Gene Record Finder User Guide

Wilson Leung

Table of Contents

Introduction

One of the key requirements for students participating in the annotation projects from the Genomics Education Partnership (GEP; https://thegep.org) is to construct gene models for all the isoforms of the genes in their project. The comparative annotation strategy used by the GEP is based on parsimony with the orthologous gene from the informant genome (i.e. *D. melanogaster*). Hence the annotation protocol starts with the hypothesis that all the isoforms of a gene in the *D. melanogaster* ortholog also exist in the target species. The annotation strategy also begins with the hypothesis that the gene structure (e.g., the number and relative positions of the exons) are conserved between the two species. Changes in gene structure must be supported by experimental evidence (e.g., RNA-Seq data) or by sequence conservation in species that are more closely-related to the target species.

While many databases (e.g., NCBI, Ensembl, FlyBase, etc.) already provide research scientists with substantial amount of information about each gene, these resources are not optimized for the GEP annotation protocol. For example, because the same exon can be used by multiple isoforms, students using only the information from the FlyBase gene record will annotate the same exon multiple times. Hence using public databases to annotate all the isoforms of a gene is a labor-intensive and potentially error-prone process.

The *Gene Record Finder* is designed to supplement the information already available from FlyBase. It enables annotators to quickly identify a unique set of exons for a gene in *D. melanogaster*. Annotators can also use this tool to retrieve the sequences for the coding exons (CDSs) and the transcribed exons of a gene. Each CDS is listed separately, which enables annotators to use the "Align two or more sequences" functionality provided by NCBI *blastx* and *tblastn* to map each unique CDS onto their project sequence, and use these alignments to construct a gene model for each isoform.

This user guide provides an overview of the program and a tutorial on how to retrieve CDS and exon sequences for a *D. melanogaster* gene with multiple isoforms.

Acknowledgements

The *Gene Record Finder* is developed by Wilson Leung at Washington University in St. Louis for the Genomics Education Partnership (GEP).

Questions about the *Gene Record Finder*

Please contact Wilson (wleung@wustl.edu) if you have any questions or encounter any problems with the *Gene Record Finder*.

Availability

The *Gene Record Finder* is available under the "**Resources & Tools**" section of the F Element project page and the Pathways project page on the GEP website.

Overview of the *Gene Record Finder* interface

In this tutorial, we will use the *D. melanogaster* gene *gawky* to illustrate some of the key features of the *Gene Record Finder*.

Retrieving a gene record

Open a web browser window and navigate to the F Element project page on the GEP website. Click on the "Gene Record Finder" link under the "Resources & Tools" section. To retrieve a gene record, enter the official FlyBase gene symbol for the *D. melanogaster* gene of interests into the textbox. (**Note that gene symbols in *D. melanogaster* are case-sensitive**.)

As you type in the search box, an autocomplete box will appear which shows up to 20 gene symbols that match your query. In addition to the gene symbol, the autocomplete search results also provide additional information such as the number of isoforms, unique exons, and CDSs. Because the official FlyBase gene symbol for the *gawky* gene is "**gw**", we will enter "gw" into the textbox (Figure 1). Click on the "Find Record" button to retrieve the record.

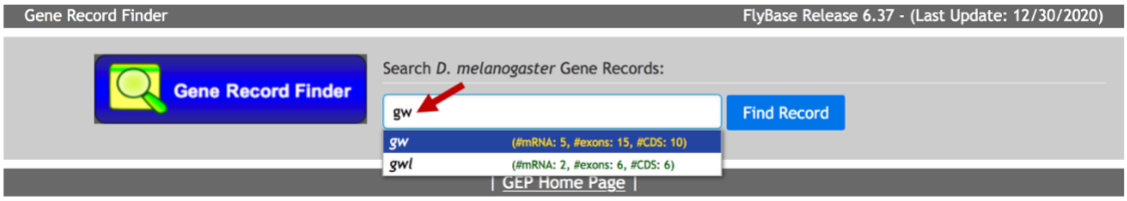


Figure 1. Search for the gene "gw" using the *Gene Record Finder*.

The *Gene Record Finder* output consists of four major sections: "Gene Details", "mRNA Details", "Transcript Details", and "Polypeptide Details" (Figure 2).

