VASSAR COLLEGE URSI: UNDERGRADUATE RESEARCH SUMMER INSTITUTE SYMPOSIUM | 2021

# Analysis Across *Drosophila* Species by Combining Muller Element Analysis and *dawdle* Gene Annotations

Trina Chou '23 and Professor Jennifer Kennell
Vassar College Department of Biology

## INTRODUCTION

In conjunction with the Genomics Education Partnership, different genomic regions of *Drosophila* fruit flies were annotated and analyzed to learn more about their evolutionary relationships. Two projects were tackled:

**1) F element project**

i) Background: The Muller F Element is the fourth chromosome in *Drosophila melanogaster* and has undergone extreme expansion in other *Drosophila* species such as *D. ananassae* and *D. bipectinata*. This chromosome is highly expressed despite having heterochromatic properties.

ii) Evidence: Annotation of two overlapping coding regions of *D. ananassae*, contig28 and contig29.

iii) Analysis: Assess the impact of expansion of the Muller F element in reference to the non-expanding Muller D element.

***2) dawdle* gene project**

i) Background: The *dawdle* gene (*daw*) is a TGF beta/Activin-type ligand involved in insulin-regulated fruit fly larval development. It's ortholog in humans is associated with several diseases, including cystic fibrosis.

ii) Evidence: Annotation of the coding regions and transcription start sites of the *daw* gene across ten *Drosophila* species.

iii) Analysis: Comparison of the synteny (local genomic neighborhood) and gene sequence across all species.

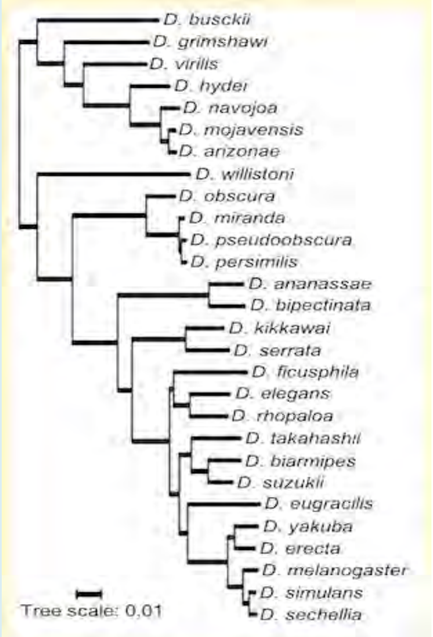


**Figure 1:** Phylogenetic tree depicting hypothesized relationships between different *Drosophila* species. (Image from the GEP)

## METHODOLOGY

### 1. Coding Region Annotation

| UCSC Genome Browser | FlyBase BLAST | FlyBase Gene Record Finder | BLASTx Alignment | UCSC Genome Browser | GEP Model Checker |
|---|---|---|---|---|---|
| • Visual inspection of the contig | • Orthologs determined from predicted protein sequences given by gene predictors | • CDS sequences noted from *D. melanogaster* ortholog | • Course start and end of each CDS located on target species | • Splice donor/acceptor rules, RNA Seq data, phase matches, and start/end codons used for fine annotation | • Final coordinates entered to check the proposed gene model |

### 2. Untranslated Region Annotation

| UCSC Genome Browser | FlyBase Gene Record Finder | BLASTn Alignment | UCSC Genome Browser | Core Promoter Motif Search |
|---|---|---|---|---|
| • Core promoter shape on *D. melanogaster* determined with Genome Browser Tracks | • Transcript sequences noted from *D. melanogaster* ortholog | • 5' untranslated exons mapped on target species | • Core promotor motifs, BLAT, RAMPAGE, RNA Seq, and Conservation data used to find TSS window | • 12 core promoter motifs located within ±300 bp of the TSS window as additional evidence |

### 3. Synteny Analysis

- UCSC Genome Browser and BLAT used to locate the two nearest predicted genes on either end of the target gene.
- Genes are inputted into BLAST to gain information about orientation and predicted proteins.
- Synteny of different species are then compared against one another to visualize differences in genomic neighborhoods.

