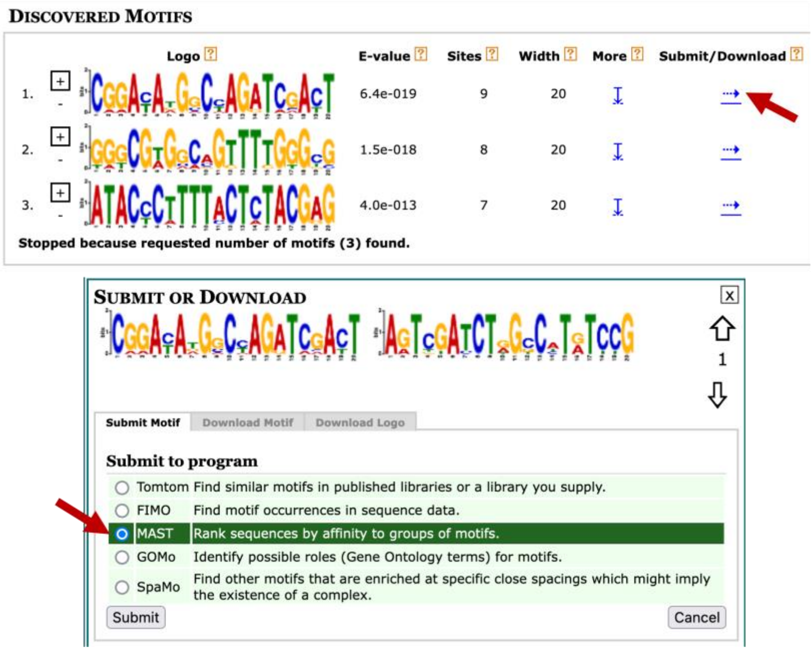
Change the portion of annotated sequence by **dragging the buttons**; hold shift to drag them individually.
Sequence Comment: loc=4:488995-501809
Combined *p*-value: 1.1e-20
Annotated Sequence

```
(+) CGGAYATGGCYAGATCGACT
8.5e-10
+++++++++++++ ++++++
7326 ------------------------CTTACGGACGCACAGACGGACATGGCCAGTTCGACTCGGCAGACCCTGATGTACATAATTTATAGAGTCGGAAATACCTCT 7431

(+) ATACCCTTTTACTCTACGAG        (-) ATACCCTTTTACTCTACGAG
9.6e-7                          5.7e-5
+++ +++++ ++++++++              +++ + ++ ++ + ++
7432 TCCAAGTAGGAATCTAGCAACCTTTTTAGTCTACGAGTTCCCGGTATAATATCTTCGTCGCATAGGAACTGATTAGTTTGGTTTTAAAAGTTTTGTGATCATTCAA 7537
```

Figure 20 Examine significant non-overlapping motifs using the *MAST* output. Click on the blue arrow next to the E-value to examine the motif matches in an input sequence (e.g., ***Asator*). Hold the shift key and drag the gray buttons to change the size of the viewing range.**

In addition to searching for motif instances against our collection of upstream sequences, we can also use the *MAST* tool to search the motifs against other collection of sequences. Go back to the web browser tab with the *MEME* search results. Click on the blue right arrow under the "**Submit/Download**" column for motif 1. Select "***MAST***" under the "Submit to program" section and then click on the "Submit" button (Figure 21).



DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.	CGGACAAGCCAGATCGACT	6.4e-019	9	20		
2.	GGGCGTGCCAGTTTTGGGCG	1.5e-018	8	20		
3.	ATACCCTTTTACTCTACGAG	4.0e-013	7	20		

Stopped because requested number of motifs (3) found.

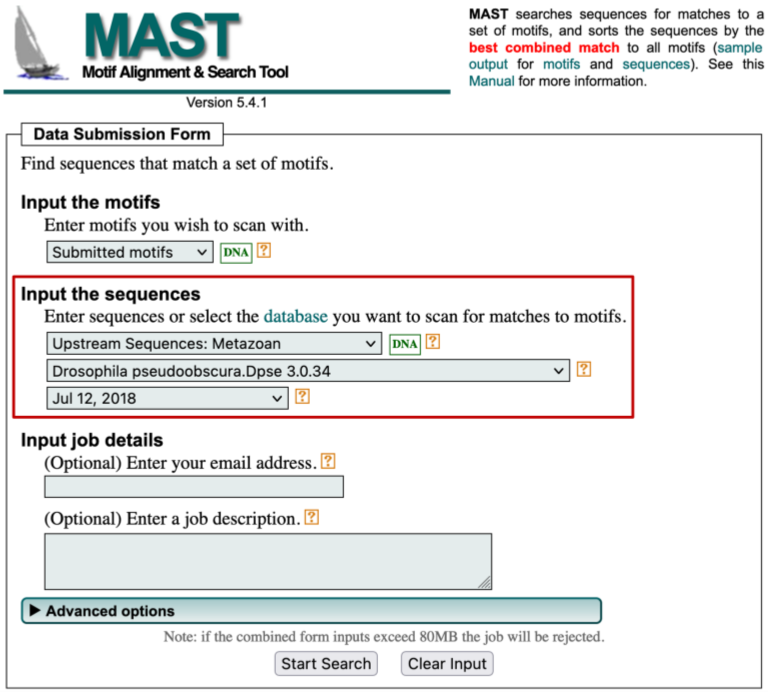
SUBMIT OR DOWNLOAD

Submit to program

- ○ Tomtom Find similar motifs in published libraries or a library you supply.
- ○ FIMO Find motif occurrences in sequence data.
- ● MAST Rank sequences by affinity to groups of motifs.
- ○ GOMo Identify possible roles (Gene Ontology terms) for motifs.
- ○ SpaMo Find other motifs that are enriched at specific close spacings which might imply the existence of a complex.