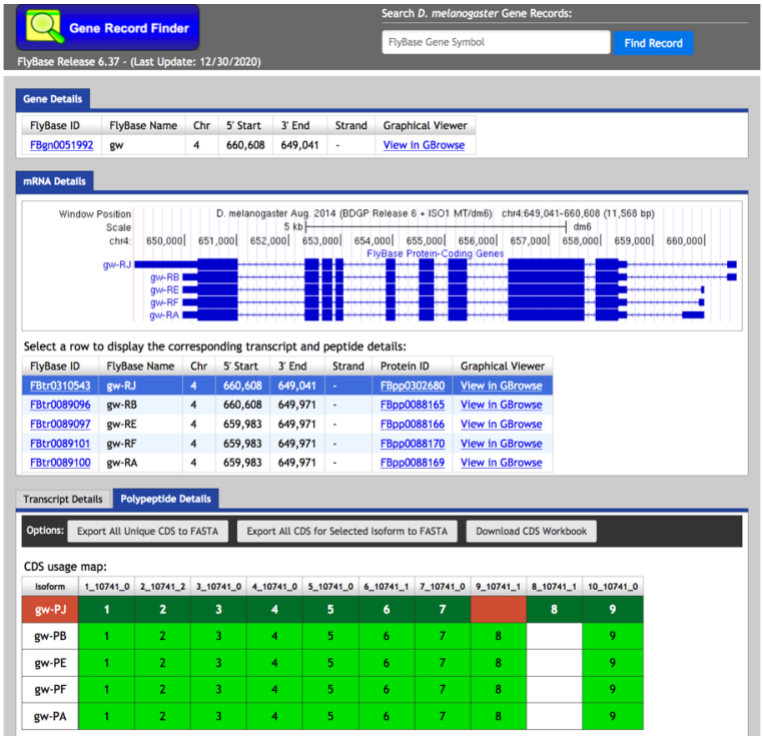
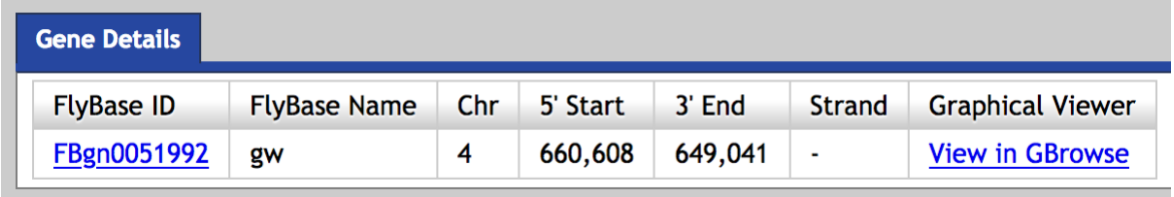


Figure 2. The *Gene Record Finder* search results page for the gene *gawky* (*gw*).

You can toggle between the “Transcript Details” and “Polypeptide Details” sections by clicking on the corresponding tab. You can enter the name of another gene into the textbox at the top right corner and then click on the “Find Record” button to retrieve the record for another gene.

Gene Details

The “Gene Details” section provides the span (i.e. the minimum position to the maximum position) that encompasses all isoforms of the specified gene in *D. melanogaster* (Figure 3). This section contains two links to FlyBase. The link under the “FlyBase ID” column (with the prefix “FBgn”) allows you to access the FlyBase gene report page for this gene. The “View in GBrowse” link under the “Graphical Viewer” column allows you to access the FlyBase *GBrowse* view of the genomic region surrounding this gene.



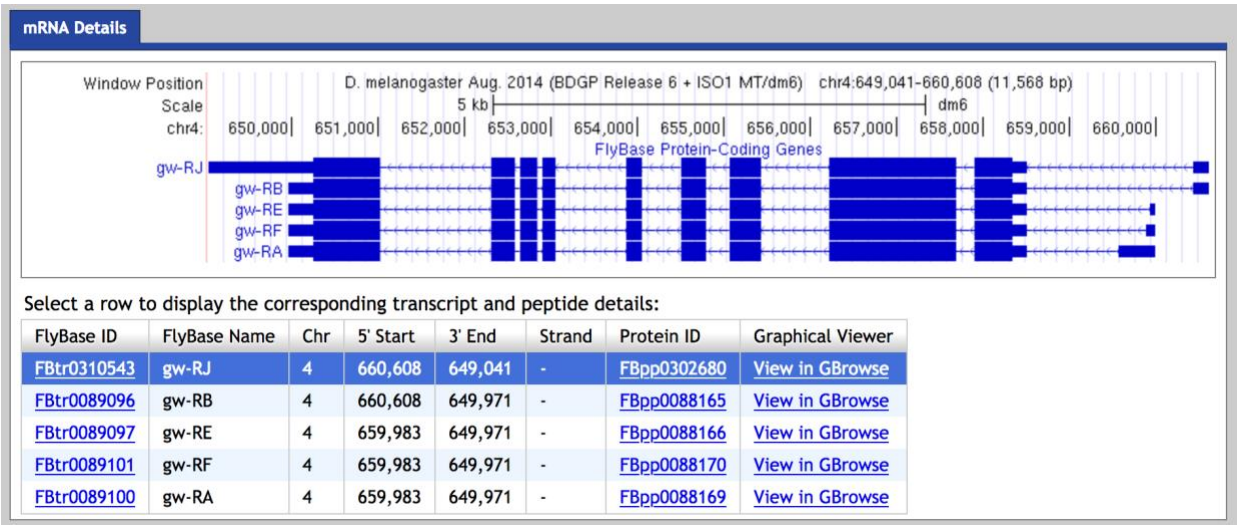
Gene Details

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0051992	gw	4	660,608	649,041	-	View in GBrowse

Figure 3. The “Gene Details” section shows that the *gw* gene is located at 660m608-649,041 on *D. melanogaster* chr4. The gene is on the minus strand relative to chr4.

mRNA Details

The “mRNA Details” section shows the list of transcripts associated with the gene. The *GEP UCSC Genome Browser* image in this section shows the positions of the untranslated regions (thin rectangles), CDSs (thick rectangles), and introns (lines) for each isoform (Figure 4).



Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0310543	gw-RJ	4	660,608	649,041	-	FBpp0302680	View in GBrowse
FBtr0089096	gw-RB	4	660,608	649,971	-	FBpp0088165	View in GBrowse
FBtr0089097	gw-RE	4	659,983	649,971	-	FBpp0088166	View in GBrowse
FBtr0089101	gw-RF	4	659,983	649,971	-	FBpp0088170	View in GBrowse
FBtr0089100	gw-RA	4	659,983	649,971	-	FBpp0088169	View in GBrowse

Figure 4. The “mRNA Details” section shows a Genome Browser image of the region surrounding the gene and the list of isoforms associated with the gene.

Clicking on the image will open a new window with the *GEP UCSC Genome Browser* for *D. melanogaster*, where the "FlyBase Transcribed Exons" and the "FlyBase Coding Exons" tracks show the positions of the unique exons and CDSs of the specified gene, respectively (Figure 5).

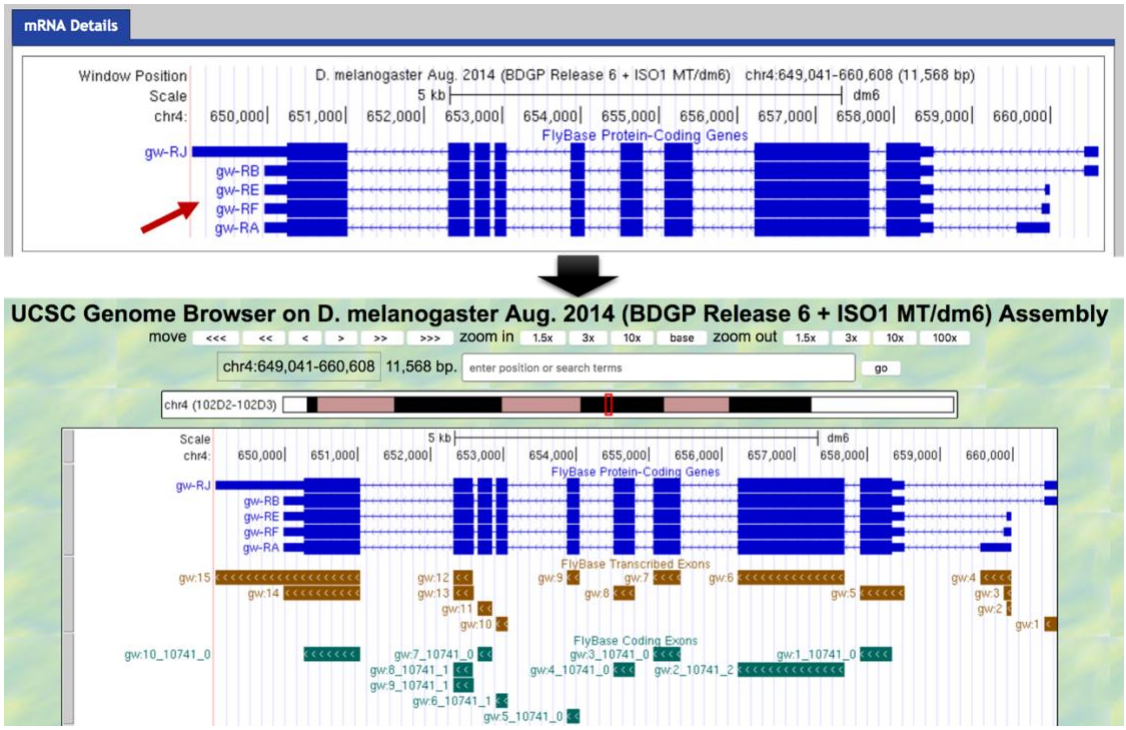


Figure 5. Click on the Genome Browser image in the "mRNA Details" section (top) to examine the relative positions of the transcribed exons and CDSs of a *D. melanogaster* gene (bottom). For example, the "FlyBase Coding Exons" track shows that the CDS 8_10741_1 overlaps with CDS 9_10741_1 in the *D. melanogaster* gene *gw*.

