Motif Discovery in *Drosophila*

*Wilson Leung*

**Prerequisites**
- Annotation of Transcription Start Sites in *Drosophila*

**Resources**

<table>
<thead>
<tr>
<th>Web Site</th>
<th>Web Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlyBase</td>
<td><a href="http://flybase.org">http://flybase.org</a></td>
</tr>
<tr>
<td>FlyFactorSurvey</td>
<td><a href="http://mccb.umassmed.edu/ffs/">http://mccb.umassmed.edu/ffs/</a></td>
</tr>
</tbody>
</table>

**Files for this Tutorial**
The exercise package is available on the GEP web site under Curriculum ➔ Course Materials ➔ Washington University ➔ Beyond Annotation

**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 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 ([http://flybase.org](http://flybase.org)) 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 FlyBase ([http://flybase.org](http://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.

![FlyBase Home Page](flybase.png)

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 time points, 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 tissue expression only" button to run the search.
The search results page shows 302 *D. melanogaster* genes that are expressed only in head tissues (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 "Results Analysis/Refinement" button to filter this list of genes and use the "HitList Conversion Tools" button to export the gene list.

**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 so we need to apply a filter to this initial gene list. Click on the "Results Analysis/Refinement" button in the gene list page and select "Chromosome arm" (Figure 4).
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 302 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 our original list of genes so that only dot chromosome genes are in our new gene list (Figure 6).

Figure 4 Use the Results Analysis/Refinement button to filter the gene list by chromosomes

Figure 5 Of the 302 genes that are expressed only in head tissues, 12 of them are located on the dot chromosome.

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

We could 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.
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 interests 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 upstream sequences are in the file `dot_genes_head_expressed.fasta` in the exercise package.)

Open a new tab and navigate to FlyBase ([http://flybase.org](http://flybase.org)). Enter the name of the first gene in the gene list (e.g., Actbeta) into "Jump to Gene" search box on the top right corner of the main navigation bar and then click "Go" (Figure 7).

![FlyBase](http://flybase.org)

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

The GBrowse 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.

![Gene Report](http://flybase.org)

*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 "GBrowse" link to view the genomic region surrounding the Actbeta gene.*
The GBrowse 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 "GBrowse" link (under "Genomic Maps") and then select the "Show 20 kbp" option under the "Scroll/Zoom" field (Figure 9).

![Figure 9 Select the "Show 20 kbp" option to show the genomic region surrounding the Actbeta gene.](image)

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.

![Figure 10 FlyBase gene report shows that the sv gene begins at 1,088,798 on the dot chromosome.](image)
We can apply the same procedure 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:

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

**Retrieve the upstream sequences**

Once we have generated the list of coordinates for the upstream regions, we can use the FlyBase "Batch Download" tool to retrieve the genomic sequences for these upstream regions. You can access the "Batch Download" tool through the FlyBase main navigation bar (under "Tools" ➔ "Retrieve/Convert Tools" ➔ "Batch Download", Figure 11).

![FlyBase](image)

*Figure 11 Use the main navigation bar to access the FlyBase Batch Download tool.*

Change the "Output Format" field to "FASTA Sequence", the "Output Options" field to "By sequence coordinates" and the "Send results to" field to "File". Copy and paste the list of upstream sequence coordinates into the "Enter IDs, Symbols, or Sequence Coordinates" text box (Figure 12). Click on the "Get FastA" button to run the program. Save the output file as **dot_genes_head_expressed.fasta**.

*Depending on the web browser settings, clicking on the "Get FastA" button might download the file onto your Desktop or Downloads folder. You could either use the file (FlyBase_FastA.txt) directly or rename the file to dot_genes_head_expressed.fasta.*
While not required, we will modify the definition lines (i.e. lines that begin with the `>` symbol) in the FASTA file so that we can more easily interpret the results during the motif discovery analysis. Open the FASTA file in a text editor (e.g., WordPad on MS Windows, TextEdit on Mac OS X) and then add the gene name to the beginning of each definition line.

