Using FlyBase RNA-Seq Tools to Investigate Gene Expression Profiles

Wilson Leung

Introduction

RNA sequencing (RNA-Seq) uses high-throughput second or third generation sequencing technologies to quantify the expression levels of transcripts within a genome under different conditions (e.g., developmental stages, tissues, treatments). The modENCODE project has produced a large collection of RNA-Seq data for *Drosophila melanogaster* (Graveley *et al.* 2011; Brown *et al.* 2014), and these datasets have been incorporated into FlyBase (St Pierre *et al.* 2014). In this walkthrough, we will use *GBrowse* and the RNA-Seq tools on FlyBase to examine the expression profiles of the gene *Gyf* on the *D. melanogaster* 4th chromosome. We will also use the FlyBase RNA-Seq Expression Similarity Search tool and the RNA-Seq Expression Profile Search tool to identify other *D. melanogaster* genes that exhibit expression profiles that are similar to *Gyf*. Our aim is to illustrate the kinds of analyses that can be done, and questions that can be explored, by generating and using genome-wide records of gene expression data with RNA-Seq.

Use *GBrowse* to visualize the expression profiles of *Gyf*

A past study has shown that the *Gyf* gene is involved in the regulation of autophagy in *D. melanogaster* (Kim *et al.* 2015). Mutations in its putative human ortholog *GIGYF2* have previously been associated with Parkinson's disease (Zhang *et al.* 2015) and Schizophrenia (Ripke *et al.* 2013).

To learn more about the *Gyf* gene, open a web browser and navigate to FlyBase (https://flybase.org). Enter "*Gyf*" into the "Jump to Gene" (J2G) search box at the right side of the main navigation bar, and then click on the "Go" button (Figure 1).

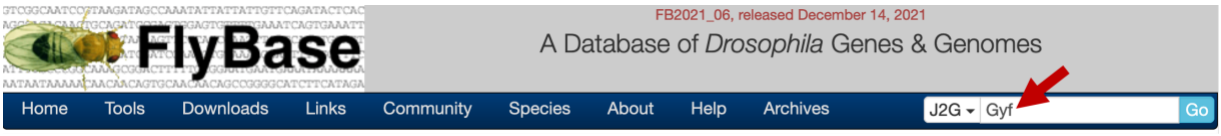

Figure 1 Use the "Jump to Gene" (J2G) search box to retrieve the FlyBase Gene Report for the *Gyf* gene.

The "Sequence location" field of the FlyBase Gene Report shows that *Gyf* is located at 853,557-867,025 on the 4th chromosome. FlyBase has incorporated RNA-Seq data produced by the modENCODE project, the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC), and the Knoblich lab into *GBrowse*. To visualize the RNA-Seq data for *Gyf*, click on the "*GBrowse*" button under the "Genomic Location" section (Figure 2).

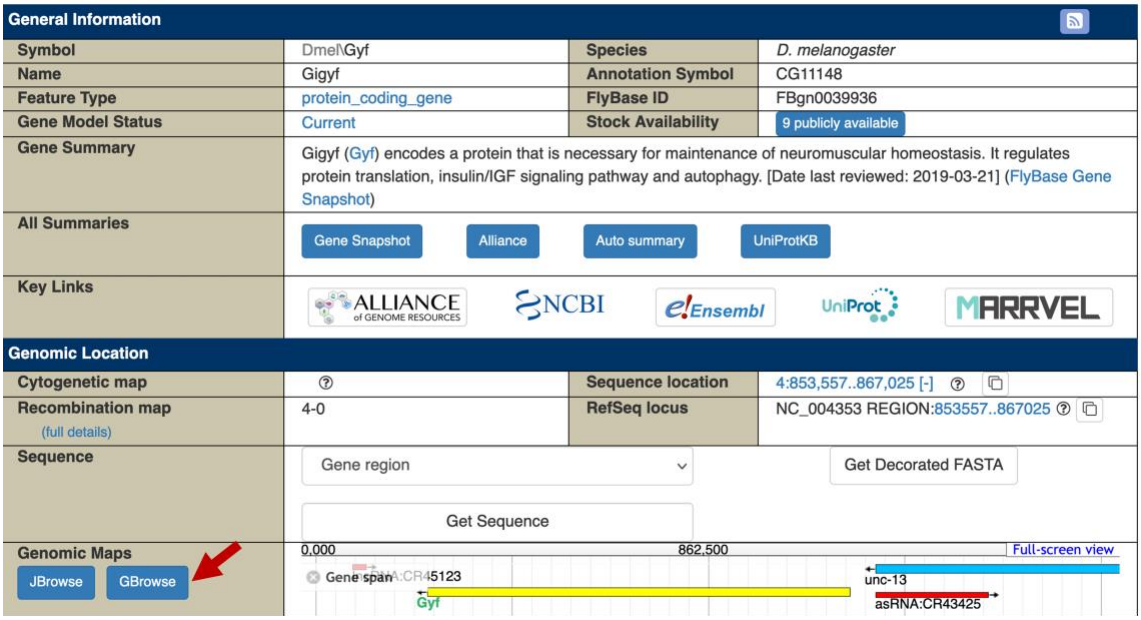

Figure 2 Click on the "*GBrowse*" button under the "Genomic Maps" header to view the *Gyf* gene in *GBrowse*.

Because *GBrowse* remembers the previous track display settings, we will hide all the evidence tracks and then turn on the subset of evidence tracks that we will use in this walkthrough. Click on the "Select Tracks" tab in *GBrowse* and then select the "All off" checkbox next to each section header. Select the checkbox next to the following evidence tracks (Figure 3):

Under "**Reference Genome Annotations (iso-1)**":

- Gene Span
- Transcript
- Natural TE

Under "**Expression Levels**" → "**RNA-Seq**":

- RNA-Seq, Developmental stages, unstranded (modENCODE)

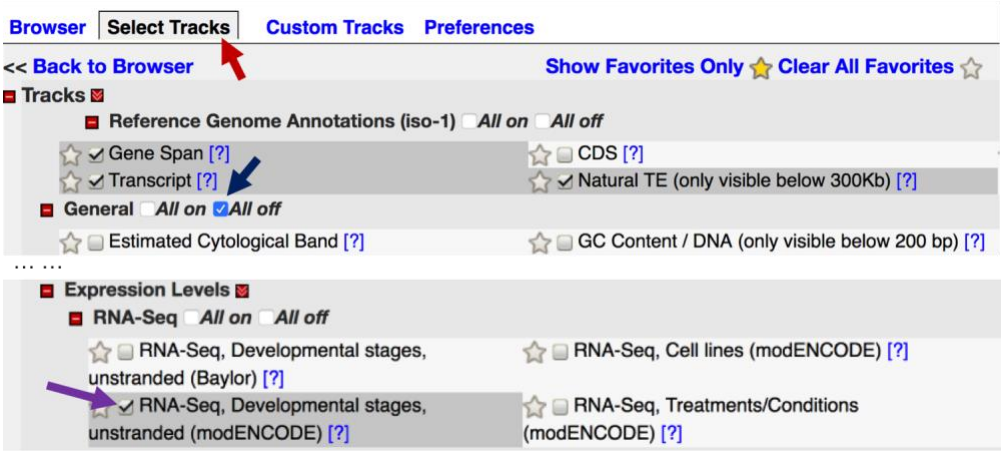

Figure 3 Click on the "Select Tracks" tab to configure the *GBrowse* display (red arrow). Select the "All off" checkbox next to the section headers (blue arrow) to hide all the evidence tracks in that section. Select the checkbox next to the evidence track to show the track in *GBrowse* (purple arrow). The RNA-Seq evidence tracks are available under the "Expression Levels" section.

Click on the "Back to Browser" button at the bottom of the page to return to the *GBrowse* view (Figure 4). The "Transcript" evidence track on *GBrowse* shows that there are four isoforms of *Gyf* in *D. melanogaster* (i.e. isoforms D, F, G, and H). For each isoform, the boxes demarcate the transcribed exons and the lines denote introns. The gray color within the transcribed exons corresponds to untranslated regions and the orange color corresponds to coding regions.

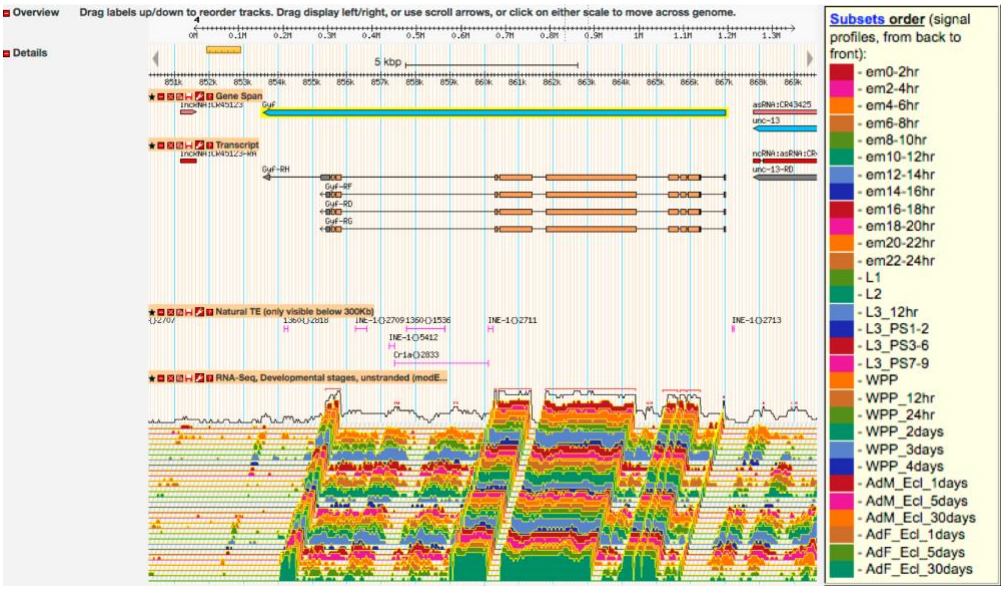

Figure 4 ***GBrowse*** **view of the** ***Gyf*** **gene with the modENCODE RNA-Seq signal profiles from 30 developmental stages. (Right) Hover the mouse over the "RNA-Seq, Developmental stages, unstranded" track to see a tooltip that describes the order of the developmental stages subtracks (listed from back to front). Hover the mouse over this tooltip to keep the legend visible.**

The "RNA-Seq, Developmental stages, unstranded" track shows the RNA-Seq signal profiles (i.e. expression levels) of *Gyf* in 30 stages of development (Figure 4, right). FlyBase *GBrowse* uses the TopoView glyph to display RNA-Seq data. The tilted (3D) display mode provides a compact way to display multiple RNA-Seq datasets, and it facilitates the visual comparisons of the expression profiles from a large number of RNA-Seq samples. By default, the expression profiles are shown in $\log_2$ scale. The black graph at the back of the TopoView track shows the maximum expression level at each genomic position among all of the subtracks (e.g., the 30 developmental stages sampled by modENCODE). The red lines above this maximum expression track corresponds to predicted exons based on the maximum RNA-Seq signal profiles (Figure 5).

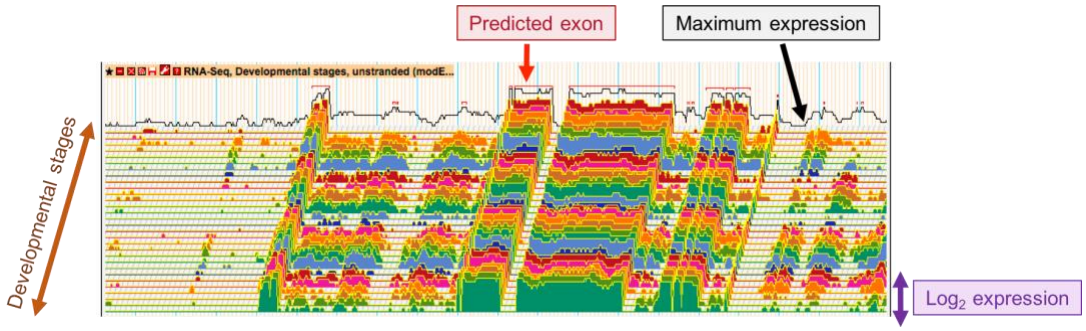

Figure 5 Key features of the TopoView glyph used to display RNA-Seq data on FlyBase ***GBrowse*****. Each color subtrack in the TopoView track depicts the expression profile of one of the developmental stages sampled by modENCODE. The y-axis of each subtrack shows the $\log_2$ expression level of that sample. The black graph at the back of the TopoView track shows the maximum expression level among all subtracks. The red lines demarcate predicted exons based on the maximum expression level.**

See the "Expression Levels" section of the "*GBrowse* Tracks" page on the FlyBase wiki for a more detailed explanation of the available RNA-Seq evidence tracks.

Examination of the "RNA-Seq, Developmental stages, unstranded" track shows that the RNA-Seq expression profiles are generally in congruence with the locations of the transcribed exons of *Gyf*. [The precursor mRNA, which is continuous from the Transcription Start Site (TSS) to termination, is rapidly processed to remove the introns, so intron sequence is depleted in the RNA-Seq data.] However, the largest intron, between exons Gyf:9 and Gyf:10 (at 855,841-860,306), shows low levels of expression (Figure 6, red arrows). This intron contains multiple transposons (see the "Natural TE" track) and the regions with the most RNA-Seq coverage overlap with these transposons. Hence there is insufficient evidence to support the hypothesis that these regions are being transcribed and incorporated into the processed mRNA, as the transcripts could have been derived from other copies of the transposons within the genome.