Each row in the mRNA data table corresponds to an isoform of the *gw* gene. (The selected isoform is highlighted in blue.) You can click on a row in the table to display the list of transcribed exons and CDSs for that isoform (Figure 6).

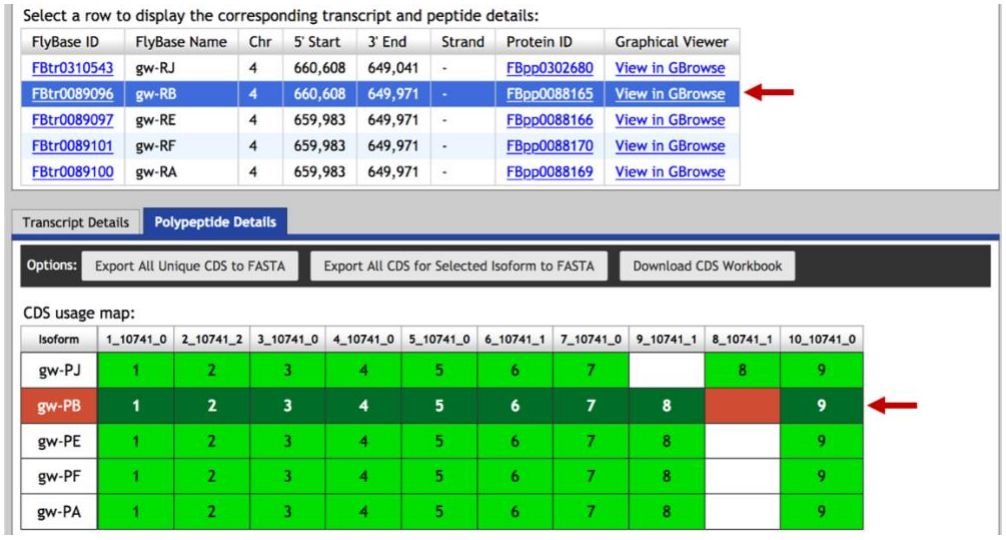


Figure 6. Click on a row in the mRNA table (e.g., gw-RB) to display the exon and CDS information for the selected isoform.

This table shows that the *gw* gene has five isoforms (gw-RJ, gw-RB, gw-RE, gw-RF, and gw-RA). There are three links on each row of the "mRNA Details" table. The links under the "FlyBase ID" column allow you to access the corresponding FlyBase transcript record, while the links under the "Protein ID" column allow you to access the corresponding FlyBase polypeptide record. The links under the "Graphical Viewer" column allow you to view the genomic region surrounding each transcript using *GBrowse* at FlyBase.

Polypeptide Details

The "Polypeptide Details" section shows the list of CDSs associated with the isoform you have selected in the "mRNA Details" section. By convention, the part of the isoform name following the last dash corresponds to the name of the isoform. Transcripts have a "-R" suffix while polypeptides have a "-P" suffix. For example, the transcript for the B isoform of *gw* is known as "gw-**R**B" while the corresponding polypeptide is known as "gw-**P**B".

Under the "Options" toolbar, you can export all the unique CDS sequences for the gene or just for the selected isoform. These sequences are in FASTA format, and can be used directly in the NCBI "align two or more sequences" (bl2seq) *BLAST* searches (Figure 7).