Figure 21 Use *MAST* to search individual motifs identified by *MEME* against other collections of sequences.

For example, we can use *MAST* to search the discovered motifs against the collection of sequences upstream of the TSS of *D. pseudoobscura* genes by changing the "Input the sequences" field to "**Upstream Sequences: Metazoan**" and then selecting the "***Drosophila pseudoobscura*.Dpse 3.0.34**" and "**Jul 12, 2018**" options (Figure 22). This search allows us to determine if a motif on the *D. melanogaster* dot chromosome is also found in *D. pseudoobscura*. Click on the "Start Search" button to launch the *MAST* analysis.



MAST searches sequences for matches to a set of motifs, and sorts the sequences by the best combined match to all motifs (sample output for motifs and sequences). See this Manual for more information.

Data Submission Form

Find sequences that match a set of motifs.

Input the motifs

Enter motifs you wish to scan with.

Submitted motifs DNA

Input the sequences

Enter sequences or select the database you want to scan for matches to motifs.

Upstream Sequences: Metazoan DNA

Drosophila pseudoobscura.Dpse 3.0.34

Jul 12, 2018

Input job details

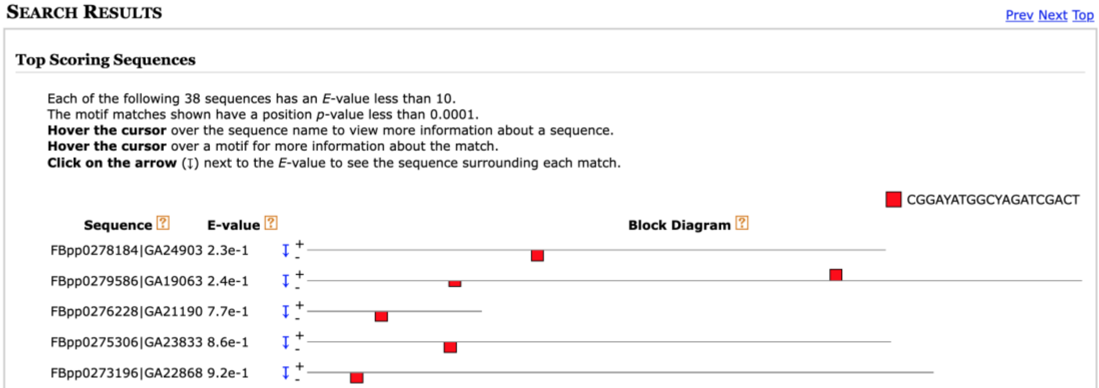
(Optional) Enter your email address.

(Optional) Enter a job description.

Note: if the combined form inputs exceed 80MB the job will be rejected.

Figure 22 Configure *MAST* to search the motif discovered by *MEME* against the collection of sequences upstream of the transcription start sites in *D. pseudoobscura*.

Once the *MAST* search is complete, right click (or control-click on macOS) on the "***MAST* HTML output**" link and open the *MAST* search results page in a new tab. The "**Search Results**" section shows *MAST* has detected 38 *D. pseudoobscura* upstream sequences with significant (E-value < 10) matches to the first motif discovered by *MEME* (Figure 23).



Search Results

Top Scoring Sequences

Each of the following 38 sequences has an *E*-value less than 10.
The motif matches shown have a position *p*-value less than 0.0001.
Hover the cursor over the sequence name to view more information about a sequence.
Hover the cursor over a motif for more information about the match.
Click on the arrow (↧) next to the *E*-value to see the sequence surrounding each match.

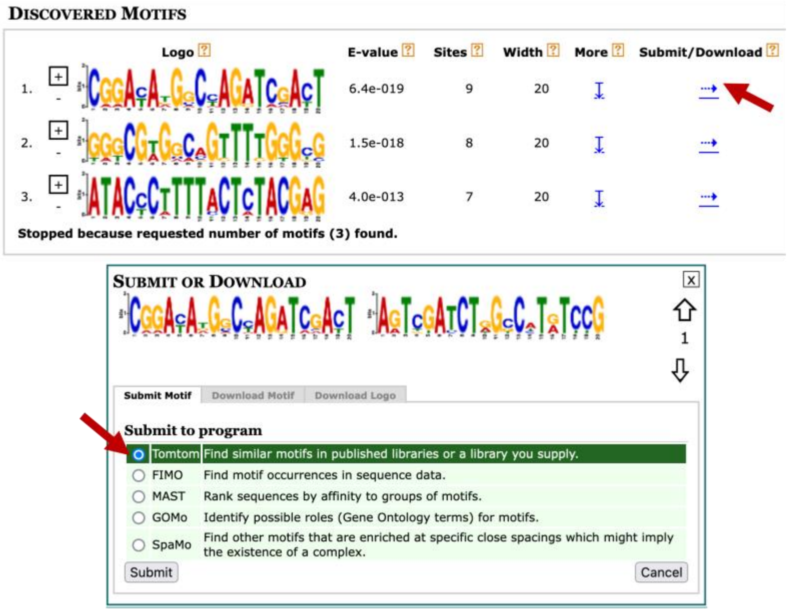
🟥 CGGAYATGGCYAGATCGACT

Sequence	E-value	Block Diagram
FBpp0278184|GA24903	2.3e-1	
FBpp0279586|GA19063	2.4e-1	
FBpp0276228|GA21190	7.7e-1	
FBpp0275306|GA23833	8.6e-1	
FBpp0273196|GA22868	9.2e-1	

Figure 23 Upstream *D. pseudoobscura* sequences with significant matches to the first motif identified by *MEME*.