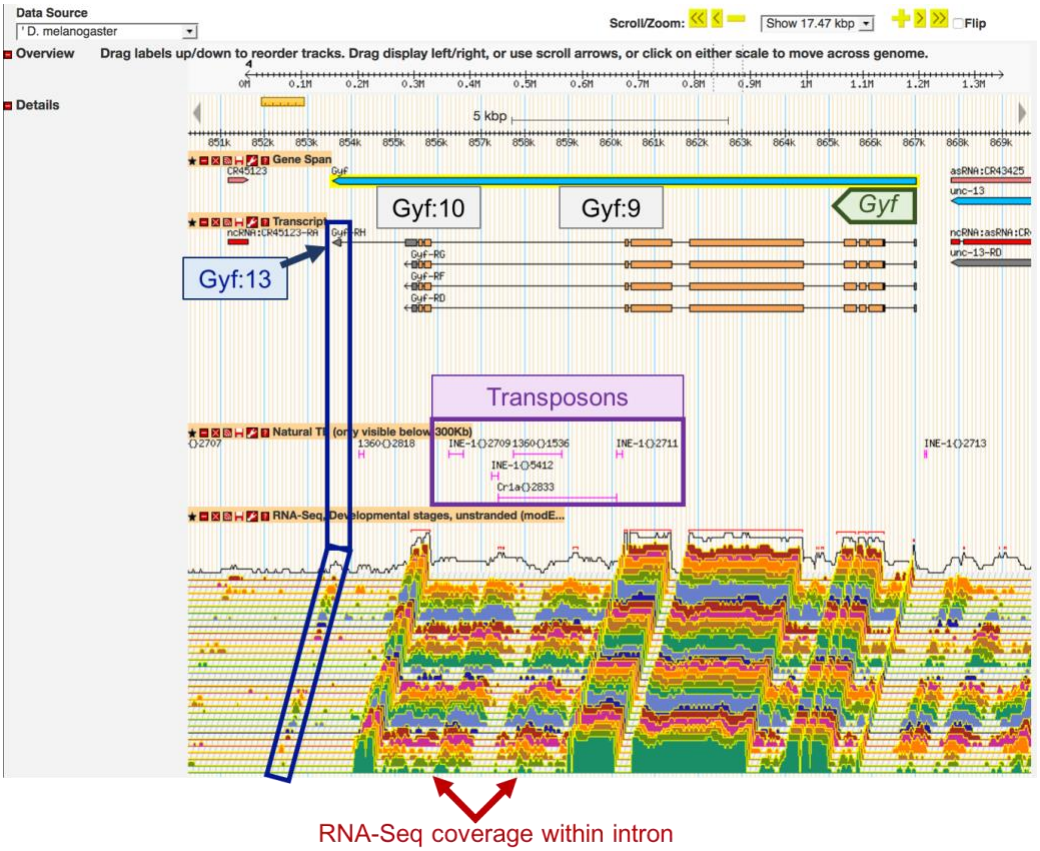

Figure 6 Comparison of RNA-Seq expression profiles from 30 developmental stages with the genomic features annotated by FlyBase. The intron between the transcribed exons Gyf:9 and Gyf:10 does show low RNA-Seq read coverage (red arrows), but the RNA-Seq read coverage could be attributed to the transposon remnants (i.e. *INE-1*, *1360*, and *Cr1a*) within the intron (purple box). The last transcribed exon of Gyf-RH (Gyf:13) is expressed only in a subset of the developmental stages (blue box).

The RNA-Seq evidence track also shows that the last transcribed exon of the H isoform of *Gyf* (Gyf-RH) is only expressed in a subset of developmental stages (Figure 6, blue box). To investigate this region more closely, enter "4:853000-854000" into the "Landmark or Region" field underneath the "Browser" tab and then click on the "Display Region" button (Figure 7).

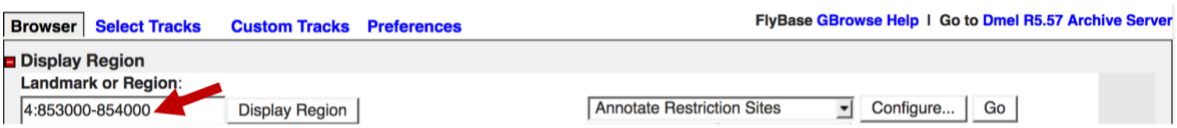

Figure 7 Enter "4:853000-854000" into the "Landmark or Region" field to navigate to the region surrounding the last transcribed exon of Gyf-RH (Gyf:13).

We can change the display options for the TopoView track so that it only shows the expression profiles of a subset of the developmental stages. For example, we can configure the TopoView track to compare the expression profiles of Gyf:13 during the first 6 hours of embryogenesis.

Click on the wrench icon ( ) in the "RNA-Seq, Developmental stages, unstranded" track label to open the track configuration panel (Figure 8, left). Select the "em0-2hr", "em2-4hr", and "em4-6hr" options under the "Select samples to show" field. Change the "Samples presentation style" to "Vertical". Increase the "Vertical spacing between samples (pixels)" to "30" in order to ensure that the signals from the different samples will not overlap with each other. Click on the "Apply changes" button to update the *GBrowse* display (Figure 8, left). The updated TopoView track shows that Gyf:13 is not expressed (or expressed at very low levels) in the 0-2 hour and the 2-4 hour embryos, but it is expressed at a higher level in the 4-6 hour embryos (Figure 8, brown arrow).

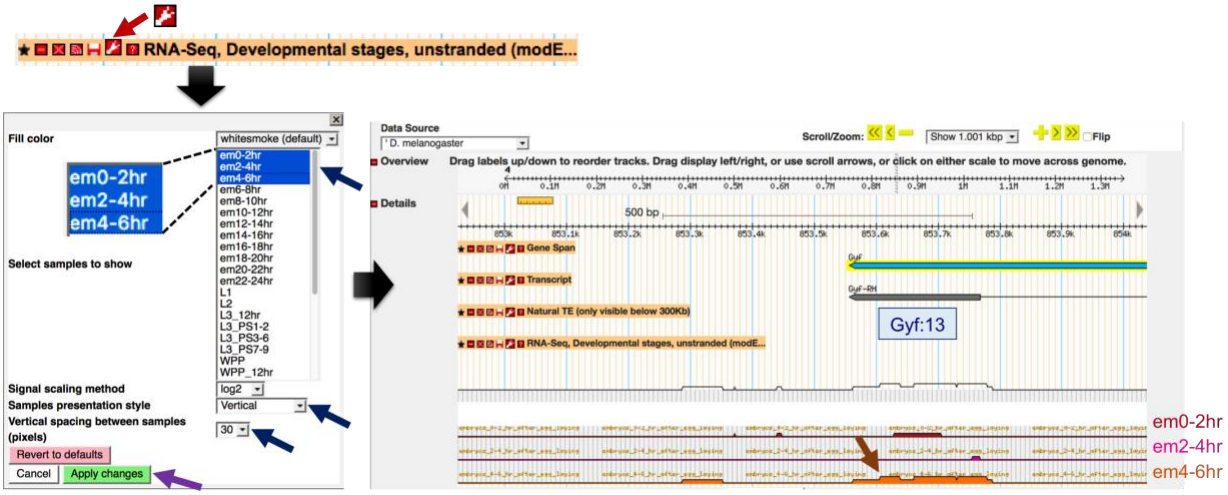

Figure 8 Configure the TopoView track to display the expression profiles during the first 6 hours of embryogenesis. Click on the wrench icon in the track label to configure the track (red arrow). We can use the configuration pane to select the samples to show and the presentation style (blue arrows). Click on the "Apply changes" button to update the display (purple arrow). The updated track shows Gyf:13 is expressed in the 4-6 hour embryos (brown arrow) but not in the 0-2 hour or 2-4 hour embryos.

The default "Tilted" presentation style of the TopoView track provides a compact overview of the expression profiles across a large number of samples. The "Vertical" presentation style shows the expression level at each genomic position without the horizontal offset, which makes it easier to compare the expression levels of individual exons. See the "*GBrowse*/*JBrowse*" section of the "FlyBase:RNA-Seq Overview" page on the FlyBase wiki for details.

Examine strand-specific RNA-Seq data using *GBrowse*

Because mutations in the human ortholog of *Gyf* (*GIGYF2*) has previously been implicated in neurological disorders such as Parkinson's disease (Zhang *et al.* 2015) and Schizophrenia (Ripke *et al.* 2013), we would like to examine the expression profiles of *Gyf* in the central nervous system (CNS) and adult heads.

Enter "*Gyf*" into the "Landmark or Region" field and then click on the "Display Region" button to zoom out to the entire gene span of *Gyf*. Click on the "Select Tracks" tab, then scroll down to the "Expression Levels" section. Select the checkbox next to the "RNA-Seq, CNS and adult head (modENCODE)" label under the "RNA-Seq by Tissue" section (Figure 9). Click on the "Back to Browser" button at the bottom of the page to view this RNA-Seq TopoView track in *GBrowse*.

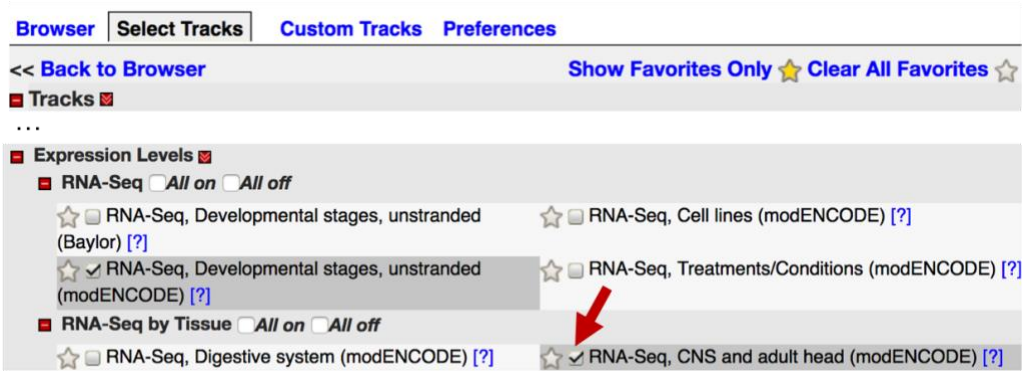

Figure 9 Select the "RNA-Seq, CNS and adult head (modENCODE)" checkbox under the "RNA-Seq by Tissue" section to display the RNA-Seq data for CNS and adult heads on FlyBase *GBrowse*.

The "RNA-Seq, CNS and adult head (modENCODE)" track will appear above the "Gene Span" track. To facilitate the interpretation of the expression profiles from these samples, we will rearrange the evidence tracks so that this RNA-Seq track appears below the "Transcript" track.

Move the mouse so that it hovers over the "RNA-Seq, CNS and adult head (modENCODE)" track label and the mouse cursor changes into a "move" icon (Figure 10). Press and hold onto the (left) mouse button and then drag the track down until this track is located below the "Transcript" track. Release the mouse button to complete the reordering of the track.

Figure 10 A move icon will appear when the mouse hovers over the track title. Drag the track label up or down within the *GBrowse* view to reorder the evidence tracks.

The default "Tilted" presentation style of the RNA-Seq TopoView track assumes the track is **placed below the evidence tracks that depict other genomic features** (e.g., Gene Span, Transcript, Natural TE). The expression profiles in the "RNA-Seq, CNS and adult head (modENCODE)" track initially appear to be inconsistent with the locations of the transcribed exons of *Gyf* because the track was placed above the "Transcript" track.

Similar to the "RNA-Seq, Developmental stages, unstranded" track, a legend that describes the list of subtracks in the "RNA-Seq, CNS and adult head (modENCODE)" TopoView track will appear when the mouse hovers over this track (Figure 11, right). The legend shows that this track consists of RNA-Seq data from eleven samples (i.e. CNS from 3rd instar larvae and pupal stage P8, heads from 1-day, 4-days, and 20-days old mated males, mated females, and virgin females).

Because the modENCODE RNA-Seq datasets for the CNS and adult heads are strand-specific, the TopoView track is divided into two sections. The top section shows the expression levels on the plus strand while the bottom section shows the expression levels on the minus strand (Figure 11). Since *Gyf* is located on the minus strand of the 4th chromosome, most of the RNA-Seq signal is in the bottom section of the TopoView track. These subtracks show that *Gyf* is expressed in the eleven CNS and adult heads RNA-Seq samples sequenced by modENCODE.