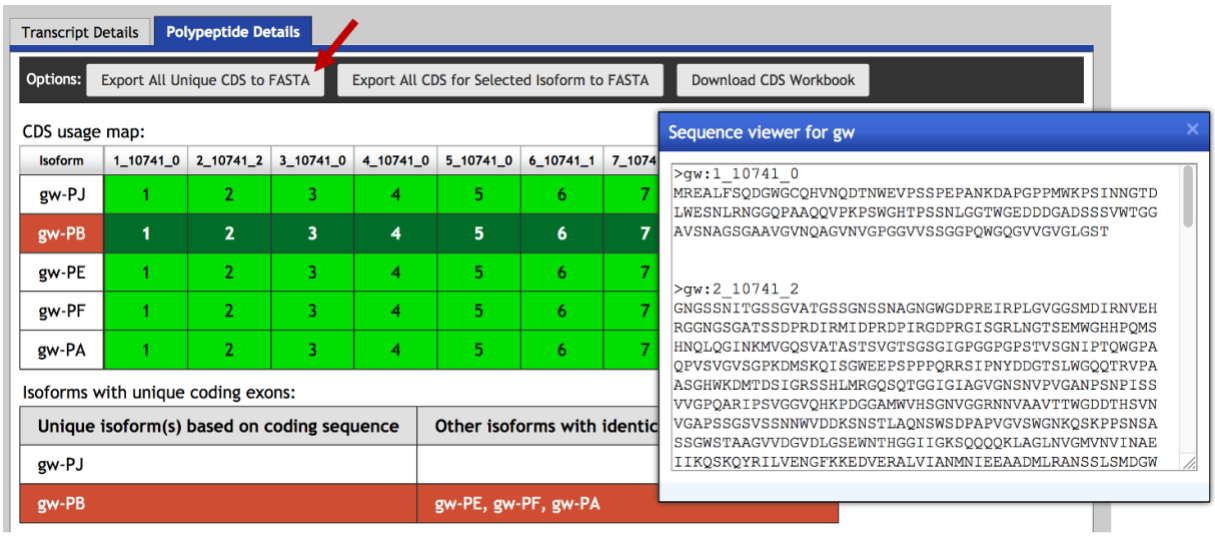


Figure 7. Click on the "Export All Unique CDS to FASTA" button in the "Options" toolbar to retrieve all the unique CDS sequences for the *gw* gene. Click on the "Export All CDS for Selected Isoform to FASTA" button to retrieve the CDS sequences for the isoform you have selected (i.e. gw-PB).

The CDS usage map summarizes the usage of each CDS in the different isoforms (Figure 8). Each column in the CDS usage map corresponds to a unique CDS in the gene and the column header contains the unique identifier for each CDS. The CDS are ordered from 5' to 3' from left to right. Each row in the CDS usage map corresponds to an isoform of *gw* and the selected isoform (e.g., gw-PB) is highlighted in dark red under the "Isoform" column.

CDS usage map:

Isoform	1_10741_0	2_10741_2	3_10741_0	4_10741_0	5_10741_0	6_10741_1	7_10741_0	9_10741_1	8_10741_1	10_10741_0
gw-PJ	1	2	3	4	5	6	7		8	9
gw-PB	1	2	3	4	5	6	7	8		9
gw-PE	1	2	3	4	5	6	7	8		9
gw-PF	1	2	3	4	5	6	7	8		9
gw-PA	1	2	3	4	5	6	7	8		9

Figure 8. The CDS usage map shows that the differences in CDS usage for the different isoforms of *gw* are limited to the 3' end of the gene. The green boxes denote the CDS that are used by each isoform, whereas the empty boxes denote CDSs that are not used by each isoform. The number within each box corresponds to the CDS number with respect to the isoform.

Each cell in the CDS usage map indicates whether the CDS in each column is used by the isoform in each row. A green box indicates that the isoform uses the CDS and a white box indicates that the isoform does not use the CDS. (For the selected isoform, the boxes are in dark green and dark red, respectively.) The number in each box corresponds to the CDS number for the isoform within each row. For example, CDS 10_10741_0 is the ninth CDS of gw-PB.

You can retrieve the sequence for a specific CDS by clicking on the column header for the CDS in the CDS usage map (Figure 9). You can click on each row to change the selected isoform.

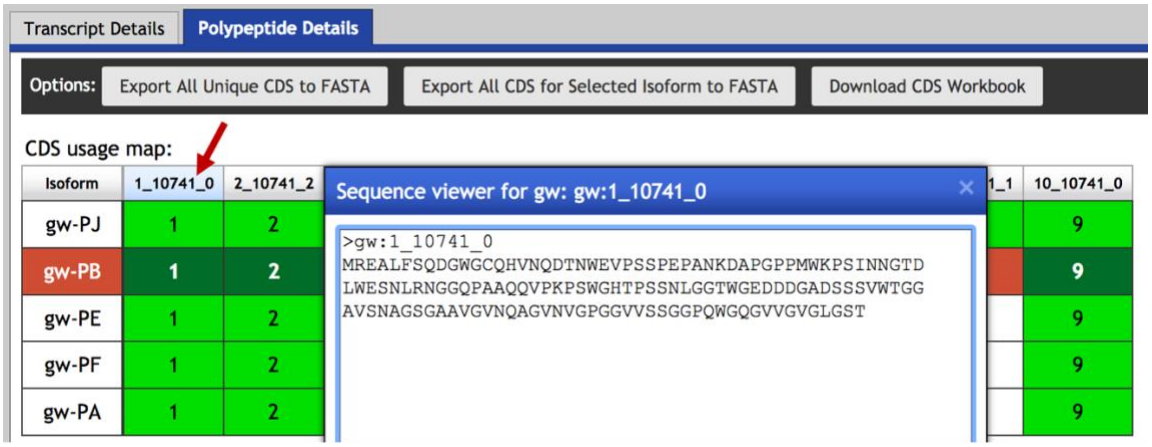


Figure 9. Click on the column header (e.g., 1_10741_0) in the CDS usage map to retrieve the sequence for the CDS.

To help students organize their annotation efforts, the *Gene Record Finder* can also generate an Excel Workbook that contains the list of unique exons for each gene. Click on the “Download CDS Workbook” button on the “Options” toolbar to download the Excel Workbook (Figure 10).