## RESULTS: F ELEMENT ANNOTATION

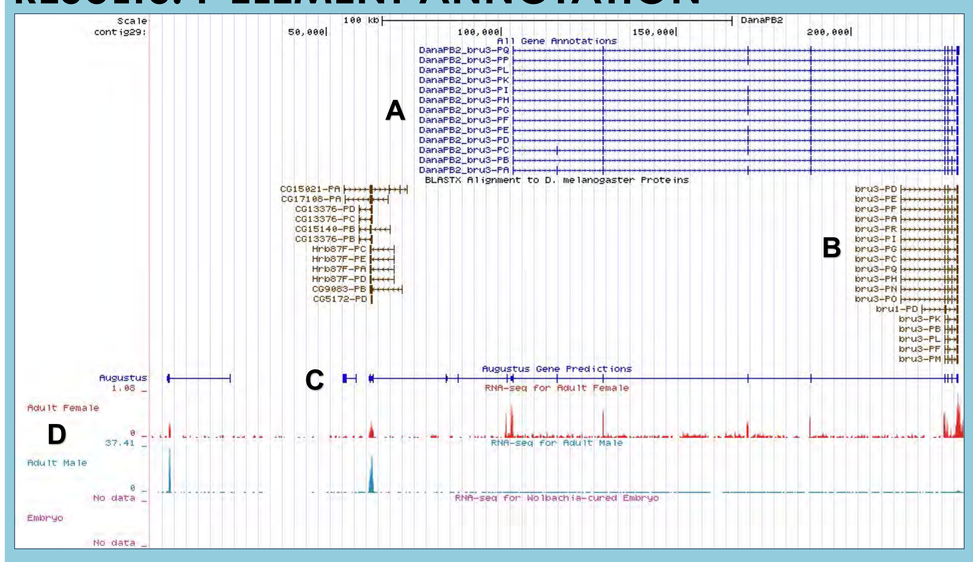


**Figure 2**: UCSC Genome Browser of *D. ananassae* Muller D contig29, which is 282,000bp long.
**A)** Resulting annotations of 13 isoforms of *bru3*
**B)** BLASTx alignments of predicted genes on the contig
**C)** Augustus gene predictor track
**D)** RNA Seq data for adult Females in red and adult males in blue

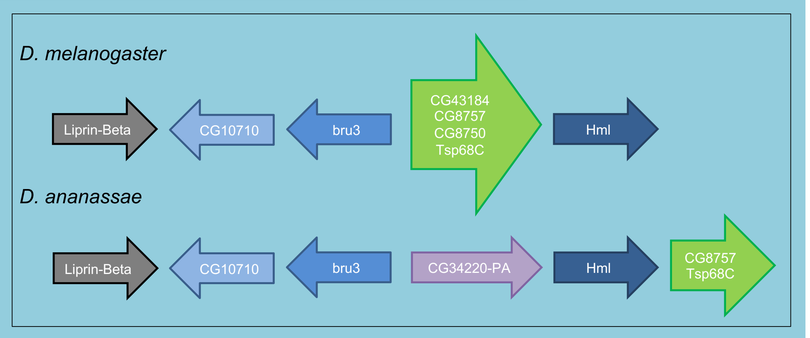


**Figure 3**: Synteny comparison of genes around *bru3* in both *D. melanogaster* and *D. ananassae.* The two genes nearest upstream of the gene remain the same in both species, however the genes downstream are different. Several smaller genes on *D. melanogaster* are found past the *Hml* gene on *D. ananassae.*

| Species | Chromosome | Scaffold | Gene | Coding Span | Total Intron Size | Total CDS Size | # of CDS | Median CDS | Median Intron |
|---|---|---|---|---|---|---|---|---|---|
| *D. melanogaster* | Muller D | 3L | bru3 | 125735 | 124484 | 1251 | 10 | 113.5 | 1444 |
| *D. ananassae* | Muller D | contig29 | bru3 | 127078 | 125839 | 1239 | 10 | 112 | 1376 |
| *D. melanogaster* | Muller D | 3L | CG34220 | 2237 | 56 | 2181 | 2 | 1090.5 | 56 |
| *D. ananassae* | Muller D | contig28 | CG34220 | 2984 | 62 | 2922 | 2 | 1461 | 62 |
| *D. melanogaster* | Muller F | 4 | anne | 8267 | 3911 | 4356 | 8 | 399 | 63 |
| *D. bipectinata* | Muller F | contig2 | anne | 33411 | 29046 | 4365 | 8 | 391.5 | 60 |
| *D. melanogaster* | Muller F | 4 | CG11155 | 9067 | 6334 | 2733 | 14 | 182.5 | 199 |
| *D. bipectinata* | Muller F | contig2 | CG11155 | 269459 | 266750 | 2709 | 14 | 176.5 | 81 |

**Table 1:** Comparison of gene characteristics of Muller D element genes compared to Muller F element genes. As seen in the highlighted column, the intron size of the Muller D genes are similar in both *D. melanogaster* and *D. ananassae.* However, there is a drastic difference in the Muller F genes between *D. melanogaster* and *D. bipectinata.*

## RESULTS: *DAWDLE* GENE ANNOTATION

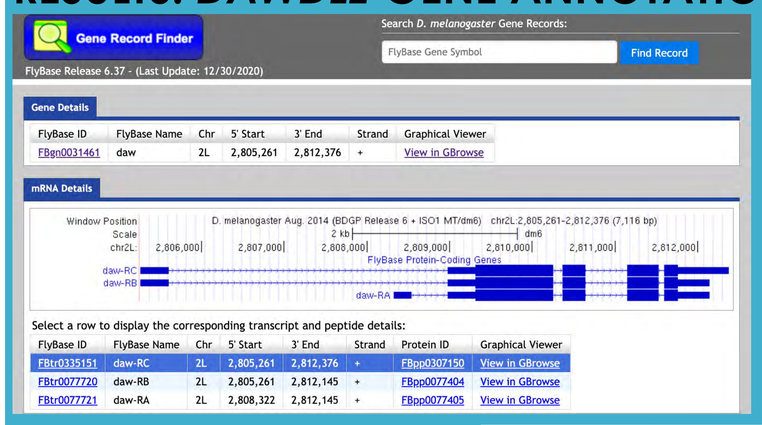


**Figure 4:** Gene Record Finder displaying the gene characteristics of the *daw* gene on *D. melanogaster*. This particular gene has 4 coding sequences (wide boxes) and three isoforms which only differ in their transcription start sites and untranslated regions (narrow boxes).