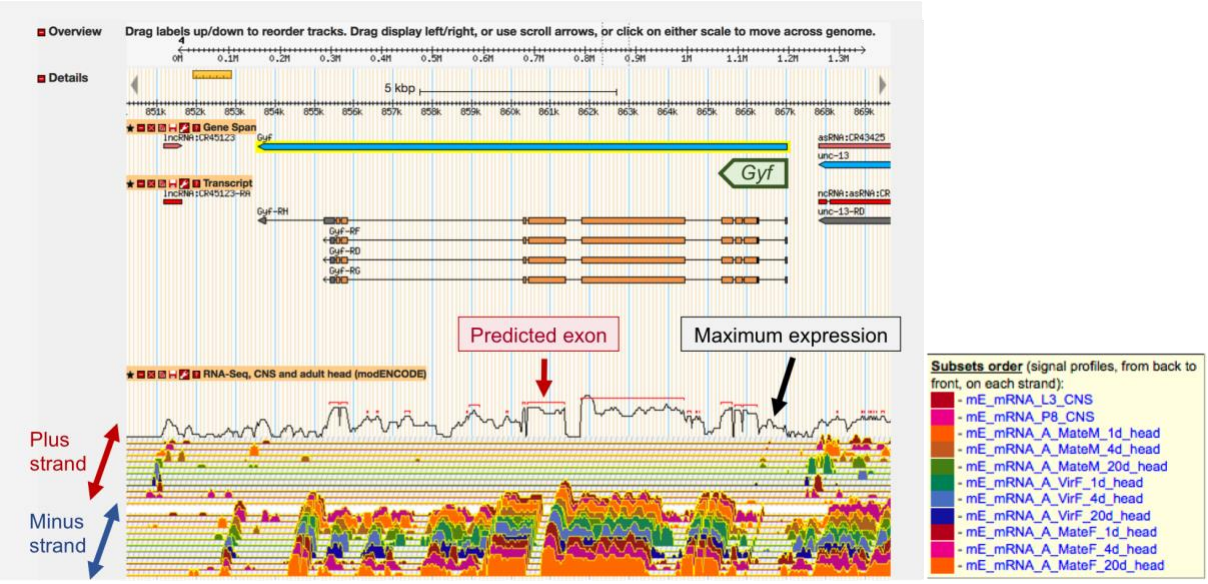

Figure 11 ***GBrowse*** **TopoView track for the strand-specific RNA-Seq data from 11 CNS and adult heads samples. (Right) The black graph at the back of the TopoView track depicts the maximum expression level among all of the subtracks in either strand. The red lines above the maximum expression track corresponds to predicted exons. For strand-specific RNA-Seq data (which uses a library preparation protocol that preserves the orientation of the transcripts), the TopoView track is divided into two sections. The top section shows the expression profiles on the plus strand and the bottom section shows the expression profiles on the minus strand. Hover the mouse over the top section to see a tooltip that lists the order of the subtracks on the plus strand. Similarly, hover the mouse over the bottom section to see the list of subtracks on the minus strand (right). Click on the wrench icon in the "RNA-Seq, CNS and adult head (modENCODE)" label to select the subtracks to show (e.g., only show subtracks that are on the minus strand). Subtracks on the plus strand have the prefix "+_" while subtracks on the minus strand have the prefix "-_"in the track configuration pane.**

Use the FlyBase Gene Report to examine the expression profiles of *Gyf*

To facilitate the identification of *D. melanogaster* genes that exhibit similar expression profiles, FlyBase has calculated the expression levels of each gene using the modENCODE RNA-Seq data for 30 developmental stages, 29 tissues, 25 treatments/conditions, and 24 cell lines (See the "RNA-Seq Profile" section of the "FlyBase:RNA-Seq Overview" page on the FlyBase wiki for details.) These gene expression profiles are available through the "High-Throughput Expression Data" subsection of the FlyBase Gene Report (under the "Expression Data" section).

Click on the *Gyf* feature in the *GBrowse* "Gene Span" track to navigate to the FlyBase Gene Report for *Gyf* (Figure 12).

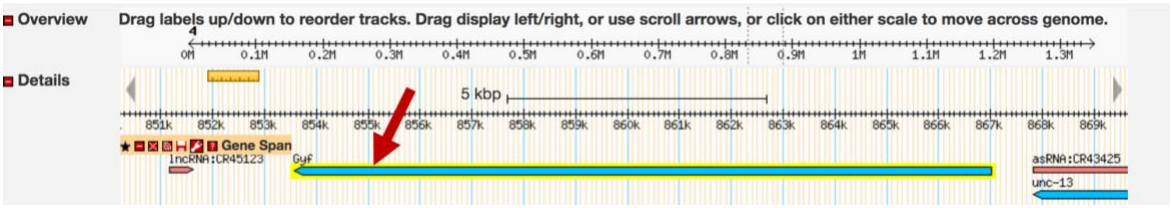

Figure 12 Click on the *Gyf* feature in the *GBrowse* "Gene Span" track to return to the FlyBase Gene Report for *Gyf*.

Click on the "Expression Data" link under the "Report Sections" panel on the right to navigate to the "Expression Data" section of the FlyBase Gene Report for *Gyf*. Click on the "High-Throughput Expression Data" header to expand the subsection (Figure 13). The gene expression levels derived from the modENCODE RNA-Seq datasets are listed under the "modENCODE [Anatomy, Development, Cell Lines, and Treatments] RNA-Seq" subsections.

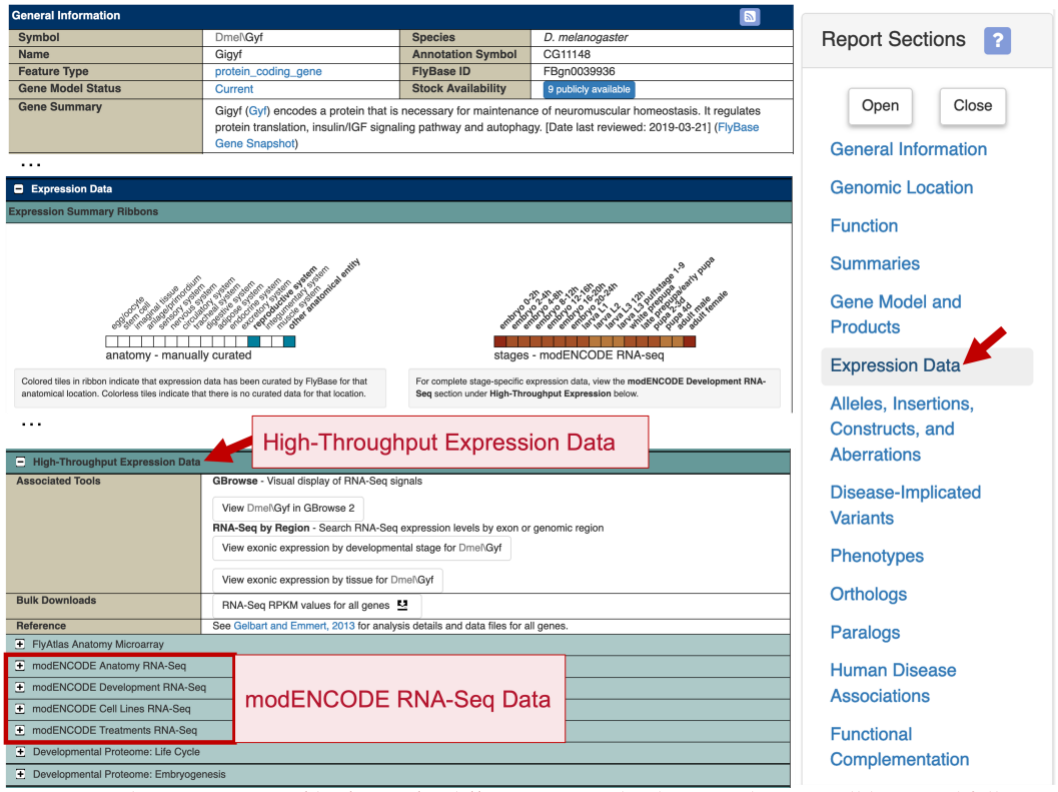

Figure 13 The expression profiles for *Gyf* in different tissues, developmental stages, cell lines, and following different treatments are available in the "High-Throughput Expression Data" subsection of the FlyBase Gene Report (red box).

To examine the expression levels of *Gyf* in different tissues, click on the "modENCODE Anatomy RNA-Seq" header to expand this subsection. The top pane of the expanded subsection contains the controls for manipulating the expression levels diagram shown in the bottom pane. To simplify the display and the query of gene expression levels, the gene expression levels are partitioned into eight expression level bins. The right side of the top pane includes a legend that shows the color and the range of expressions for each expression level bin (warmer colors denote higher levels of expression). Below is a brief overview of the available display options:

Styles:

- **linear**: The *x*-axis of the expression level histogram is in a linear scale (default)
- **log**: The *x*-axis of the expression level histogram is in $\log_2$ scale
- **heatmap**: Expression levels as a heat map (i.e. same width, different color)

Scales (i.e. the range of values on the *x*-axis):

- **gene maximum expression** (default):
  - Ranges from 0 to the maximum expression level of the gene among all of the samples in this RNA-Seq experiment
- **low expression bin max**:
  - Ranges from 0 to the maximum value of the "low" expression bin (10)
- **moderately high expression bin max**:
  - Ranges from 0 to the maximum value of the "moderately high" expression bin (50)
- **very high expression bin max**:
  - Ranges from 0 to the maximum value of the "very high" expression bin (1000)

Depending on the Styles setting, the bottom pane will show either a histogram or a heat map with the expression levels of *Gyf* in the 29 tissues sampled by modENCODE (Figure 14). The values in the bottom panel correspond to the expression level of the gene in each tissue. Values greater than the maximum value of the selected scale are in parentheses.

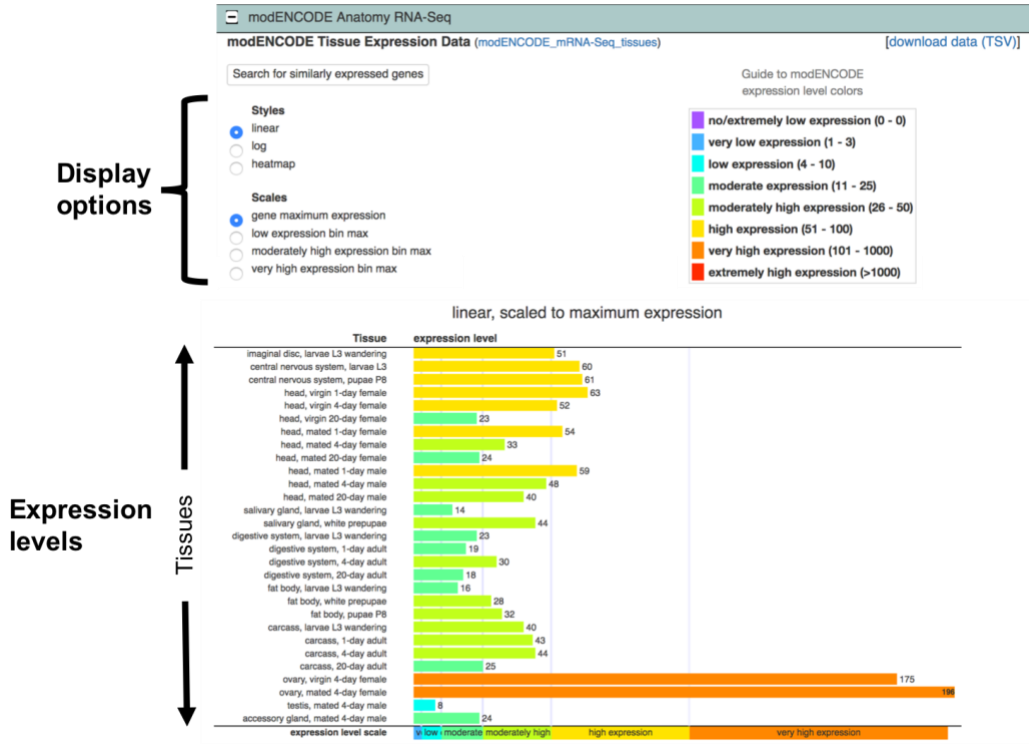

Figure 14 The "modENCODE Anatomy RNA-Seq" subsection shows the expression levels of *Gyf* in 29 tissues. By default, the expression levels are shown in a linear scale and the values on the x-axis range from 0 to the maximum expression level of the gene among the 29 tissues. The color of the bars in the "Expression levels" pane correspond to the expression level bin.

Interpreting the expression values of *Gyf*

The values in the bottom pane of the "modENCODE Anatomy RNA-Seq" subsection correspond to the **R**eads **P**er **K**ilobase of the exon models per **M**illion mapped reads (RPKM) in each tissue. The use of RPKM values facilitates the comparisons of gene expression levels **within** a sample because it normalizes the RNA-Seq read count for each gene by the size of the RNA-Seq library, the length of the gene (summed exons), and the length of the RNA-Seq reads for that sequence sample. FlyBase partitions the RPKM values into eight expression level bins (from "no/extremely low" expression to "extremely high" expression). Each expression level bin is assigned a different color as illustrated by the legend on the top right corner of the "modENCODE Anatomy RNA-Seq" pane. See the FlyBase High Throughput Expression Pattern Data reference report FBrf0221009 for details.

RPKM measures the expression level of a gene relative to the expression levels of all the other genes **within** a sample. The RPKM values of the same gene across multiple samples (e.g., among different tissues) or multiple experiments (e.g., among tissues, developmental stages, treatments) cannot be compared directly. This is because RNA-Seq measures the relative abundance of transcripts within a sample, instead of the absolute abundance. Hence the RPKM values are affected by both biological (e.g., range of expression for all the genes within a sample) and technical (e.g., efficacy in removing ribosomal RNAs from the sample prior to sequencing, fragment bias, depth of sequencing) variations within a sample. More sophisticated techniques are needed to estimate changes in expression levels from multiple RNA-Seq samples (see below).

The histogram in the "modENCODE Anatomy RNA-Seq" subsection shows that *Gyf* has "very high" expression in the ovaries of 4-day old mated adult females (RPKM = 196) compared to the expression of the other genes in the *D. melanogaster* genome (Figure 14). Similarly, *Gyf* has "very high" expression in the ovaries of 4-day old virgin adult females (RPKM = 175) compared to the other *D. melanogaster* genes. However, because RPKM is a within-sample normalization technique, these RPKM values do not provide any information as to the relative expression levels of *Gyf* in the ovaries of 4-day old mated adult females compared to the ovaries of 4-day old virgin females. Consistent with our previous observations using *GBrowse* (Figure 11), *Gyf* also shows "moderate" to "high" expression levels in the CNS and adult heads.

See the "RNA Quantitation from RNA-Seq Data" presentation on the GEP website for a more detailed explanation of the limitations of RPKM. See the "Differential gene expression analysis" section of the review by Conesa and colleagues for additional details on the different approaches and tools (e.g., *RSEM, DESeq2, edgeR*) for performing differential expression analysis of multiple samples and conditions (Conesa *et al.* 2016).

Use the RNA-Seq Expression Similarity Search tool to identify genes with similar expression profiles

Genes that exhibit similar expression profiles could be part of the same pathway or regulated by the same factors. We can use the FlyBase RNA-Seq Expression Similarity Search tool to identify genes that show expression profiles similar to those seen in a particular gene. (See the "RNA-Seq Similarity" section of the "FlyBase:RNA-Seq Overview" page on the FlyBase wiki for details.) To identify genes that exhibit tissue-specific expression profiles similar to those seen for *Gyf*, scroll up to the top of the FlyBase page, and select Tools ➜ Genomic Tools ➜ RNA-Seq Search ➜ RNA-Seq Similarity on the main navigation bar (Figure 15, top).