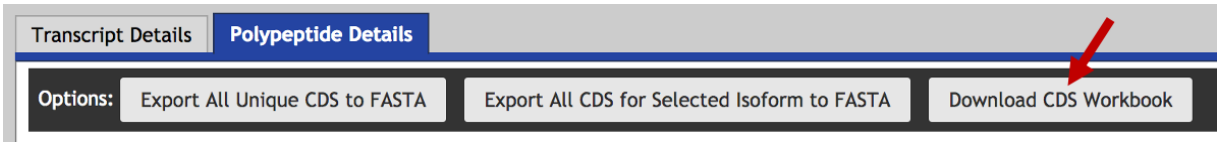
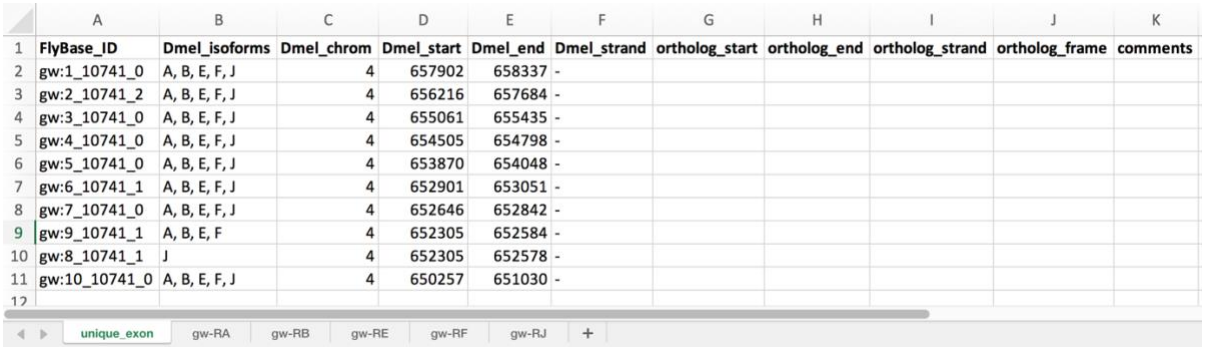


Figure 10. Click on the "Download CDS Workbook" button to download an Excel Workbook configured for keeping track of CDS coordinates.

The Excel workbook consists of multiple worksheets: the “unique_exon” worksheet contains the list of unique exons for the gene and the remaining worksheets contain the list of exons found in each isoform. For example, there are five worksheets besides the “unique exon” worksheet that corresponds to the A, B, E, F, and J isoforms of *gw*, respectively (Figure 11).



	A	B	C	D	E	F	G	H	I	J	K
1	FlyBase_ID	Dmel_isoforms	Dmel_chrom	Dmel_start	Dmel_end	Dmel_strand	ortholog_start	ortholog_end	ortholog_strand	ortholog_frame	comments
2	gw:1_10741_0	A, B, E, F, J	4	657902	658337	-					
3	gw:2_10741_2	A, B, E, F, J	4	656216	657684	-					
4	gw:3_10741_0	A, B, E, F, J	4	655061	655435	-					
5	gw:4_10741_0	A, B, E, F, J	4	654505	654798	-					
6	gw:5_10741_0	A, B, E, F, J	4	653870	654048	-					
7	gw:6_10741_1	A, B, E, F, J	4	652901	653051	-					
8	gw:7_10741_0	A, B, E, F, J	4	652646	652842	-					
9	gw:9_10741_1	A, B, E, F	4	652305	652584	-					
10	gw:8_10741_1	J	4	652305	652578	-					
11	gw:10_10741_0	A, B, E, F, J	4	650257	651030	-					
12											

unique_exon | gw-RA | gw-RB | gw-RE | gw-RF | gw-RJ | +

Figure 11. The Excel workbook generated by the *Gene Record Finder*. The workbook for *gw* consists of six worksheets: a “unique_exon” worksheet and five additional worksheets for each isoform of gw.

When you enter the ortholog coordinates and the strand (“+” or “-”) for an exon in the “unique_exon” worksheet, the corresponding exon entry in the other isoform worksheets will update automatically (Figure 12). (Note that, irrespective of the orientation of the gene, the “ortholog_start” coordinate should be less than or equal to the “ortholog_end” coordinate.)