| Species | Most Upstream | Nearest Upstream | Target Gene | Nearest Downstream | Most Downstream |
|---|---|---|---|---|---|
| *D. melanogaster* | *CG15399* | *syt1* | *daw* | *CG2964* | *G6P* |
| | - | - | + | + | - |
| *D. yakuba* | *CG15399* | *syt1* | *daw* | *CG2964* | *G6P* |
| | - | - | + | + | - |
| *D. erecta* | *CG15399* | *syt1* | *daw* | *CG2964* | *G6P* |
| | - | - | + | + | - |
| *D. ananassae* | *CG44141* | *CG4587* | *daw* | *Zw* | *CG2964* |
| | + | + | + | - | + |
| *D. persimilis* | *CG44141* | *CG4587* | *daw* | *CG2964* | *G6P* |
| | - | - | - | - | + |
| *D. suzukii* | *CG15399* | *syt1* | *daw* | *CG2964* | *G6P* |
| | + | + | - | - | + |
| *D. willistoni* | *CycE* | *CG4587* | *daw* | *G6P* | *Duox* |
| | - | - | - | + | + |
| *D. mojavensis* | *CycE* | *CG4587* | *daw* | *nrv2* | *nrv1* |
| | + | + | + | + | + |
| *D. virilis* | *Sfp24F* | *CG4587* | *daw* | *Gasp* | *CG9171* |
| | + | - | - | + | + |
| *D. grimshawi* | *CycE* | *CG4587* | *daw* | *CG18661* | *Try29F* |
| | - | - | - | - | - |

**Table 2:** Comparison of synteny around the target gene across all ten species, including the strand on which the gene was found. The two nearest genes in either direction of the target gene were recorded.

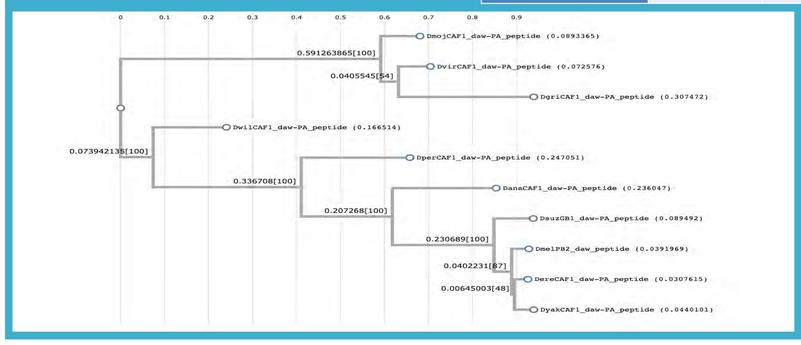


**Figure 5:** Phylogenetic tree depicting the relationships between the ten species based on the protein sequences generated from the *daw* annotations. Since the coding sequences of the isoforms are identical, only the protein sequences for one of the isoforms were compared. *D. melanogaster* is shown at the bottom, with the furthest related species shown at the top.

## DISCUSSION: F ELEMENT ANALYSIS

- Of the 18 *bru3* isoforms found on *D. melanogaster*, only 13 were found on *D. ananassae.* Five isoforms were lost due to the loss of five specific coding regions.
- *CG34220* is found next to *bru3* in only *D. ananassae*, however it is still on chromosome 3L in *D. melanogaster*.
- Comparison of overall gene characteristics show that Muller D genes have not undergone intron expansion in the same way Muller F genes have.

## DISCUSSION: *DAWDLE* GENE ANALYSIS

- Synteny in close species is conserved. As the species become more distantly related, synteny starts to break down.
- *D. melanogaster* has no synteny similarities around *daw* with four out of the nine other species.
- The hypothesized relationships between the species are reflected by the relationships dictated by the *daw* gene protein sequences, except for *D. virilis* being shown as more closely related to *D. grimshawi* than to *D. mojavensis*.

## References

- The Genomics Education Partnership website has links to all the information and tools used for this project (http://thegep.org)

## Acknowledgements

I would like to thank Professor Kennell for her guidance, mentorship, and support throughout this research. I would also like to thank Salome Ambokadze and Veronica Gomez for their collaboration efforts on several contigs. In addition, I would like to thank the GEP for making collaboration regarding these projects possible.