Enter "*Gyf*" into the "Sample gene" field, and then select "modENCODE_Tissues" under the "Experiment" field. To select all the tissues in the Categories field, click on the first entry ("imaginal disc, larvae L3 wandering"), hold down the **shift key**, scroll to the bottom of the list and select the last entry ("accessory gland, mated 4-day male"). Click on the "Submit Query" button to begin the search (Figure 15, bottom).

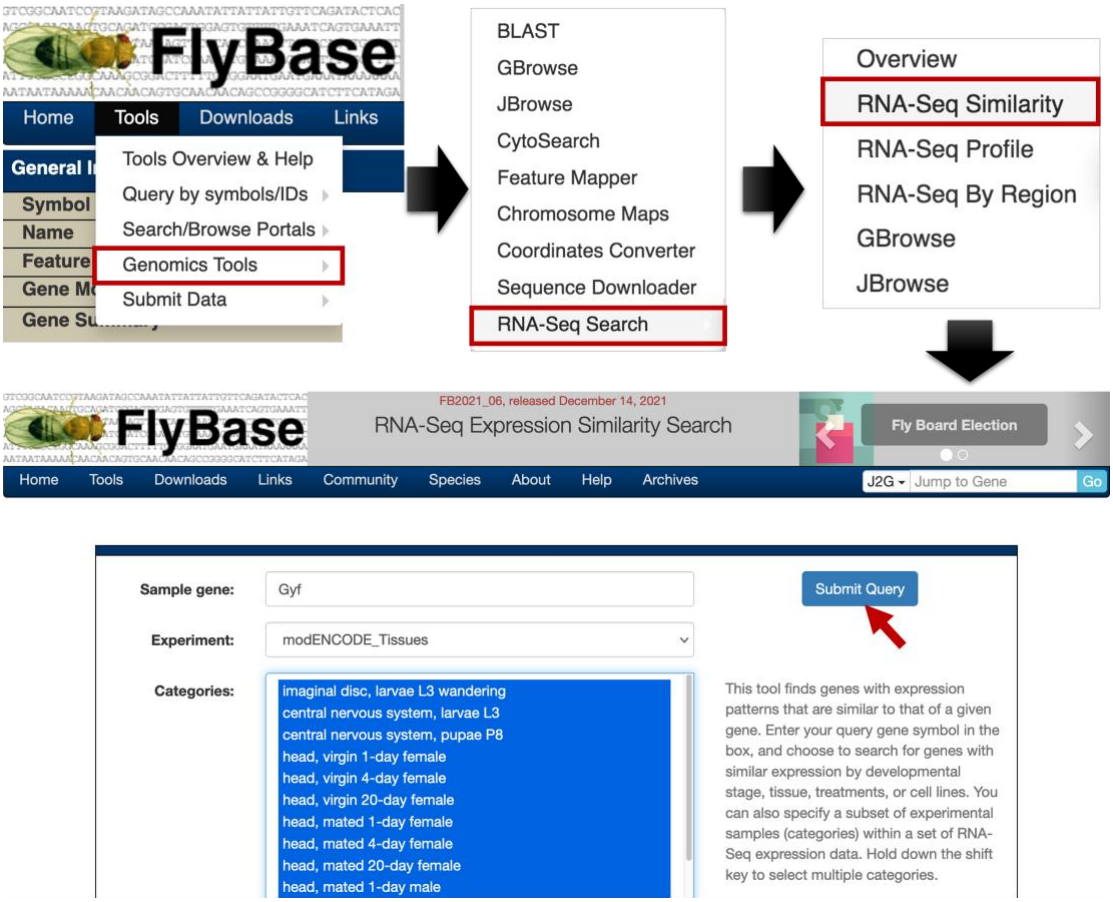

Figure 15 Access the "RNA-Seq Similarity" tool from the main navigation bar. Configure the tool to search for genes that exhibit expression profiles similar to *Gyf* based on all of the available modENCODE_Tissues RNA-Seq experiments.

The search results page lists the top 100 genes that have the highest Spearman's rank correlation coefficient with the tissue-specific expression profiles of *Gyf*. The results table is sorted by the Spearman's correlation (ρ) in descending order (Figure 16).

The "Gene" column in the results table contains the FlyBase gene symbols and links to the corresponding FlyBase Gene Report. The query gene of the Expression Similarity search (i.e. *Gyf*) is highlighted in green. The "Profile" column provides a graphical illustration of the gene expression profiles across the different tissues. The "Correlation" column lists the Spearman's correlation between the expression profiles of *Gyf* and the gene in each row. Genes with higher Spearman's correlations have tissue-specific expression profiles that are more similar to those seen in *Gyf*. The next two columns show the Gene Ontology (GO) terms in the "molecular function" and "biological process" domains that are associated with the gene in each row. The GO terms associated with the query gene (*Gyf*) are highlighted in yellow. The table shows that the *Gyf* gene has been associated with several molecular functions (e.g., protein binding, translation repressor activity) and biological processes (e.g., muscle cell cellular homeostasis).

Examination of the GO terms in the "Biological process" column indicates that many of the genes that show high Spearman's correlation with the tissue-specific expression profiles of *Gyf* are involved in neurogenesis and the development of the central nervous system. Among all of the *D. melanogaster* genes, the tissue-specific expression profiles of the *Fs(2)Ket* gene are the most similar to the expression profiles of *Gyf* (Spearman's correlation = 89.24%; Figure 16). The molecular function column shows that the *Fs(2)Ket* gene product exhibits small GTPase binding and nuclear localization sequence binding activities. The biological process column shows that *Fs(2)Ket* is involved in chorion-containing eggshell formation, the mitotic cell cycle, and protein import into the nucleus, among other reported activities.

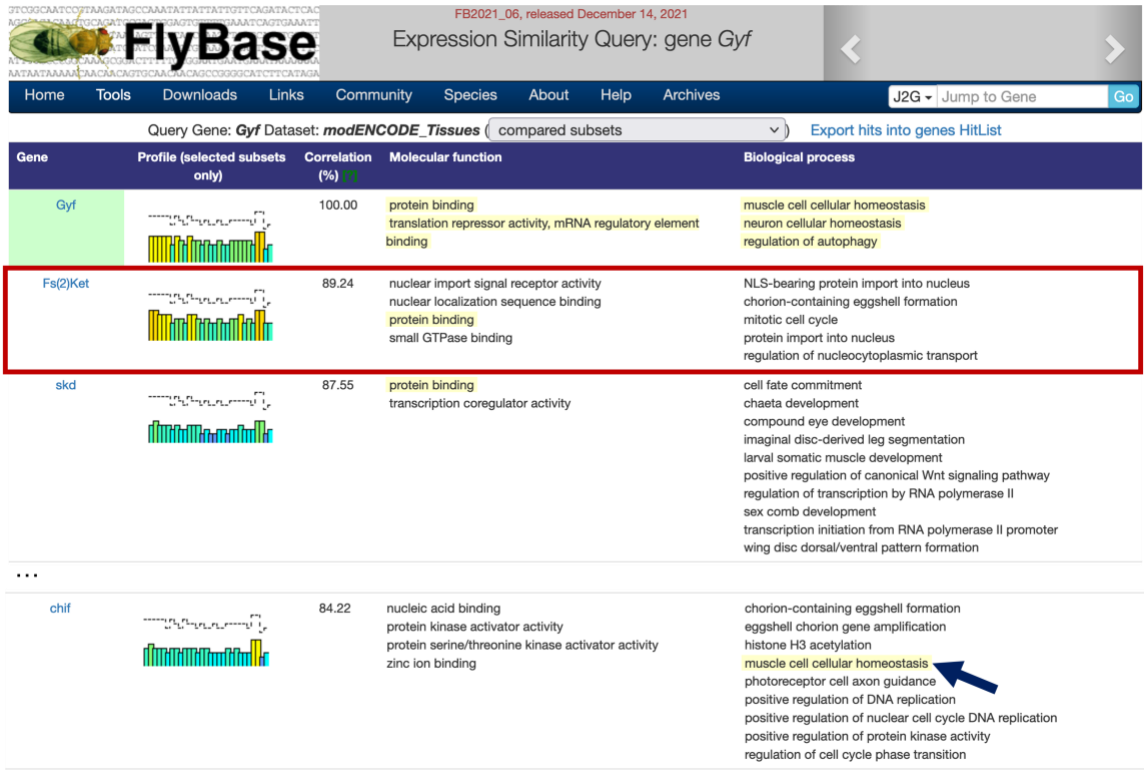

Figure 16 Search results from the Expression Similarity Search tool lists the top 100 *D. melanogaster* genes showing the highest Spearman's correlation with the tissue-specific expression profiles of *Gyf*. Among the approximately 14,000 *D. melanogaster* genes, the tissue-specific expression profiles of *Fs(2)Ket* is the most similar to the expression profiles of *Gyf* (Spearman's correlation = 89.24%; red box). Gene Ontology (GO) terms found in both the query gene and the gene in each row are highlighted in yellow (e.g., muscle cell cellular homeostasis for the gene *chif*, which has an 84.22% correlation; blue arrow).

To learn more about this gene, click on the "Fs(2)Ket" link under the "Gene" column to navigate to the corresponding FlyBase Gene Report (Figure 17, top). The "General Information" section of the report shows that the name of this gene is *Female sterile (2) Ketel*, and the "Genomic Location" section indicates that this gene is located on the left arm of chromosome 2 (2L).

Scroll down to the "modENCODE Anatomy RNA-Seq" section (under "Expression Data" ➜ "High-Throughput Expression Data"). Consistent with the high Spearman's correlation, the overall tissue-specific expression profile of *Fs(2)Ket* is similar to that seen for *Gyf*. However, *Fs(2)Ket* shows "very high" expression levels in the imaginal discs and in the CNS of 3rd instar larvae (Figure 17, bottom), compared to only "high" expression levels in *Gyf* (Figure 14).

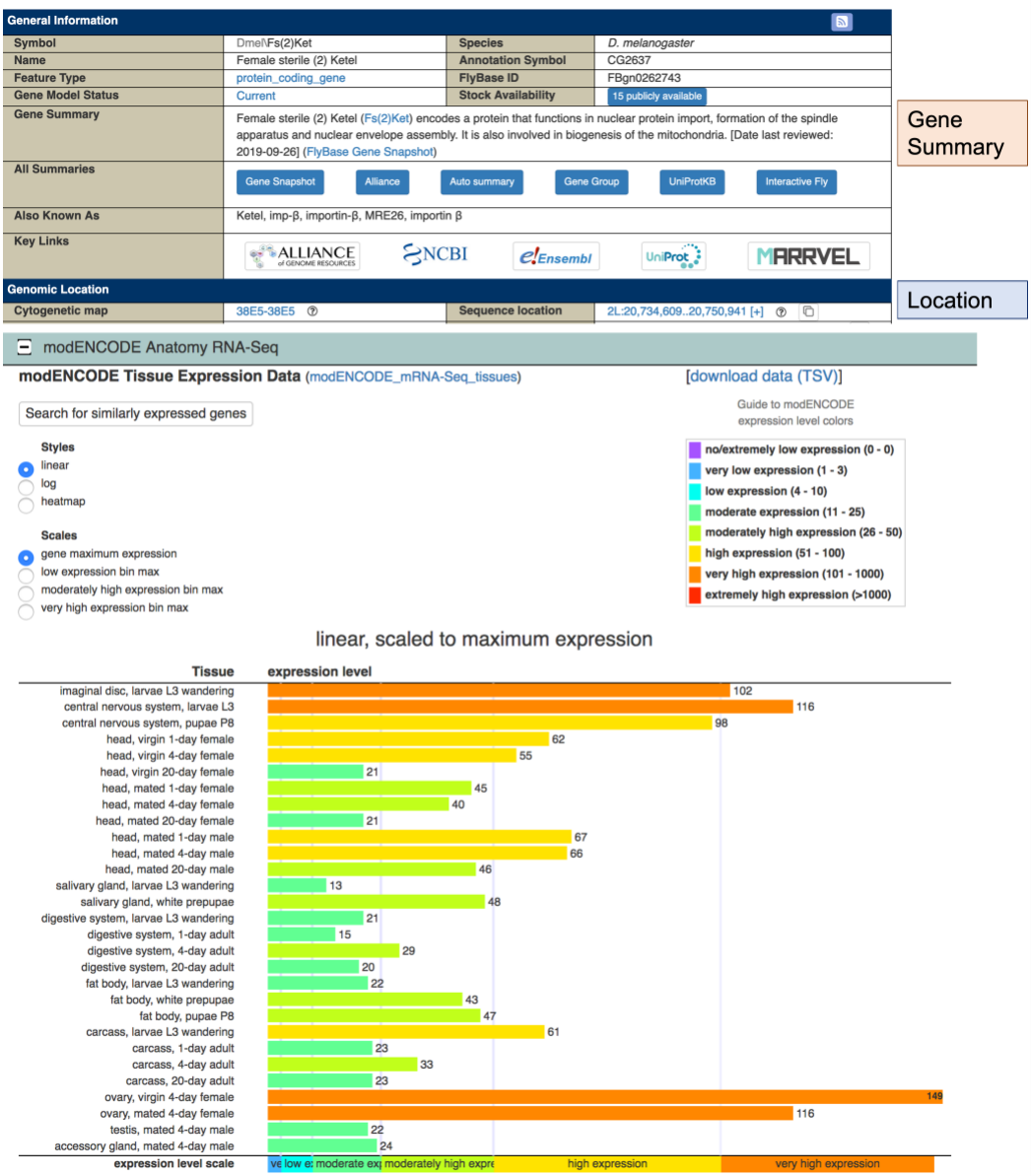

Figure 17 The FlyBase Gene Report for *Fs(2)Ket* shows that the gene *Female sterile (2) Ketel* is located on chr2L. The Gene Summary section describes the known functions of this gene (i.e. functions in nuclear protein import, formation of the spindle apparatus, assembly of the nuclear envelope). (Bottom) The "modENCODE Anatomy RNA-Seq" subsection shows the *Fs(2)Ket* tissue-specific expression profile, which looks similar to that for *Gyf* (Figure 14) in most respects.