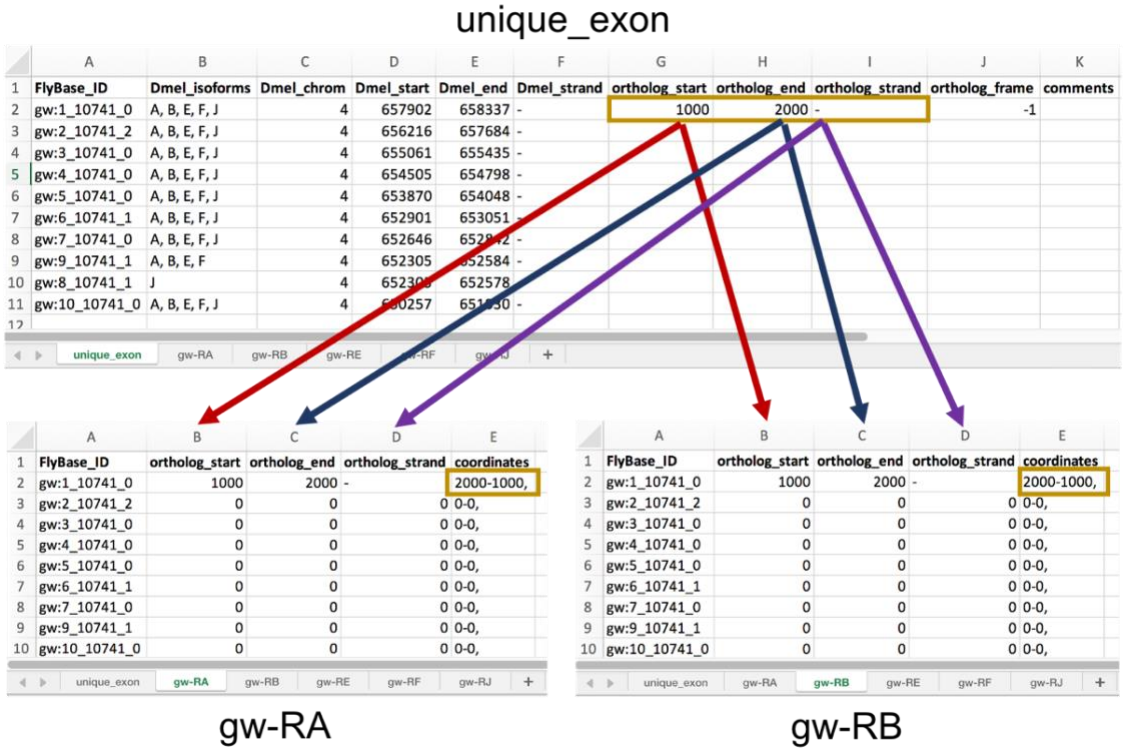


Figure 12. Entries in the ortholog_start, ortholog_end, and ortholog_strand columns in the “unique_exon” worksheet will automatically propagate to the isoform worksheets.

The “coordinates” column of each isoform worksheet contains the exon coordinates that have been pre-formatted for use in the *Gene Model Checker*. Once you have completed the “unique_exon” worksheet, you can copy all the values under the “coordinates” column of the Excel Workbook and paste them into the “Coding Exon Coordinates” field of the *Gene Model Checker* to verify your gene model.

Isoforms with unique coding exons

This section only appears in the *Gene Record Finder* output if the gene has multiple isoforms with identical polypeptide sequences. Each row corresponds to a unique set of CDSs used by one or more isoforms. The isoforms listed under the “Other isoforms with identical coding sequences” column use the same set of CDSs as the isoform listed under the “Unique isoform(s) based on coding sequence” column in the corresponding row. The selected isoform is highlighted in dark red.

For example, the unique isoforms table shows that the E, F, and A isoforms of *gw* use the same set of CDSs as the B isoform, and that the J isoform uses another unique set of CDSs (Figure 13).

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences
gw-PJ	
gw-PB	gw-PE, gw-PF, gw-PA

Figure 13. The “Other isoforms with identical coding sequences” column shows the list of isoforms that use the same set of CDSs as the isoform listed in the corresponding row under the “Unique isoform(s) based on coding sequence” column. The selected isoform is highlighted in dark red (i.e. gw-PB).

Sequence table

The last section of the “Polypeptide Details” tab is the CDS sequence table. This table lists the position, phase, and the size of each CDS. You can use the column headers to sort the table by “5' Start”, “3' End” or “Size (aa)” in either ascending or descending order. You can also click on a row to retrieve the corresponding CDS sequence (Figure 14).

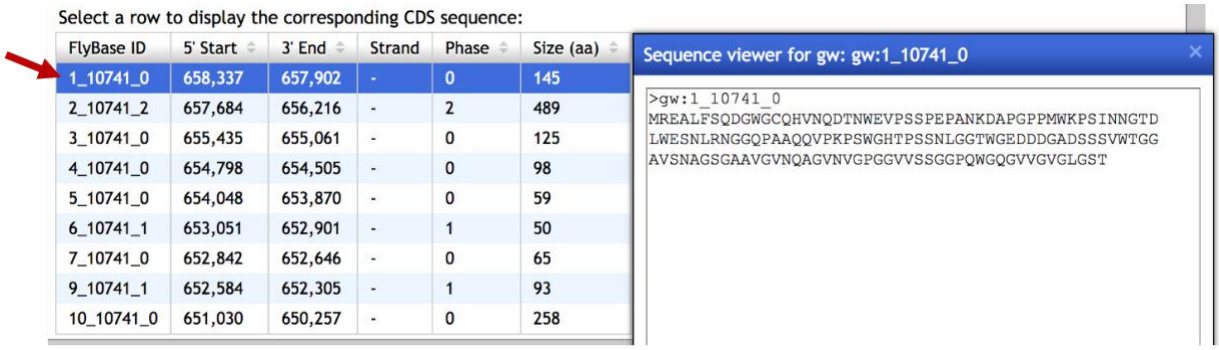


Figure 14. Click on a row in the CDS sequence table to retrieve the corresponding CDS sequence (e.g., CDS 1_10741_0).