The "**Sequence**" column shows the accession number and the name of the gene that contains the motif. The first part of the name (e.g., FBpp0278184) corresponds to the FlyBase ID and the second part of the name (e.g., GA24903) corresponds to the FlyBase annotation symbol.

In addition to determining the frequency and distributions of the motifs, we can also determine if any of the motifs identified by *MEME* are similar to known motifs in *Drosophila* using the *Tomtom* program (Gupta *et al.* 2007). Go back to the *MEME* results page and then click on the blue right arrow under the "**Submit/Download**" column for motif 1. Select "***Tomtom***" under the "Submit to program" section and then click on the "Submit" button (Figure 24).



Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Download
1.		6.4e-019	9	20	↧	⇢
2.		1.5e-018	8	20	↧	⇢
3.		4.0e-013	7	20	↧	⇢

Stopped because requested number of motifs (3) found.

Submit or Download

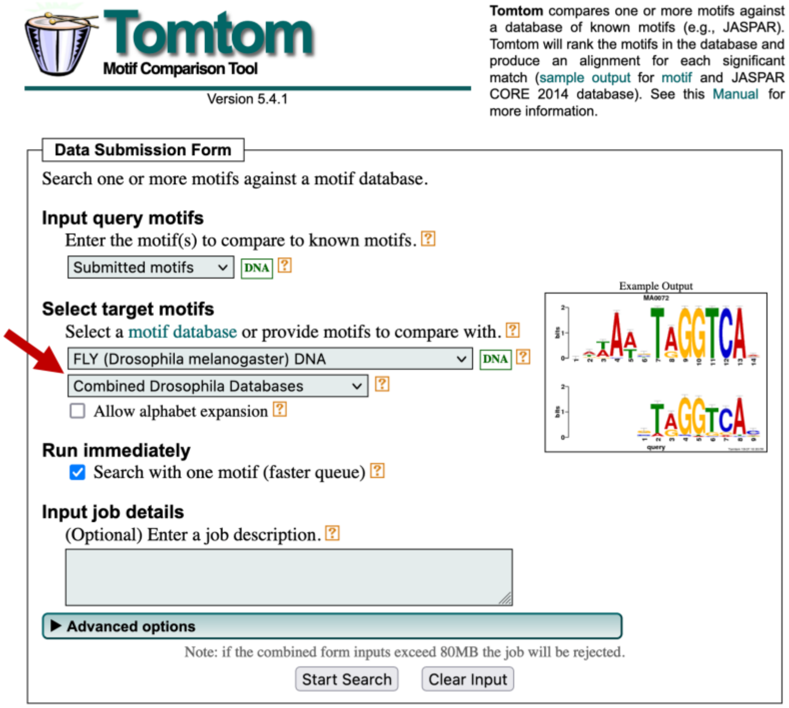
Submit to program

● Tomtom Find similar motifs in published libraries or a library you supply.
○ FIMO Find motif occurrences in sequence data.
○ MAST Rank sequences by affinity to groups of motifs.
○ GOMo Identify possible roles (Gene Ontology terms) for motifs.
○ SpaMo Find other motifs that are enriched at specific close spacings which might imply the existence of a complex.

Figure 24 Use *Tomtom* to search a motif discovered by *MEME* against a database of known motifs.

In this example, we will search the first motif identified by *MEME* against multiple databases of known motifs in *D. melanogaster*. Select "**FLY (*Drosophila melanogaster*) DNA**" and "**Combined Drosophila Databases**" under the "Select target motifs" field (Figure 25). Click on the "Start Search" button.

Using this configuration, *Tomtom* will search our motif against five motif databases (OnTheFly_2014, Fly Factor Survey, FLYREG, iDMMPMM, and DMMPMM). These databases contain motifs that are supported by different types of experimental evidence. For example, the transcription factor binding sites in the Fly Factor Survey database were identified primarily using the bacterial one-hybrid (B1H) system (Zhu *et al.* 2011).



Tomtom compares one or more motifs against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match (sample output for motif and JASPAR CORE 2014 database). See this Manual for more information.

Data Submission Form

Search one or more motifs against a motif database.

Input query motifs
Enter the motif(s) to compare to known motifs. ?
Submitted motifs DNA ?

Select target motifs
Select a motif database or provide motifs to compare with. ?
FLY (Drosophila melanogaster) DNA DNA ?
Combined Drosophila Databases ?
☐ Allow alphabet expansion ?

Run immediately
☑ Search with one motif (faster queue) ?

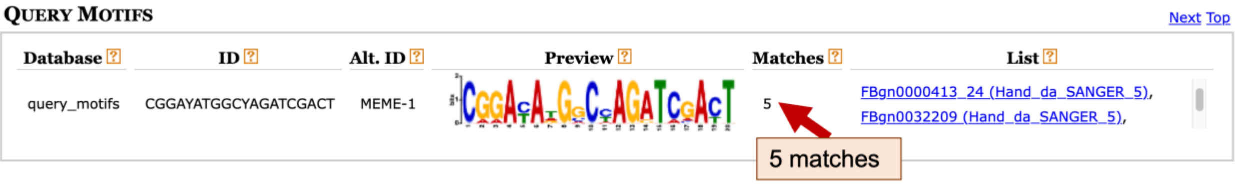
Input job details
(Optional) Enter a job description. ?

▶ Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Figure 25 Use the *Tomtom* program to search motif 1 against the set of known *Drosophila* DNA motifs.