The *Fs(2)Ket* gene shows "moderate" to "high" expression levels in adult heads, and "very high" expression levels in the CNS of 3rd instar larvae, which suggests that this gene might be involved in the nervous system. Examination of the GO terms in the "Biological Process" subsection [under the "Gene Ontology (GO)" section] shows that *Fs(2)Ket* does not have known associations with the nervous system. Hence, we will expand the scope of our investigation to the orthologs of *Fs(2)Ket* in other organisms.

Use Gene2Function to infer the functions of *Fs(2)Ket* based on its human ortholog

To identify the putative orthologs of *Fs(2)Ket*, open a new web browser tab and navigate to the FlyBase homepage. Click on the "Inter-species" header at the bottom left column of the page (under "External Resources"), and then click on the "G2F" icon (Figure 18).

Figure 18 Under the "External Resources" panel at the bottom left corner of the FlyBase home page, click on the "Inter-species" header and then click on the "G2F" icon to access the Gene2Function (G2F) website.

The Gene2Function (G2F) website provides a unified interface to the information stored in curated model organism databases, such as the databases for human (HGSC), rat (RGD), mouse (MGI), frog (Xenbase), zebrafish (ZFIN), *Drosophila* (FlyBase), nematodes (WormBase), and yeast (SGD). This interface enables users to quickly identify the putative orthologs of a gene in different model organisms, and to ascertain if the putative human ortholog is associated with human diseases.

To identify the putative orthologs of *Fs(2)Ket*, select "Fly" under the "Species" field of the "Search by Gene" section, enter "*Fs(2)Ket*" into the "Gene" field, verify that the "Return only best match when there is more than one match per input gene or protein" option is selected under the "Filters" field, and then click on the "Search By Gene" button (Figure 19).

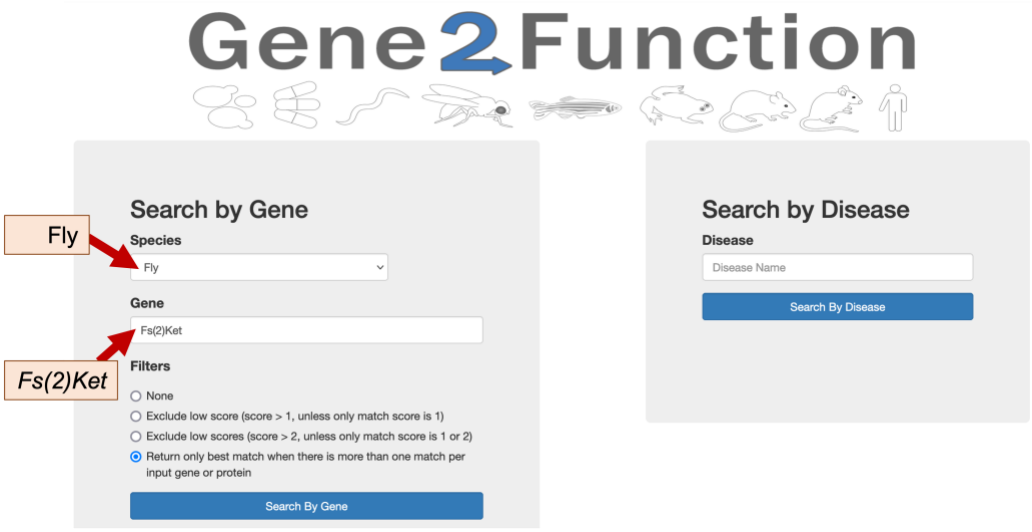

Figure 19 The Gene2Function search page can retrieve records based on either the name of the gene in one of the supported organisms (left), or the name of a human disease (right). For example, we can select "Fly" and enter the gene name "*Fs(2)Ket*" to retrieve the putative orthologs of this gene in other model organisms (red arrows).

The table in the "Orthologs Overview" section of the "Gene Search Results" page shows the putative orthologs of *Fs(2)Ket* in eight other model organisms (Figure 20). The "NCBI Gene ID" column contains links to the gene records in the NCBI Entrez database, while the "Species specific gene ID" column contains links to the records in the primary source database (e.g., FlyBase). The table also includes links to additional metadata associated with each gene, such as GO terms, publications, protein interactions, and protein alignments.

Figure 20 The "Orthologs Overview" section of the G2F "Gene Search Results" page provides a summary of the putative orthologs of the query gene in the other model organism databases. The metadata (e.g., publications, GO terms) associated with each ortholog is available on the right side of the table. The query gene used in the G2F search is highlighted in blue (e.g., *Fs(2)Ket* in *D. melanogaster*). The human ortholog of *Fs(2)Ket* is *KPNB1* (red box), and there are three human disease terms associated with this ortholog (red arrow).

See the G2F help page and Hu *et al.* 2017 for additional information on the key functionalities and goals of G2F. See the Hu *et al.* 2011 manuscript for details on the DRSC Integrative Ortholog Prediction Tool (DIOPT) used to identify the putative orthologs, and on the confidence ranks associated with the ortholog assignments.

The table shows that *KPNB1* is the putative ortholog of *Fs(2)Ket* in human, and there are three human disease terms associated with this human ortholog. Click on the "3" link under the "Human Disease Count" column to identify the human disease terms associated with *KPNB1*, (Figure 20, red arrow). The "Disease Annotation" page shows that *KPNB1* is associated with three disease terms: multiple sclerosis, optic cup area, and very long-chain saturated fatty acid levels (fatty acid 20:0). The "Source" column indicates that these disease associations were determined by Genome-Wide Association Studies (GWAS). Links to the original publications that support these associations are listed under the "Other Values" column (Figure 21).

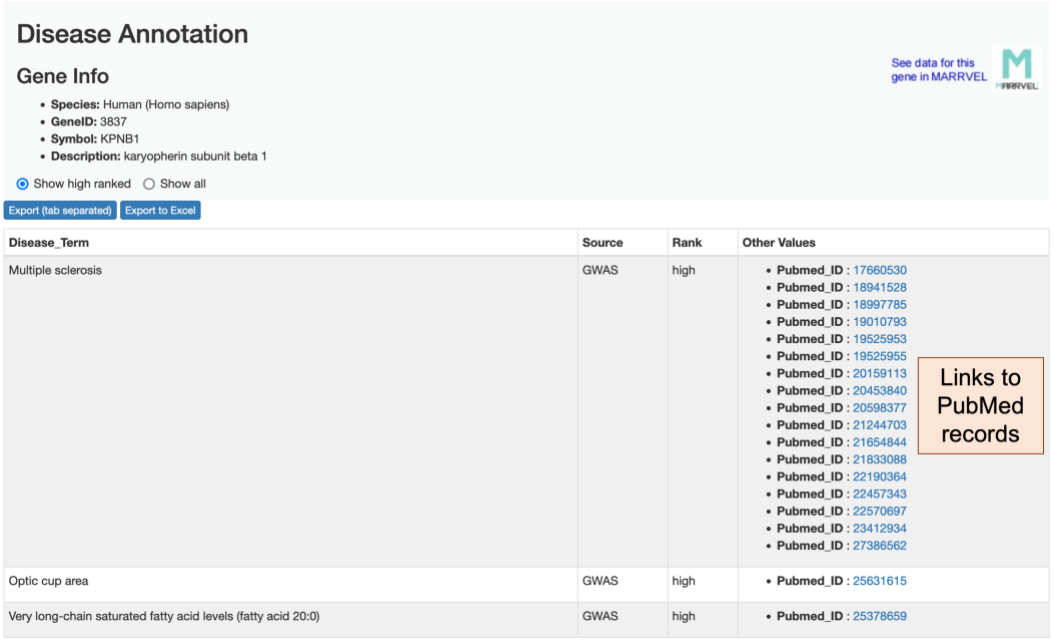

Figure 21 The "Disease Annotation" page shows the three disease terms that are associated with the human gene *KPNB1*. The "Pubmed_ID" links under the "Other Values" column correspond to the publications that support these disease associations.

The Disease Annotation page shows 17 publications that support the association between *KPNB1* and the CNS disorder multiple sclerosis (MS). Examination of the PubMed record associated with the "optic cup area" disease term (Pubmed_ID: 25631615) shows that this association is derived from a GWAS study on primary open-angle glaucoma. This study identifies genes that affect the morphology of the optic nerve head (also known as the optic disc). *KPNB1* is found to be associated with the cup area of the optic nerve head (Springelkamp *et al.* 2015).

The FlyBase record for the *D. melanogaster* gene *Fs(2)Ket* does not show any known functions in the CNS. However, the tissue-specific expression profiles of *Fs(2)Ket* and the disease associations of its human ortholog *KPNB1* suggest *Fs(2)Ket* might play a role in the CNS and in neurogenesis.

The GWAS data shown in the "Disease Annotation" page is derived from the information in the GWAS Catalog. See the Documentation page on the GWAS Catalog website for additional details on the methods used to curate the GWAS datasets, and the ontology used to describe disease traits.

Interpreting the correlations reported by the Expression Similarity Search tool

Go back to the web browser tab with the RNA-Seq Expression Similarity search results for *Gyf*. The gene that shows the second highest correlation with the tissue-specific expression profiles of *Gyf* is *skd* (Spearman's correlation = 87.55%). However, the illustration in the "Profile" column shows that *skd* has substantially lower levels of expression than *Gyf* (as denoted by color of the bars that reflect the expression level bins, and the height of the bars in the histogram).

This discrepancy can be explained by the criteria used by the FlyBase RNA-Seq Expression Similarity Search tool to calculate the Spearman's correlation of gene expression profiles. The correlation is based on the "shape" (i.e. peaks and troughs) of the expression profiles among all of the samples rather than the actual expression levels (i.e. RPKM values). Consequently, genes with high Spearman's correlation could nonetheless show large differences in expression levels.

To illustrate this concept, click on the "*skd*" link under the "Gene" column to open the FlyBase Gene Report in a new tab. Scroll down to the "modENCODE Anatomy RNA-Seq" section (under "Expression Data" ➜ "High-Throughput Expression Data"). The panel shows that the expression levels of *skd* range from "very low" to "moderately high" (Figure 22, right). By contrast, the expression levels of *Gyf* range from "low" to "very high" (Figure 22, left). While *skd* shows lower expression levels than *Gyf*, the two expression profiles have a similar "shape", which explains the high Spearman's correlation between these two genes.

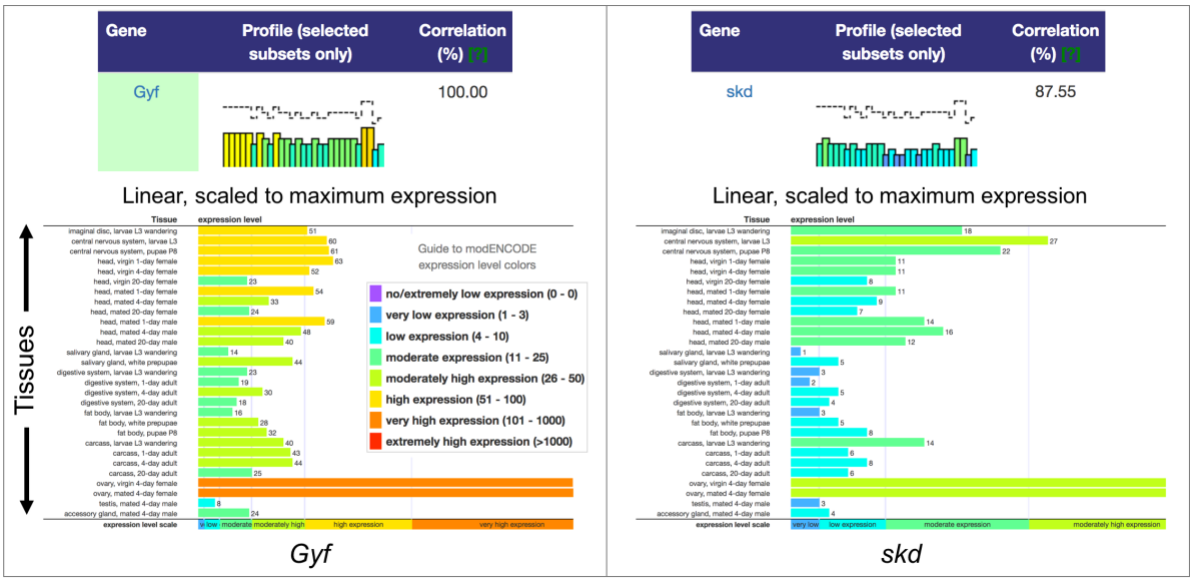

Figure 22 Correlations between expression profiles are based on the "shape" (i.e. peaks and troughs across the different tissues) of the expression profile, not the expression (RPKM) values. Hence the tissue-specific expression profiles of *Gyf* and *skd* have a Spearman's correlation of 87.55%, despite the lower overall expression levels of *skd* compared to *Gyf*.

Use the FlyBase Genes HitList to analyze multiple genes with similar expression profiles

In addition to investigating each gene on the Expression Similarity search results page separately, we can also export the entire list of genes to the FlyBase Genes HitList for analysis. Click on the "Export hits into genes HitList" link above the results table to transfer *Gyf* and the 100 genes that show the highest Spearman's correlations with the tissue-specific expression profiles of *Gyf* to the Genes HitList (Figure 23).