For example, we will prepend the gene name "Actbeta" to the definition line for the feature at 1084797-1088797:

```
>Actbeta 4:1084797..1088797 species=Dmel; strand='+'; 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 Batch Download tool, we are ready to 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 web site [http://meme-suite.org/](http://meme-suite.org/) and click on the "MEME" icon (Figure 13). We will perform the initial MEME search using mostly default parameters. Verify that the "Select the motif discovery mode" field is set to "Normal 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**." We should also verify 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 \((\text{Stewart et al. 2012})\), 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).

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).
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 are available in the \texttt{dot\_head\_expressed\_MEME.html} file (inside the \texttt{MEME\_results} folder) in the exercise package.

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.

Figure 16 Links to the MEME search results. Right click on the links to save the output files or to open the HTML output in a new tab.

Right click (control-click on the Mac) 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 input sequences that were used to construct each motif.
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).

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

Because transcription factors usually 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.

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 the Mac) 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.

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).
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" and "Feb 27, 2017" 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.

Figure 22 Configure MAST to search the motifs discovered by MEME against the collection of sequences upstream of the transcription start sites in *D. pseudoobscura*. 
Once the MAST search is complete, we can right click (or control-click on the Mac) on the "MAST html output" link and open the MAST search results page in a new tab to examine the results. The "Search Results" section shows MAST has detected 40 *D. pseudoobscura* upstream sequences with significant (E-value < 10) matches to the first motif discovered by MEME (Figure 23).

![Figure 23 Upstream *D. pseudoobscura* sequences with significant matches to the first motif identified by MEME.](image)

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., XP_002138657.1) corresponds to the RefSeq accession number and the second part of the name (e.g., Dpse\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).

![Figure 24 Use Tomtom to search a motif discovered by MEME against a database of known motifs.](image)
In this example, we will search the first motif identified by MEME against a database 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, FlyFactorSurvey, 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 FlyFactorSurvey database were identified primarily using the bacterial one-hybrid (B1H) system (Zhu et al. 2011).

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, 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.

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 two of the matches are to motifs in the "OnTheFly 2014 Drosophila" database and three of the matches are to motifs in the FlyFactorSurvey database. Examination of the sequence logos under the "**Matches to 1**" section shows that the motif matches to the "OnTheFly 2014 Drosophila" database are identical to the motif matches to the FlyFactorSurvey database.

For example, the motif OTF0469.1 in the “OnTheFly 2014 Drosophila” database is the same as the FBgn0000413_24 motif in the FlyFactorSurvey database (Figure 27). Two of the motif matches to the FlyFactorSurvey database (FBgn0000413_24 and FBgn0032209) also have the same sequence logo. Similarly, the sequence logo for the motif OTF0489.1 in the “OnTheFly 2014 Drosophila” database is identical to the FlyFactorSurvey motif FBgn0035454. Hence there are only two distinct motifs in the Tomtom search results.

We can click on the link under the **Name** field to learn more about each motif. For example, clicking on the FBgn0000413_24 link (Figure 28) will bring us to the corresponding motif record in the FlyFactorSurvey database (Figure 29).
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). Click on the link under the Name field to obtain additional information about the motif.

Figure 29 The second Tomtom match corresponds to the binding site for the daughterless gene.

We find that this motif corresponds to known DNA binding sites for the daughterless (da) gene and that the DNA binding domain of this transcription factor has a helix-loop-helix (HLH) structure. We can learn more about this gene by clicking on the "P11420" link under the "Uniprot ID" row. The "GO - Biological process" section of the UniProt record shows that the daughterless gene is involved in nervous system development and in sex determination (Figure 30).

Figure 30 The Gene Ontology (GO) section of the UniProt gene record for the daughterless gene shows that this gene plays an important role in the development of the nervous system and in sex determination.
Go back to the FlyFactorSurvey tab. In addition to the gene summary, the FlyFactorSurvey entry also lists the known heterodimeric partners of da and their corresponding motifs. In this case, the actual motif match identified by Tomtom is called **Hand_da_Sanger_5**.

Scroll down to the row in which the motif Hand_da_Sanger_5 is listed (approximately at the middle of the page), we find that the motif is a binding site for the heterodimer of *daughterless* and *Hand* (Figure 31). This explains why the motif identified by MEME matches to both FBgn0000413.24 and FBgn0032209 in the Tomtom search result; FBgn0032209 is the accession number for the *Hand* gene.

---

### Figure 31
The motif Hand_Da_SANGER_5 is a DNA binding site for the daughterless and Hand heterodimer.

---

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 GBrowse to determine the coordinates of the regions upstream of the TSS. We can retrieve the genomic sequences for these upstream regions using the FlyBase Batch Download 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.*, FlyFactorSurvey), we can further investigate the set of proteins (*e.g.*, transcription factors) that are associated with these specific binding sites.
Bibliography