This *Tomtom* search might take a few minutes to complete. Once the results are available, click on the "***Tomtom* HTML output**" link to view the search result. We can see under the **Query Motifs** section that *Tomtom* has identified five matches to motif 1 (Figure 26). A closer examination of the list of matches shows that two of the matches (FBgn0000413_24 and FBgn0032209) have the same alternate name (Hand_da_SANGER_5), indicating that these two matches refer to the same motif record.



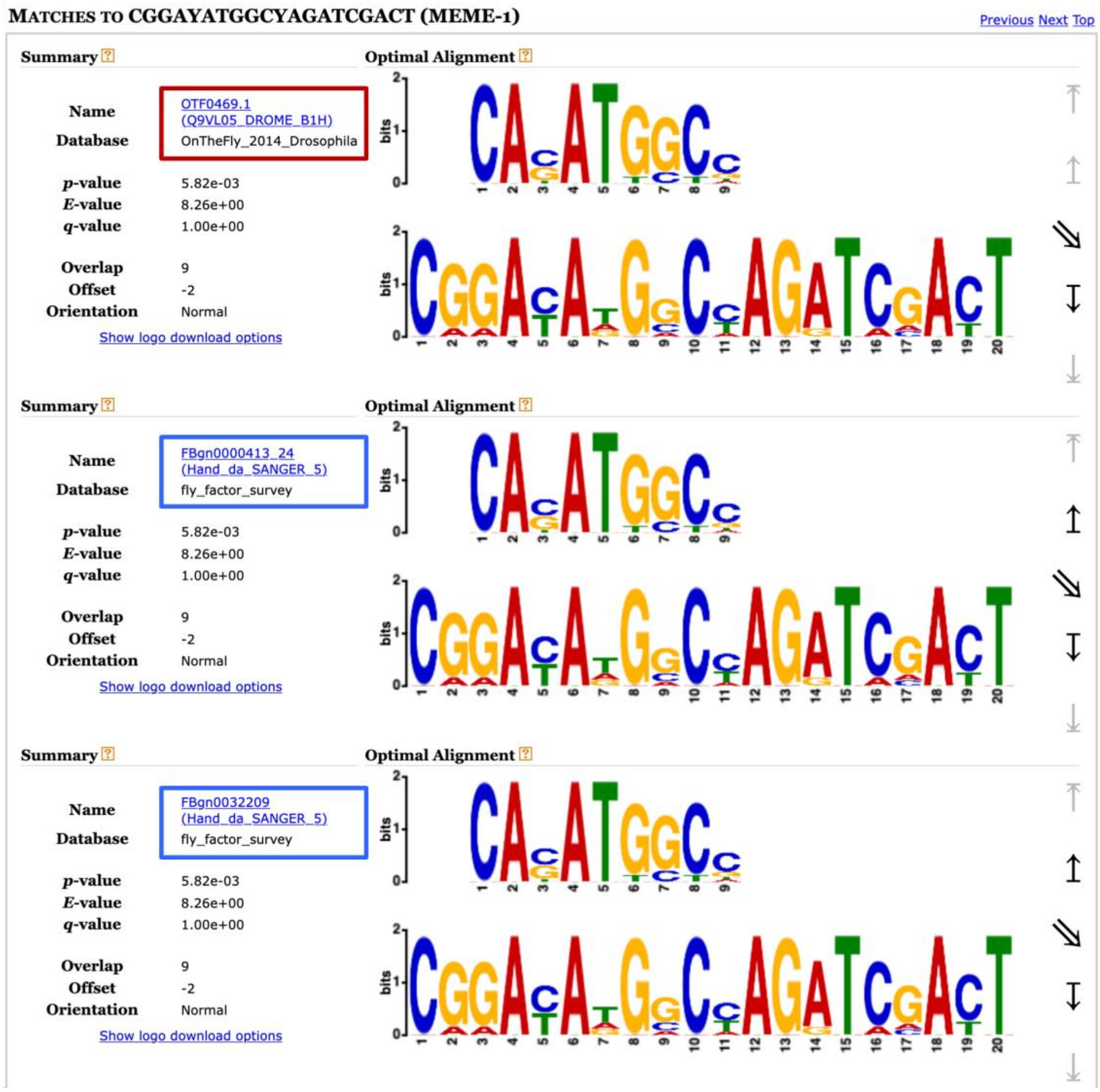
QUERY MOTIFS

Database ?	ID ?	Alt. ID ?	Preview ?	Matches ?	List ?
query_motifs	CGGAYATGGCYAGATCGACT	MEME-1		5	FBgn0000413_24 (Hand_da_SANGER_5), FBgn0032209 (Hand_da_SANGER_5),

Figure 26 *Tomtom* identified five matches to motif 1. Two of the matches have the same alternate name (Hand_da_SANGER_5). (Use the scrollbar in the "List" column to see the other three matches.)

Examination of the “**Target Databases**” section shows that three of the matches are to motifs in the Fly Factor Survey database and two of the matches are to motifs in the “OnTheFly_2014_Drosophila” database. The sequence logos under the “**Matches to 1**” section shows that two of the motif matches from the Fly Factor Survey database (FBgn0000413_24 and FBgn0032209) are identical.

In addition, the two motif matches to the “OnTheFly_2014_Drosophila” database are identical to the motif matches to the Fly Factor Survey database. For example, the sequence logo for the motif OTF0469.1 in the “OnTheFly_2014_Drosophila” database is the same as the FBgn0000413_24 and FBgn0032209 motifs in the Fly Factor Survey database (Figure 27). Similarly, the sequence logo for the “OnTheFly_2014_Drosophila” motif OTF0489.1 is identical to the Fly Factor Survey motif FBgn0035454. Hence there are only two distinct motifs in the *Tomtom* search results.