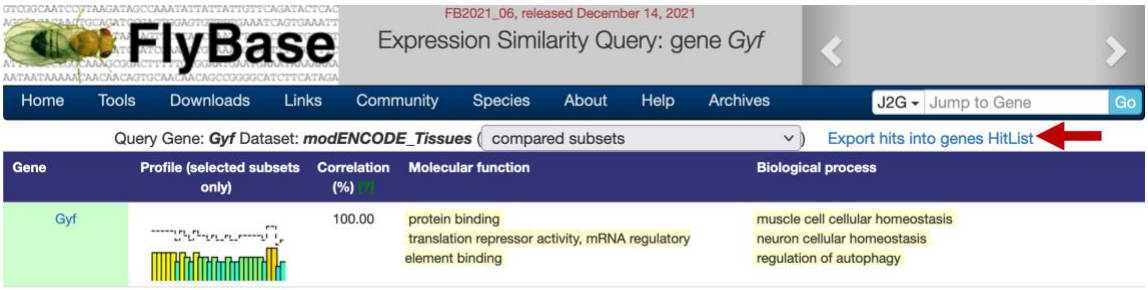

Figure 23 Click on the "Export hits into genes HitList" link to transfer the list of genes from the Expression Similarity search result page to the FlyBase Genes HitList.

The Genes HitList provides a summary of the 101 genes in the Expression Similarity search results. The "Convert", "Export", and "Analyze" buttons above the HitList allow us to query and manipulate the HitList.

Click on the "Table" button under the "View As" section for a more compact view of the HitList. Each row in the Genes HitList table contains the gene symbol, gene name, annotation ID, and the cytological location of the gene. It also lists the genetic resources (e.g., alleles, stocks) that might be available (Figure 24).

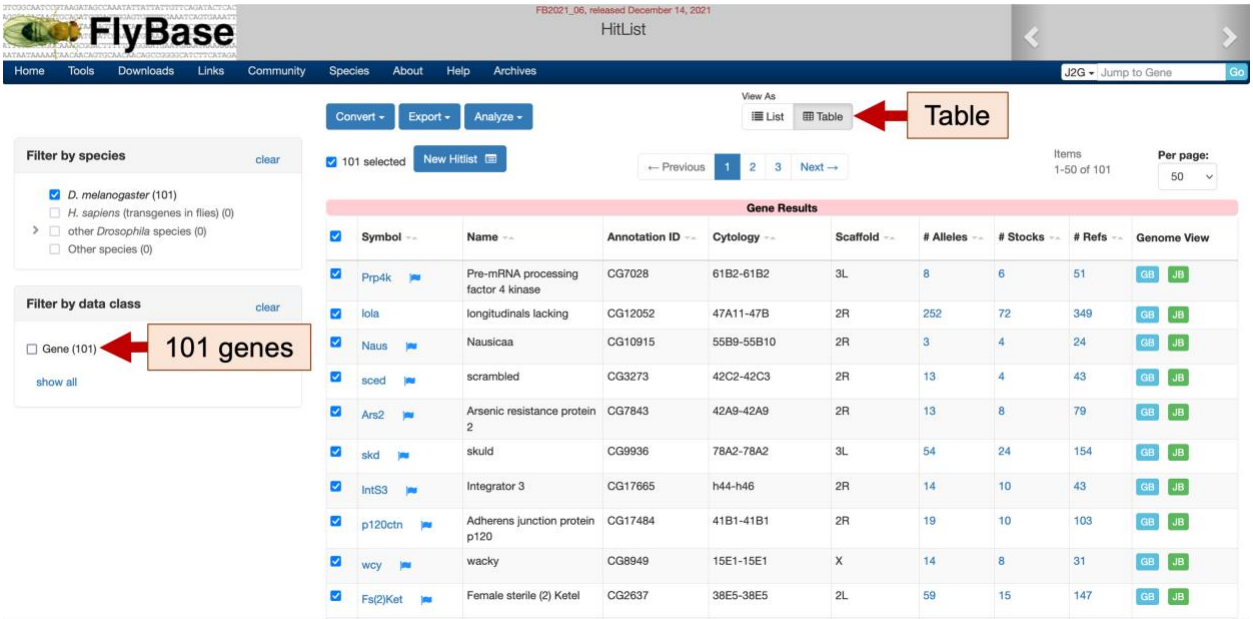

Figure 24 The 101 genes (i.e. *Gyf* and the top 100 genes with tissue-specific expression profiles most similar to *Gyf*) are shown in the FlyBase Genes HitList. Click on the "Table" button to view the HitList in tabular format (red arrow). Click on the links under the "Symbol" column to access the corresponding FlyBase Gene Report. The blue flag next to the gene symbol indicates that there was a recent update to the gene record.

To analyze this Genes HitList, click on the "Analyze" button above the HitList table. A drop-down menu will appear which allow us to analyze the genes in the HitList based on the frequency of GO terms, expression profiles, conserved domains, chromosome arms, as well as protein and genetic interactions. Click on the "Chromosome arm" link to determine the distribution of these genes among the different *D. melanogaster* chromosomes (Figure 25).

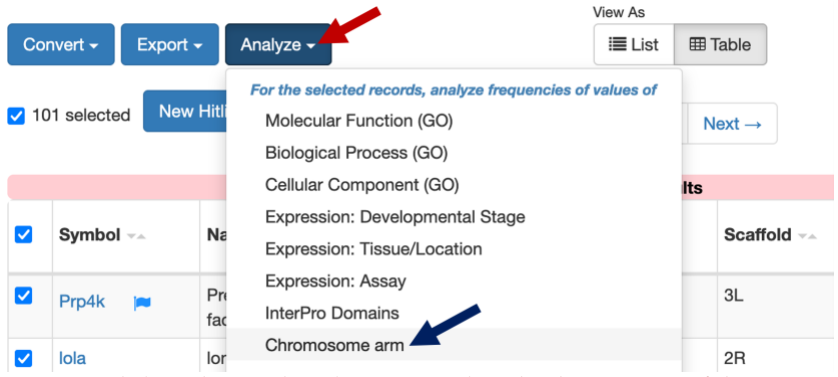

Figure 25 Click on the "Analyze" button to explore the characteristics of the genes in the HitList table (red arrow). Click on the "Chromosome arm" link to see the number of genes on the HitList that are located on each chromosome (blue arrow).

The "Values Frequency" table shows that 35 of the genes on the HitList are located on the X chromosome, while four of the genes are located on the 4th chromosome (Figure 26, top). Click on the "4" link under the "Related records" column to filter the HitList so that it only contains the four genes located on the 4th chromosome. The "Refinement results" page shows that the 4th chromosome genes *Crk*, *Tdg*, and *Slip1* exhibit tissue-specific expression profiles that are the most similar to the expression profiles of *Gyf* (Figure 26, bottom).

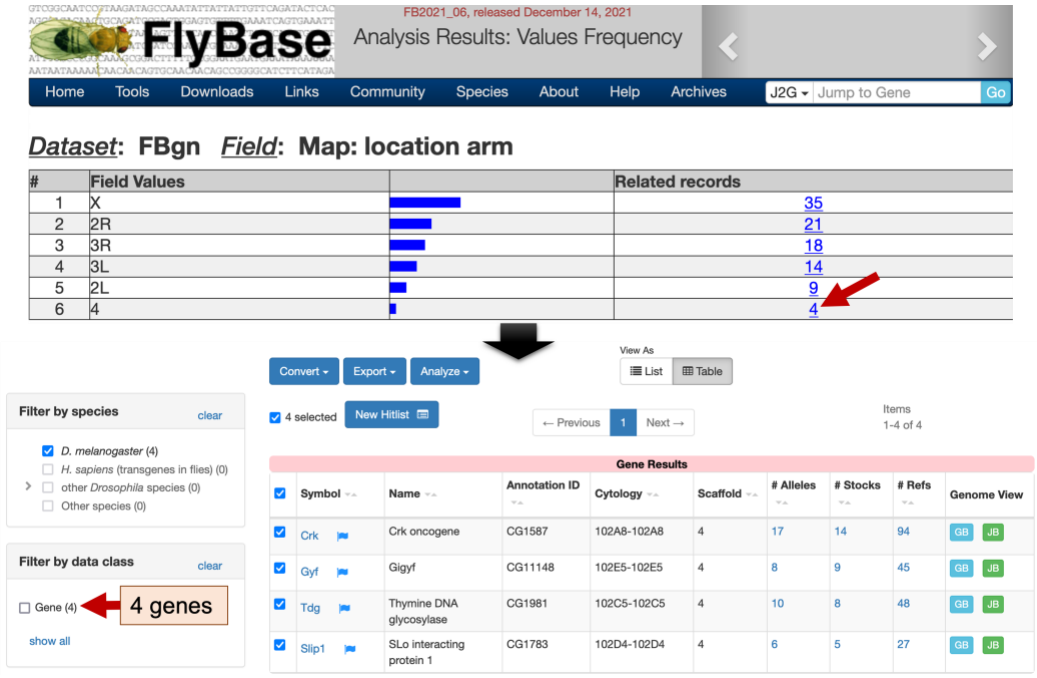

Figure 26 The chromosome arm "Value Frequency" table shows the number of genes in the HitList that are located on each chromosome. Click on the corresponding link in the "Related records" column to select genes with a particular field value (e.g., the "location arm" field has a value of 4; red arrow). (Bottom) After applying the filter, the refined HitList contains only the four genes that are located on the 4th chromosome.

Search for similar expression profiles using a subset of the RNA-Seq samples

In addition to searching for similarly expressed genes using all the available RNA-Seq samples (e.g., the 29 tissues sequenced by modENCODE), we can also use the FlyBase RNA-Seq Expression Similarity Search tool to compare the expression profiles of a subset of the RNA-Seq samples. For example, we can use this interface to identify genes that show expression profiles similar to *Gyf* in the imaginal disc, central nervous system, ovary, testis, and accessory glands.

To perform this search, open a new web browser window and navigate to the FlyBase RNA-Seq Expression Similarity Search tool (Figure 15). Enter "*Gyf*" into the "Sample gene" field, select "modENCODE_Tissues" under the "Experiment" field, and select the following samples under the "Categories" field (Figure 27):

- imaginal disc, larvae L3 wandering
- central nervous system, larvae L3
- central nervous system, pupae P8
- ovary, virgin 4-day female
- ovary, mated 4-day female
- testis, mated 4-day male
- accessory gland, mated 4-day male

You can use the control key on MS Windows (the command key on macOS) to select multiple samples on the list individually. Click on the "Submit Query" button to perform the search.

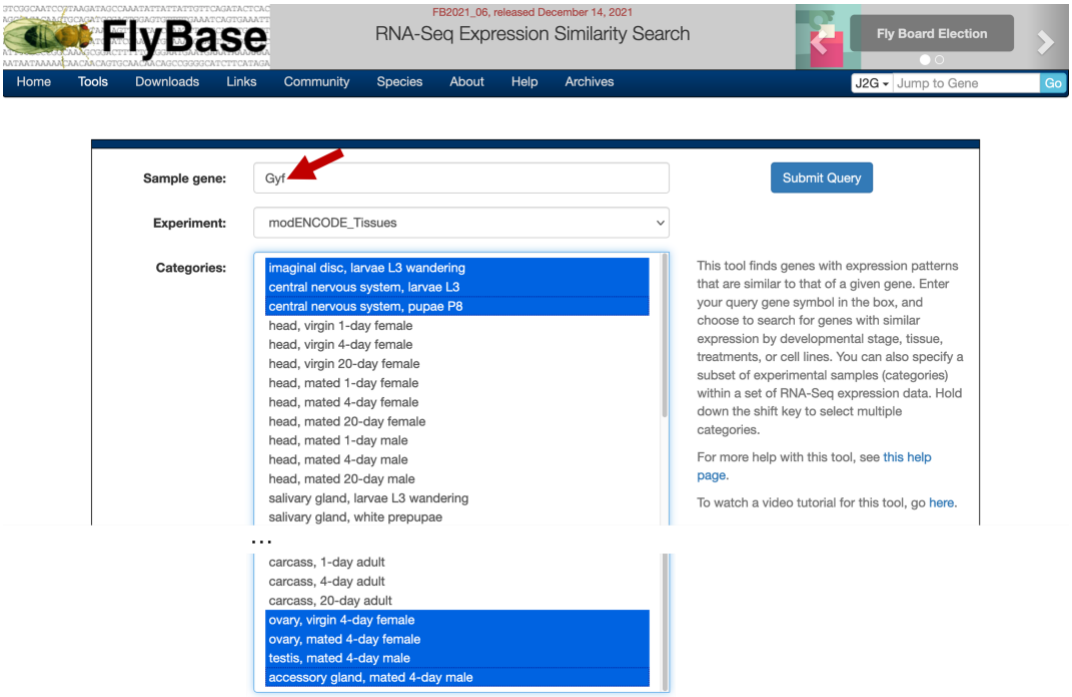

Figure 27 Configure the RNA-Seq Expression Similarity Search interface to identify genes that show strong Spearman's correlation with *Gyf* in seven tissues.

The results page shows two genes (*bon* and *Sin3A*) that have 100% correlation with the expression profiles of *Gyf* in these seven tissue samples. The GO terms in the "Molecular function" column indicate that both *bon* and *Sin3A* have chromatin binding activity, suggesting a role in controlling gene expression. The GO terms in the "Biological process" column shows that the *bon* gene is involved in protein ubiquitination, while *Sin3A* is involved in the determination of adult lifespan (Figure 28, red arrows). Hence both genes might also be related to the role of *Gyf* in the regulation of autophagy. (A single gene product can have several molecular functions, and be involved in multiple biological processes.)