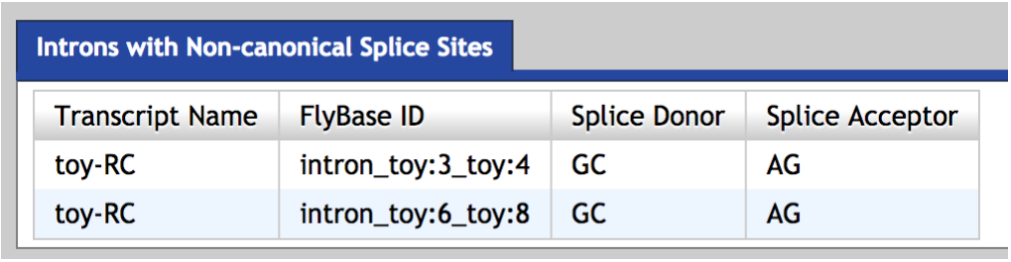
Transcript Details

This section has the same format as the "Polypeptide Details" tab. The only difference is that the Transcript Details tab contains information for all the transcribed exons (which includes both translated and untranslated regions), and it consists of nucleotide sequences instead of amino acid sequences.

Introns with Non-canonical Splice Sites

This section will only appear in the *Gene Record Finder* output if the gene has at least one intron with a non-canonical splice donor or acceptor site (Figure 15). The identifier in the "FlyBase ID" column is based on the names of the transcribed exons (separated by an underscore) that are adjacent to the intron with non-canonical splice sites.

For example, the FlyBase ID "intron_toy:3_toy:4" indicates that the intron is located between the transcribed exons "toy:3" and "toy:4" of the *toy* gene. The "Splice Donor" and "Splice Acceptor" columns show the splice site sequences for the intron. The table in this section will update automatically when you select a different isoform.



Introns with Non-canonical Splice Sites

Transcript Name	FlyBase ID	Splice Donor	Splice Acceptor
toy-RC	intron_toy:3_toy:4	GC	AG
toy-RC	intron_toy:6_toy:8	GC	AG

Figure 15. The "Introns with Non-canonical Splice Sites" section shows the C isoform of the *toy* gene has two introns with non-canonical GC splice donor sites (i.e. the intron between exons toy:3 and toy:4, and the intron between exons toy:6 and toy:8).

Detailed explanation of the fields in each data table

Gene Details

Field	Explanation
FlyBase ID	Link to the FlyBase gene record
FlyBase Name	Gene symbol assigned by FlyBase
Chr	The chromosome where the gene is located
5’ Start	5' start coordinate of the gene span
3’ End	3' end coordinate of the gene span
Strand	The orientation of the gene relative to the chromosome
Graphical Viewer	Link to FlyBase *GBrowse* view of the region surrounding the gene

mRNA Details

Field	Explanation
FlyBase ID	Link to the FlyBase transcript record
FlyBase Name	Name of the FlyBase transcript
Chr	The chromosome where the transcript is located
5’ Start	5' start coordinate of the transcript
3’ End	3' end coordinate of the transcript
Strand	The orientation of the transcript relative to the chromosome
Protein ID	Link to the FlyBase protein record derived from the transcript
Graphical Viewer	Link to FlyBase *GBrowse* view of the region surrounding the transcript

Introns with Non-canonical Splice Sites

Field	Explanation
Transcript Name	Name of the FlyBase transcript
FlyBase ID	Intron ID: intron_<exon:1>_<exon:2>
Splice Donor	Nucleotide sequence of the splice donor site
Splice Acceptor	Nucleotide sequence of the splice acceptor site

Polypeptide Details (Sequence Table)

Field	Explanation
FlyBase ID	Unique identifier for the CDS
5’ Start	The 5' start coordinate of the CDS
3’ End	The 3' end coordinate of the CDS
Strand	The orientation of the CDS relative to the chromosome
Phase	Number of bases until the first base of the first complete codon (i.e. acceptor phase)
Size (aa)	Length of the CDS sequence (number of amino acids)

Transcript Details (Sequence Table)

Field	Explanation
FlyBase ID	Unique identifier for the transcribed exon
5’ Start	The 5' start coordinate of the transcribed exon
3’ End	The 3' end coordinate of the transcribed exon
Strand	The orientation of the exon relative to the chromosome
Size (bp)	Length of the exon sequence (number of nucleotides)