MATCHES TO CGGAYATGGCYAGATCGACT (MEME-1)

Previous Next Top

Summary

Name	OTF0469.1 (Q9VL05_DROME_B1H)
Database	OnTheFly_2014_Drosophila
p-value	5.82e-03
E-value	8.26e+00
q-value	1.00e+00
Overlap	9
Offset	-2
Orientation	Normal

Show logo download options

Summary

Name	FBgn0000413_24 (Hand_da_SANGER_5)
Database	fly_factor_survey
p-value	5.82e-03
E-value	8.26e+00
q-value	1.00e+00
Overlap	9
Offset	-2
Orientation	Normal

Show logo download options

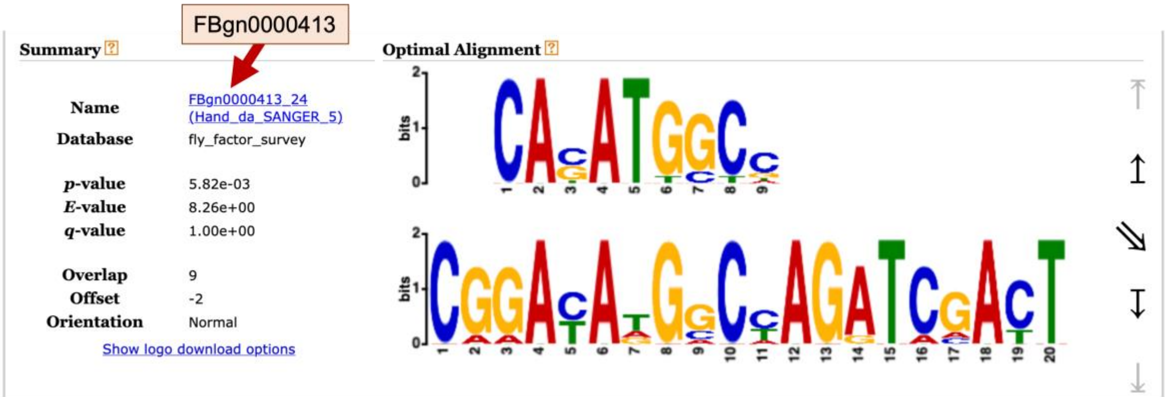
Summary

Name	FBgn0032209 (Hand_da_SANGER_5)
Database	fly_factor_survey
p-value	5.82e-03
E-value	8.26e+00
q-value	1.00e+00
Overlap	9
Offset	-2
Orientation	Normal

Show logo download options

Figure 27 The motif OTF0469.1 (top) from the “OnTheFly_2014_Drosophila” database has the same sequence logo as the motifs FBgn0000413_24 (middle) and FBgn0032209 (bottom) from the Fly Factor Survey database.

For motif records in the Fly Factor Survey database, the name of the motif includes the FlyBase ID for the gene associated with the motif. For example, for the Fly Factor Survey motif match **FBgn0000413_24**, the part of the name before the underscore (**FBgn0000413**) corresponds to the FlyBase gene ID (Figure 28).


FBgn0000413

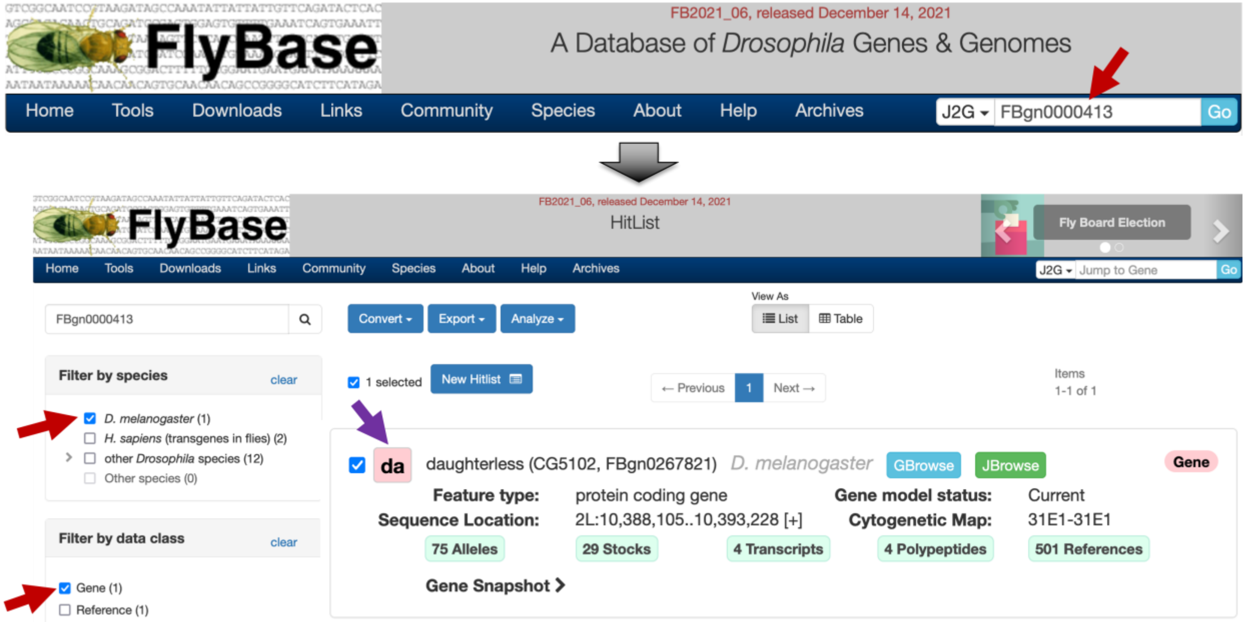
Summary

Name	FBgn0000413_24 (Hand_da_SANGER_5)
Database	fly_factor_survey
***p*-value**	5.82e-03
***E*-value**	8.26e+00
***q*-value**	1.00e+00
Overlap	9
Offset	-2
Orientation	Normal

Show logo download options

Figure 28 The "Matches to 1" section compares the sequence logo of the known motif in the motif database (top) against the motif identified by *MEME* (bottom). For records in the Fly Factor Survey database, the portion of the motif name before the underscore corresponds to the FlyBase gene ID (with the FBgn prefix).