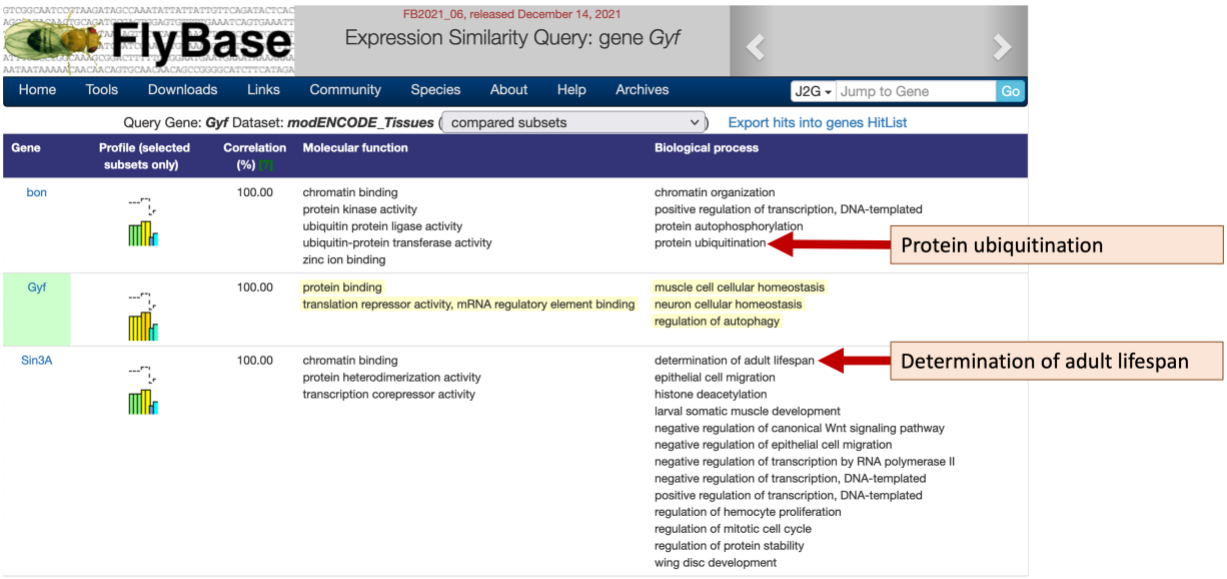

Figure 28 The Expression Similarity search results based on the expression profiles of *Gyf* in seven tissue categories (samples). The "Biological process" column shows that both *bon* and *Sin3A* might be involved in autophagy (red arrows).

Use the RNA-Seq Expression Profile Search tool to identify genes with a particular expression profile

Our investigation thus far has focused on the expression profiles of a particular gene (*Gyf*), followed by the use of the Expression Similarity Search tool to identify other genes with similar expression profiles. However, we can also search for genes that match a particular expression profile across multiple tissues, developmental stages, treatments, and cell lines using the FlyBase RNA-Seq Expression Profile Search tool.

Click on the "*Gyf*" link in the Expression Similarity search results page to return to the FlyBase Gene Report for *Gyf*. Scroll down to the "Expression Data" section and then expand the four modENCODE RNA-Seq subsections under "High-Throughput Expression Data" (Figure 29):

- modENCODE Anatomy RNA-Seq
- modENCODE Development RNA-Seq
- modENCODE Cell Lines RNA-Seq
- modENCODE Treatments RNA-Seq

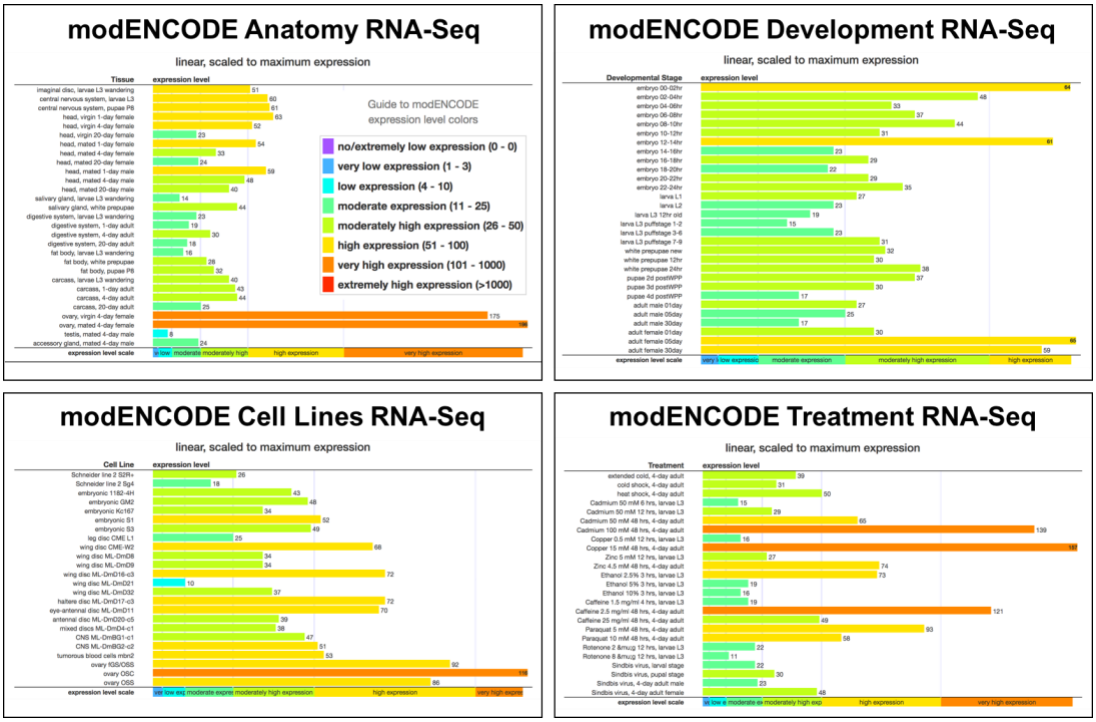

Figure 29 The expression profiles for *Gyf* in different tissues (anatomy), developmental stages, cell lines, and treatments.

The "modENCODE Treatment RNA-Seq" subsection shows that the expression levels of *Gyf* are affected by different treatments. In order to assess the changes in gene expressions in response to environmental stimuli, *D. melanogaster* was fed different chemicals (e.g., ethanol, caffeine, paraquat), exposed to heavy metals (e.g., cadmium, copper, zinc) or the Sindbis virus, or subjected to heat or cold shock prior to RNA sequencing. (See the FlyBase record FBlc0000236 for details.) The "modENCODE Treatment RNA-Seq" subsection shows that the *Gyf* gene has "very high" expression when 2-day old adults are exposed to high concentrations of cadmium (Cadmium 100 mM 48 hrs) or copper (Copper 15 mM 48 hrs), or when they were fed caffeine (Caffeine 2.5 mg/ml 48 hrs) for 48 hours.

We can use the FlyBase RNA-Seq Expression Profile Search tool to identify other genes that exhibit tissue-specific and treatment-specific expression profiles that are similar to those seen for *Gyf*, suggesting similar regulatory responses. This search tool is available through the main navigation bar on FlyBase (under Tools ➜ Genomic Tools ➜ RNA-Seq Search ➜ RNA-Seq Profile; Figure 30).

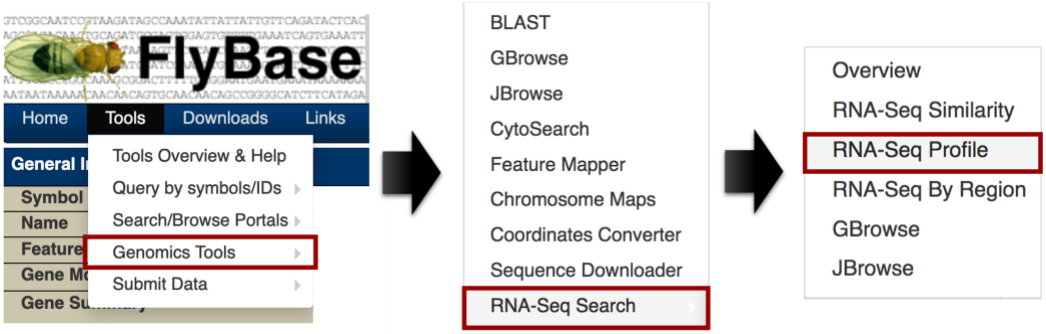

Figure 30 Access the "RNA-Seq Profile" search tool from the main navigation bar on FlyBase.

The RNA-Seq Expression Profile Search tool allows us to either search each modENCODE RNA-Seq experiment separately (i.e. by stage, tissue, treatment, or cell line), or search multiple RNA-Seq experiments in a combined search. The checkboxes in the "search using several modENCODE expression datasets in conjunction" section allow us to specify the experiments to use, and the sections below contain the list of available samples within each experiment.

We will examine the tissue and treatment RNA-Seq datasets to search for genes that show expression profiles that are similar to those in *Gyf* (Figure 29). Unselect the "stage" checkbox, and then select the "tissue" and "treatment" checkboxes under the "search using several modENCODE expression datasets in conjunction" section (Figure 31). The panels below will change so that it shows the list of available tissues under the "modENCODE expression by tissue data" section, and the list of available treatments under the "modENCODE expression by treatment data" section.

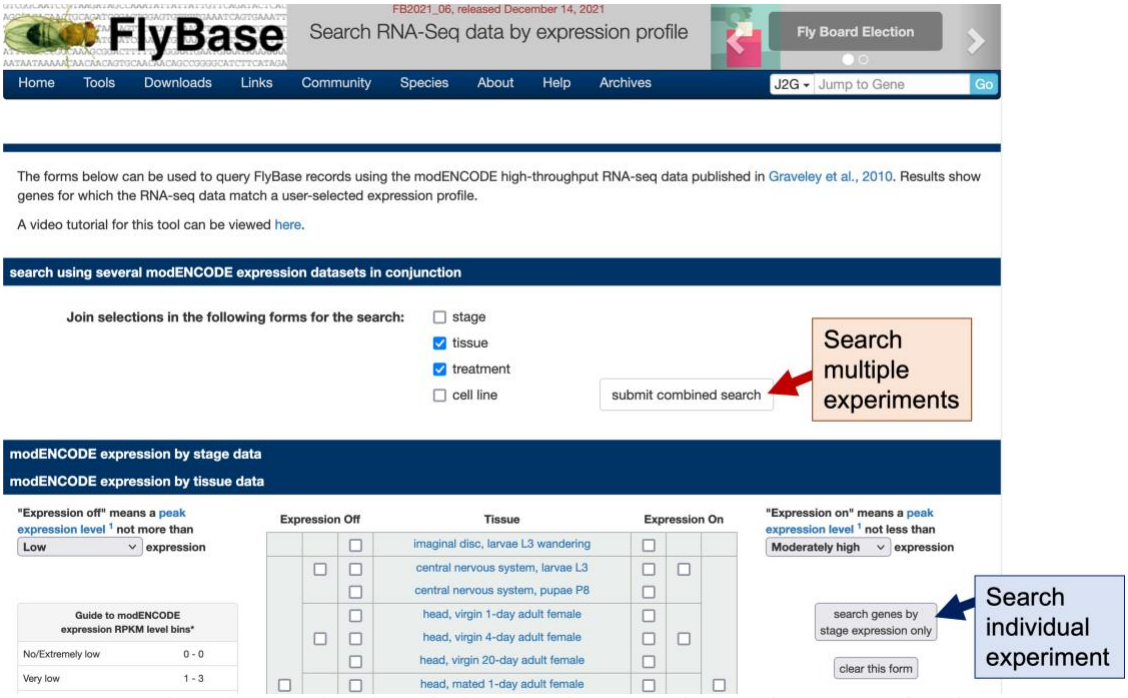

Figure 31 Use the "submit combined search" button (red arrow) in the "search using several modENCODE expression datasets in conjunction" section to search for similar expression profiles based on multiple RNA-Seq experiments. Use the "search genes by stage [sic] expression only" button (blue arrow) to search for similar expression profiles within an experiment.

Each experiment section contains three controls (Figure 32). The drop-down list on the left defines the criteria for classifying a gene as not being expressed (i.e. "Expression off"). The table in the middle is used to specify the expression profiles of interest. The drop-down list on the right defines the criteria for classifying a gene as being expressed (i.e. "Expression on").

The definition of whether a gene is "on" or "off" is based on the expression level bins. As explained above, FlyBase partitions the RPKM expression values into 8 bins (from "no/extremely low" expression to "extremely high" expression). By default, genes that show "low" expression levels or below (i.e. RPKM values <= 10) are considered to be "off", while genes that show "moderately high" expression levels or above (i.e. RPKM values >= 26) are considered to be "on".

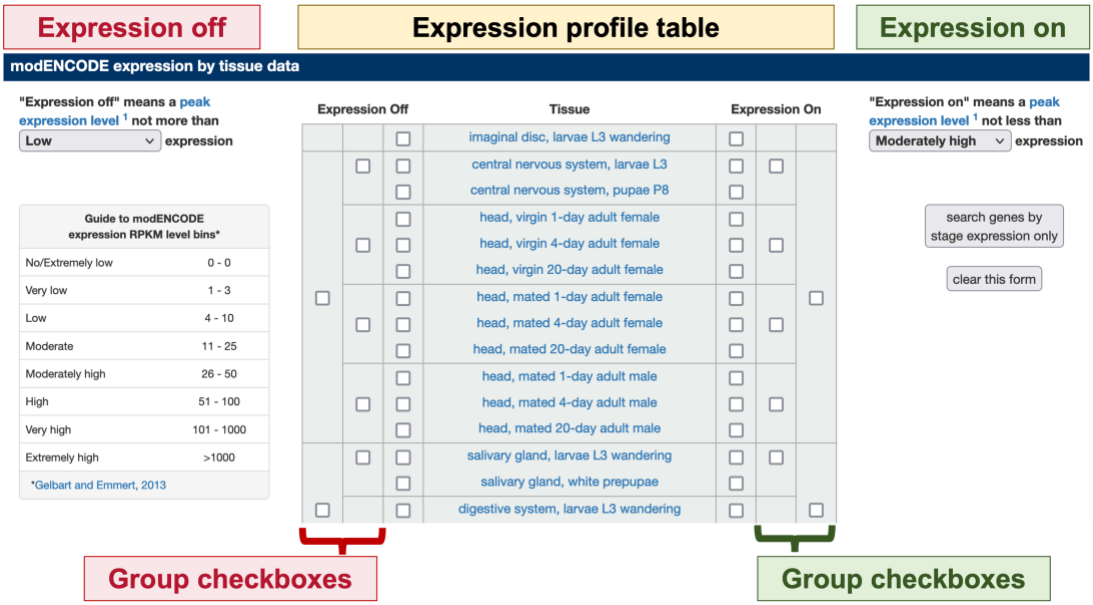

Figure 32 There are three controls within each experiment section that can be used to define the expression profiles of interest. The drop-down lists specify the expression level bins used to define whether a gene is "on" or "off". The table in the middle allows us to specify the expression profile of each sample within the experiment based on these expression criteria. The outer group checkboxes can be used to specify the expression profiles of multiple samples.