To learn more about the gene associated with this motif, open a new web browser tab and navigate to FlyBase. Enter "**FBgn0000413**" into the "Jump to Gene" (J2G) search box on the top right corner of the navigation bar, and then click "**Go**". On the HitList page, select "***D. melanogaster***" under the "Filter by species" panel, and select "**Gene**" under the "Filter by data class" panel. Click on the "***da***" button to navigate to the gene record (Figure 29).


FB2021_06, released December 14, 2021

A Database of *Drosophila* Genes & Genomes

J2G FBgn0000413 Go

FB2021_06, released December 14, 2021

HitList

FBgn0000413

Filter by species clear

☑ *D. melanogaster* (1)
□ *H. sapiens* (transgenes in flies) (2)
□ other *Drosophila* species (12)
□ Other species (0)

Filter by data class clear

☑ Gene (1)
□ Reference (1)

☑ 1 selected New Hitlist

← Previous 1 Next →

Items 1-1 of 1

☑ da daughterless (CG5102, FBgn0267821) *D. melanogaster* GBrowse JBrowse Gene

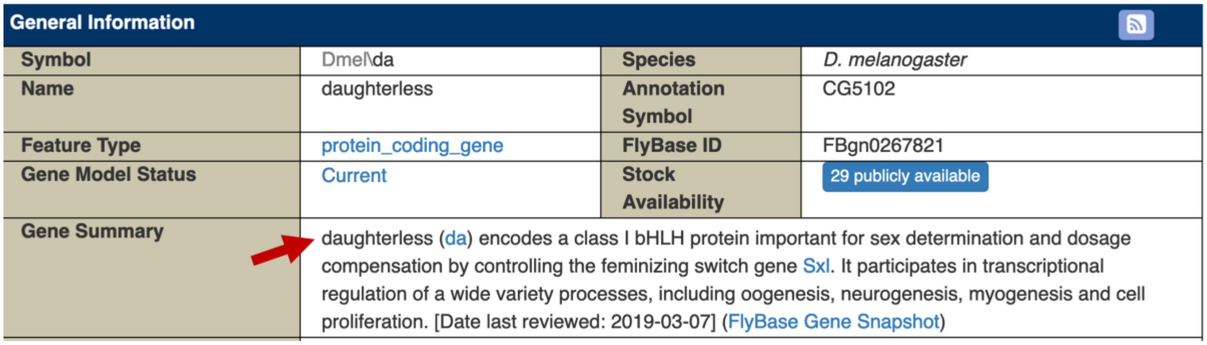
Feature type: protein coding gene **Gene model status:** Current
Sequence Location: 2L:10,388,105..10,393,228 [+] **Cytogenetic Map:** 31E1-31E1

75 Alleles 29 Stocks 4 Transcripts 4 Polypeptides 501 References

Gene Snapshot

Figure 29 Navigate to the FlyBase Gene Report associated with the first *Tomtom* motif match. (Top) Search for the FlyBase ID "FBgn0000413" using the "Jump to Gene" search box. (Bottom) After applying the species and data class filters (red arrows) to the FlyBase HitList, the search results page for "FBgn0000413" shows that the first *Tomtom* match corresponds to the binding site for the *da* (daughterless) gene. Click on the "*da*" button (purple arrow) to navigate to the corresponding FlyBase Gene Report.

The Gene Summary section of the FlyBase Gene Report for *da* indicates that the gene encodes a protein with a basic helix-loop-helix (bHLH) domain. This gene is involved in sex determination, dosage compensation, and in the transcriptional regulation of oogenesis, neurogenesis, myogenesis and cell proliferation (Figure 30).

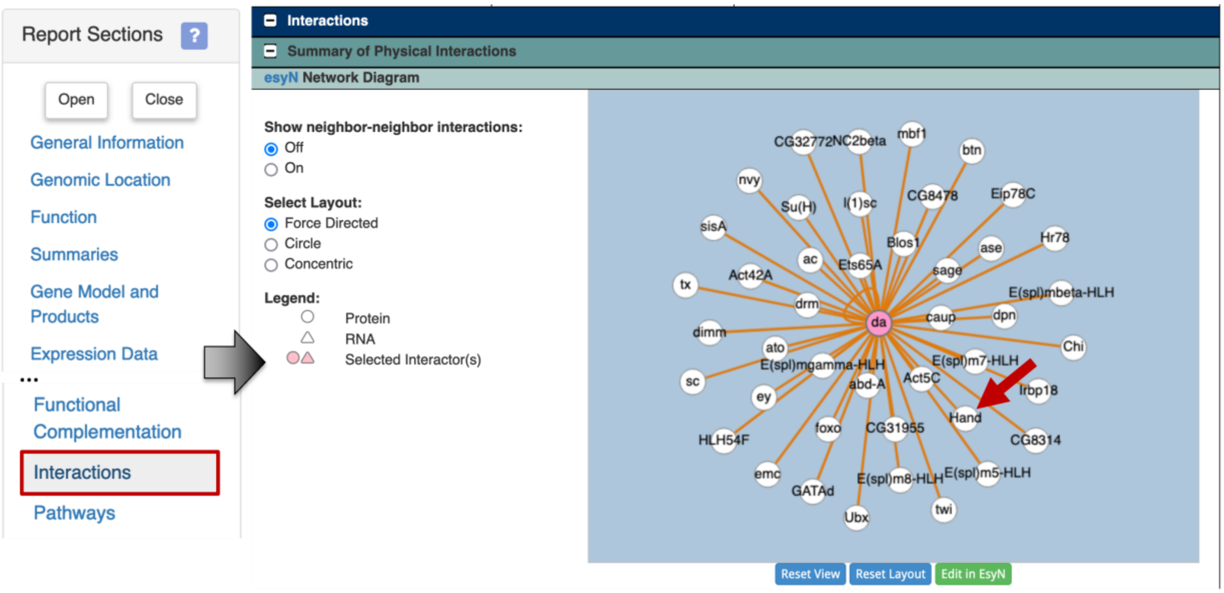


General Information

Symbol	Dmel\da	**Species**	*D. melanogaster*
Name	daughterless	**Annotation Symbol**	CG5102
Feature Type	protein_coding_gene	**FlyBase ID**	FBgn0267821
Gene Model Status	Current	**Stock Availability**	29 publicly available
Gene Summary	daughterless (da) encodes a class I bHLH protein important for sex determination and dosage compensation by controlling the feminizing switch gene Sxl. It participates in transcriptional regulation of a wide variety processes, including oogenesis, neurogenesis, myogenesis and cell proliferation. [Date last reviewed: 2019-03-07] (FlyBase Gene Snapshot)		

Figure 30 The Gene Summary section of the FlyBase Gene Report for the *da* gene shows that this gene plays an important role in sex determination, dosage compensation, and transcriptional regulation.

To determine the list of proteins that are known to interact with Daughterless, click on the "Interactions" link on the "Report Sections" sidebar. The esyN Network Diagram under the "Summary of Physical interactions" section shows that Daughterless has known physical interactions with multiple proteins. One of the proteins which interact with the Daughterless protein is Hand (Figure 31).



Report Sections

Open Close

General Information
Genomic Location
Function
Summaries
Gene Model and Products
Expression Data
...
Functional Complementation
Interactions
Pathways

Interactions

Summary of Physical Interactions

esyN Network Diagram

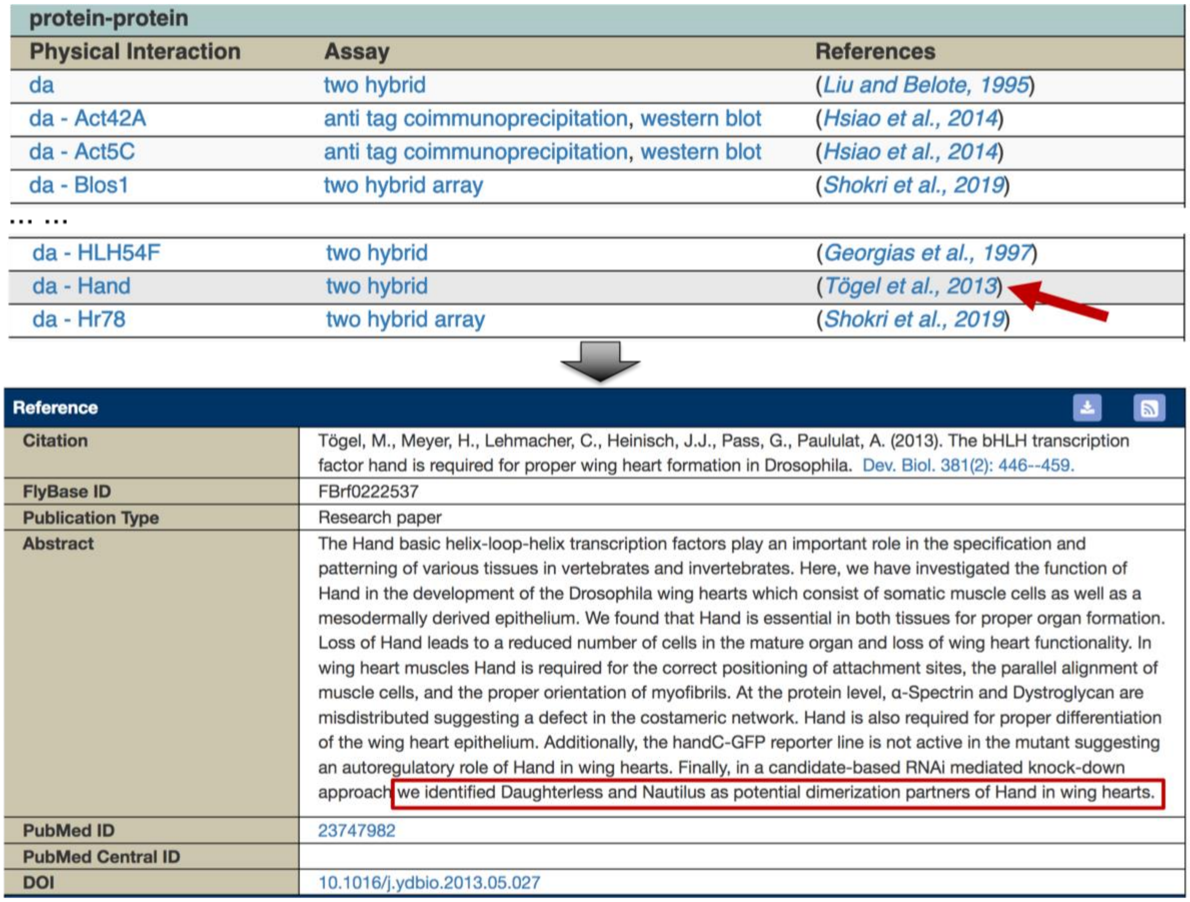
Show neighbor-neighbor interactions:
● Off
○ On

Select Layout:
● Force Directed
○ Circle
○ Concentric

Legend:
○ Protein
△ RNA
●▲ Selected Interactor(s)

Figure 31 The "Summary of Physical Interactions" section of the FlyBase Gene Report shows that *da* interacts with *Hand* (red arrow).