There are three sets of checkboxes under the "Expression off" and the "Expression on" columns in the "Expression profile table." The outer group checkboxes allow us to select multiple samples at once, while the innermost checkboxes allow us to specify the expression for each sample. Samples not covered by the "Expression off" and "Expression on" selections are omitted from the expression profile search criteria (i.e. the gene can either be on or off in that sample).

The group checkboxes in the "Expression off" column act as a logical "**AND**" in the Expression Profile search while the group checkboxes in the "Expression on" column act as a logical "**OR**". Selecting the group checkboxes in the "Expression off" column means that **all** of the selected samples have RPKM values that are at or below the "Expression off" level bin. By contrast, selecting the group checkboxes in the "Expression on" column means that **at least one of the samples** within the group has an RPKM value at or above the "Expression on" level bin. Consequently, in order to select genes that are "on" in all of the samples within a group, one would need to select the "Expression on" checkboxes for each sample within a group individually. (See the "RNA-Seq Part II: Using RNA-Seq Profile Search" video on the FlyBase YouTube channel for details.)

In this example, we will search for genes that exhibit the same expression profiles as *Gyf* in a subset of tissues and treatments, setting the values found for that gene. Under the "modENCODE expression by tissue data" section, select the "Moderate" option from the "Expression off" drop-down list, and verify that the "Moderately high" option is selected in the "Expression on" drop-down list. Configure the expression profiles table to search for genes that are "off" in testis and accessory glands of 4-day old adult males, and are "on" in the central nervous system, head tissues of 1-day and 4-day old adults, and the ovaries of 4-day old adult females (Figure 33).

Figure 33 Configure the Expression Profile search criteria for the modENCODE tissue RNA-Seq samples. Select the innermost checkboxes for the central nervous system larvae L3 and pupae P8 samples under the "Expression on" section in order to identify genes that are expressed in both samples.

Scroll down to the "modENCODE expression by treatment data" section. Change the "Expression off" criteria to "Moderate" and verify that the "Expression on" criterion is set to "Moderately high". Configure the expression profiles table to search for genes that are "off" under Rotenone (an insecticide) treatment in 3rd instar larvae, and are "on" under exposure to heavy metals, caffeine, and paraquat (an herbicide) in 4-day old adults (Figure 34).

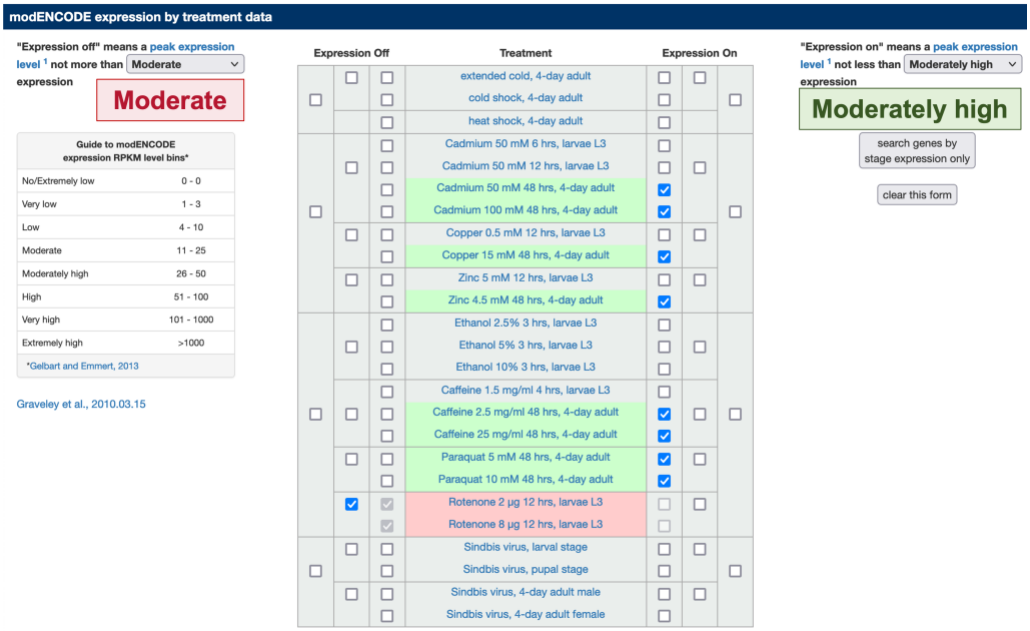

Figure 34 Configure the expression profiles search criteria for the modENCODE treatment RNA-Seq samples.

Scroll up to the top of the page and then click on the "submit combined search" button under the "search using several modENCODE expression datasets in conjunction" section (Figure 35, top). The Genes HitList shows 32 genes that satisfy these expression profiles search criteria (Figure 35, bottom). We can use the "Analyze" and "Convert" buttons above the table to analyze this Genes HitList (see the "Use the FlyBase Genes HitList to analyze multiple genes with similar expression profiles" section on page 18 for details).

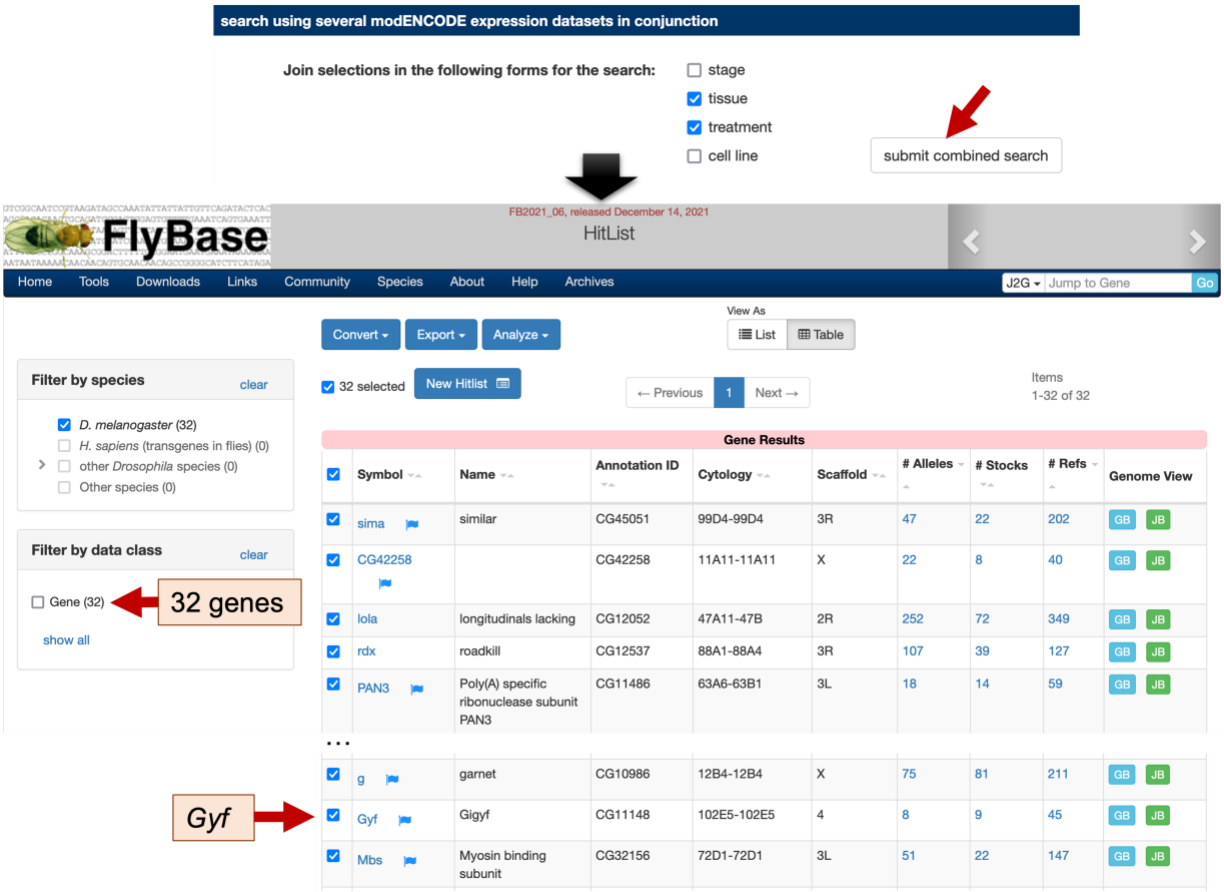

Figure 35 Click on the "submit combined search" button to search for genes that exhibit the tissue-specific and treatment-specific expression profiles specified in the experiment sections (see Figure 33 and Figure 34). (Bottom) This search identifies 32 genes that match the specified expression profiles, including *Gyf* (red arrow).

Genes that exhibit similar expression profiles might be regulated by the same transcription factor. The "Motif Discovery in *Drosophila*" walkthrough on the GEP website illustrates how we can use *MEME* to identify sequence motifs (e.g., transcription factor binding sites) that are enriched in a set of genes from the FlyBase Genes HitList.

Conclusions

This walkthrough demonstrates how we can use the FlyBase RNA-Seq tools to investigate the expression profiles of the *Gyf* gene on the *D. melanogaster* 4th chromosome, and to identify other genes that exhibit similar expression profiles. We used the TopoView tracks on FlyBase *GBrowse* to examine the unstranded RNA-Seq data from 30 developmental stages, and the strand-specific RNA-Seq data from the CNS and adult heads tissues. We then examined the "High-Throughput Expression Data" section of the FlyBase Gene Report to ascertain the gene expression levels of *Gyf* in different tissues.

We used the FlyBase RNA-Seq Expression Similarity Search tool to identify other *D. melanogaster* genes that exhibit similar tissue-specific expression profiles. Among all of the *D. melanogaster* genes, the tissue-specific expression profiles of *Fs(2)Ket* show the highest Spearman's correlation with the expression profiles of *Gyf*. Using the Gene2Function website, we found that the ortholog of *Fs(2)Ket* in human (*KPNB1*) has previously been associated with the neurological disorder multiple sclerosis. We also used the RNA-Seq Expression Similarity Search tool to identify two genes (*bon* and *Sin3A*) that show perfect correlations with the expression profiles of *Gyf* in a subset of tissues (i.e. imaginal disc, CNS, ovary, testis, and accessory glands). Finally, we used the RNA-Seq Expression Profile Search tool to identify genes that exhibit similar expression profiles across multiple modENCODE experiments (i.e. tissues and treatments).

While there are more sophisticated tools for performing RNA-Seq expression analyses, the suite of FlyBase RNA-Seq tools provides a powerful and more user-friendly platform to access and analyze the RNA-Seq data produced by the modENCODE project. These tools allow you to explore the expression patterns of your favorite gene, identifying genes that show similar expression patterns, and thus may share regulatory motifs. Based on the gene lists produced by the FlyBase RNA-Seq tools, we can use motif discovery tools such as *MEME* (Bailey *et al.* 2015) to identify common transcription factor binding sites, and use annotation tools such as *DAVID* (Huang *et al.* 2009) to perform functional annotations, classifications, and clustering. These investigations could potentially allow us to identify genes that are part of the same pathway, or are regulated by the same factors.

Additional resources

See the following resources for additional details on the RNA-Seq tools provided by FlyBase:

- FlyBase RNA-Seq overview page
- FlyBase RNA-Seq series on YouTube
- FlyBase 102
  - See the "RNA-Seq Search with query analysis" section
- FlyBase:Tools Overview wiki page
  - See the "Genomic Search Tools and Browsers" section

Literature Cited

Bailey, T. L., J. Johnson, C. E. Grant, and W. S. Noble, 2015 The *MEME* Suite. Nucleic Acids Res. 43: W39-49.

Brown, J. B., N. Boley, R. Eisman, G. E. May, M. H. Stoiber *et al.*, 2014 Diversity and dynamics of the *Drosophila* transcriptome. Nature 512: 393–399.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera *et al.*, 2016 A survey of best practices for RNA-Seq data analysis. Genome Biol. 17: 13.

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.

Hu, Y., A. Comjean, S. E. Mohr, FlyBase Consortium, and N. Perrimon, 2017 Gene2Function: An Integrated Online Resource for Gene Function Discovery. G3 Bethesda Md 7: 2855–2858.

Hu, Y., I. Flockhart, A. Vinayagam, C. Bergwitz, B. Berger *et al.*, 2011 An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12: 357.

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009 Systematic and integrative analysis of large gene lists using *DAVID* bioinformatics resources. Nat. Protoc. 4: 44–57.

Kim, M., I. Semple, B. Kim, A. Kiers, S. Nam *et al.*, 2015 *Drosophila* Gyf/GRB10 interacting GYF protein is an autophagy regulator that controls neuron and muscle homeostasis. Autophagy 11: 1358–1372.

Ripke, S., C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kähler *et al.*, 2013 Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. 45: 1150–1159.

Springelkamp, H., A. Mishra, P. G. Hysi, P. Gharahkhani, R. Höhn *et al.*, 2015 Meta-analysis of Genome-Wide Association Studies Identifies Novel Loci Associated With Optic Disc Morphology. Genet. Epidemiol. 39: 207–216.

St Pierre, S. E., L. Ponting, R. Stefancsik, P. McQuilton, and FlyBase Consortium, 2014 FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Res. 42: D780-788.

Zhang, Y., Q.-Y. Sun, R.-H. Yu, J.-F. Guo, B.-S. Tang *et al.*, 2015 The contribution of *GIGYF2* to Parkinson's disease: a meta-analysis. Neurol. Sci. Off. J. Ital. Neurol. Soc. Ital. Soc. Clin. Neurophysiol. 36: 2073–2079.