To learn more about the interaction between Daughterless and Hand, scroll down to the table under the "protein-protein" section (Figure 32). The table shows that the physical interaction between Daughterless and Hand (da - Hand) was identified using the two hybrid assay by Tögel and colleagues (Tögel *et al.* 2013; FBrf0222537). This study suggests that Daughterless and Hand are dimerization partners, and they are both required for the development of *Drosophila* wing hearts.



protein-protein

Physical Interaction	Assay	References
da	two hybrid	(*Liu and Belote, 1995*)
da - Act42A	anti tag coimmunoprecipitation, western blot	(*Hsiao et al., 2014*)
da - Act5C	anti tag coimmunoprecipitation, western blot	(*Hsiao et al., 2014*)
da - Blos1	two hybrid array	(*Shokri et al., 2019*)
… …		
da - HLH54F	two hybrid	(*Georgias et al., 1997*)
da - Hand	two hybrid	(*Tögel et al., 2013*)
da - Hr78	two hybrid array	(*Shokri et al., 2019*)

Reference

Citation	Tögel, M., Meyer, H., Lehmacher, C., Heinisch, J.J., Pass, G., Paululat, A. (2013). The bHLH transcription factor hand is required for proper wing heart formation in Drosophila. Dev. Biol. 381(2): 446--459.
FlyBase ID	FBrf0222537
Publication Type	Research paper
Abstract	The Hand basic helix-loop-helix transcription factors play an important role in the specification and patterning of various tissues in vertebrates and invertebrates. Here, we have investigated the function of Hand in the development of the Drosophila wing hearts which consist of somatic muscle cells as well as a mesodermally derived epithelium. We found that Hand is essential in both tissues for proper organ formation. Loss of Hand leads to a reduced number of cells in the mature organ and loss of wing heart functionality. In wing heart muscles Hand is required for the correct positioning of attachment sites, the parallel alignment of muscle cells, and the proper orientation of myofibrils. At the protein level, α-Spectrin and Dystroglycan are misdistributed suggesting a defect in the costameric network. Hand is also required for proper differentiation of the wing heart epithelium. Additionally, the handC-GFP reporter line is not active in the mutant suggesting an autoregulatory role of Hand in wing hearts. Finally, in a candidate-based RNAi mediated knock-down approach we identified Daughterless and Nautilus as potential dimerization partners of Hand in wing hearts.
PubMed ID	23747982
PubMed Central ID	
DOI	10.1016/j.ydbio.2013.05.027

Figure 32 The "protein-protein" section of the FlyBase Gene Report for *da* shows that the physical interactions between Daughterless and Hand was determined by a study from Tögel and colleagues.

This explains why the motif identified by *MEME* has the label **Hand_da_Sanger_5**, and matches to both FBgn0000413_24 and FBgn0032209 in the *Tomtom* search result. The FlyBase ID for the *Hand* gene is FBgn0032209.

We can apply the same approach to investigate the other motif match [FBgn0035454 (CG12029_SANGER_10)] identified by the *Tomtom* search. This investigation is left as an exercise for the reader.

Conclusions

This walkthrough illustrates the use of multiple web databases and tools to identify conserved motifs in a collection of sequences upstream of the transcription start site. Using the FlyBase RNA-Seq Expression Profile Search tool, we identified a set of genes with similar expression patterns. Using the FlyBase search results table, we can apply additional filters to this list of genes (e.g., by focusing on only dot chromosome genes or genes with a particular gene ontology term). In addition, we can export the gene list in many different formats for downstream analyses.

We then used the FlyBase Gene Report and *JBrowse* to determine the coordinates of the regions upstream of the TSS. We can retrieve the genomic sequences for these upstream regions using the FlyBase *Sequence Downloader* tool.

Running *MEME* on this collection of upstream sequences allow us to identify over-represented motifs. We can examine the distribution of the motif instances using the *MAST* tool. To characterize this set of motifs, we used *Tomtom* to compare the motifs discovered by *MEME* against multiple databases of known motifs in *Drosophila*. For *MEME* motifs that match known motifs in the motif databases (e.g., Fly Factor Survey), we can further investigate the set of proteins (e.g., transcription factors) that are associated with these specific binding sites.

Bibliography

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 *MEME SUITE*: tools for motif discovery and searching. Nucleic Acids Res. 37: W202-208.

Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin *et al.*, 2011 The developmental transcriptome of *Drosophila melanogaster*. Nature 471: 473–479.

Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, 2007 Quantifying similarity between motifs. Genome Biol. 8: R24.

Stewart, A. J., S. Hannenhalli, and J. B. Plotkin, 2012 Why transcription factor binding sites are ten nucleotides long. Genetics 192: 973–985.

Tögel, M., H. Meyer, C. Lehmacher, J. J. Heinisch, G. Pass *et al.*, 2013 The bHLH transcription factor hand is required for proper wing heart formation in *Drosophila*. Dev. Biol. 381: 446–459.

Zhu, L. J., R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh *et al.*, 2011 FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res. 39: D111-117.