Exploring Evolution of Heterochromatic Genes and Signaling Pathways in *Drosophila* using Comparative Analysis

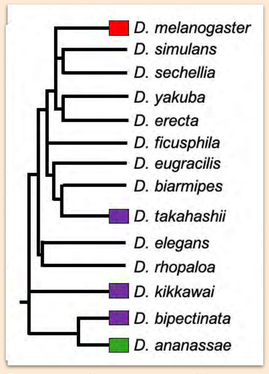
Salome Ambokadze '23 and Prof. Jennifer Kennell; Department of Biology

Project 1: Studying the Expansion of Muller F Element in *Drosophila* by Annotating the Coding Spans of its Genes

- The 4th chromosome, or the F element, exhibits unique properties in *Drosophila melanogaster*: it is packaged as heterochromatin (high repeat and methylation content, late replication), but exhibits gene expression levels of euchromatin.
- The F element is known to have disproportionately expanded in at least 4 *Drosophila* species: *D. ananassae, D. bipectinata, D. kikkawai, D. takahashii.*

	Size	Repeat content	Partial karotype
- *D. melanogaster*	1.3Mb	28%	4
- *D. ananassae*	17.8Mb	75%	4L 4R

Comparison of *D. mel* and *D. ana* F element length and % of the repeating elements, measured in Megabases. *D. ana* F element is both longer and richer in repeat content.



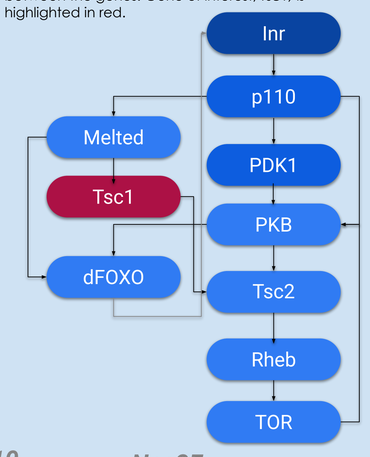
Phylogenetic tree summarizing the evolutionary relationship between the target *Drosophila* species and *D. melanogaster.*

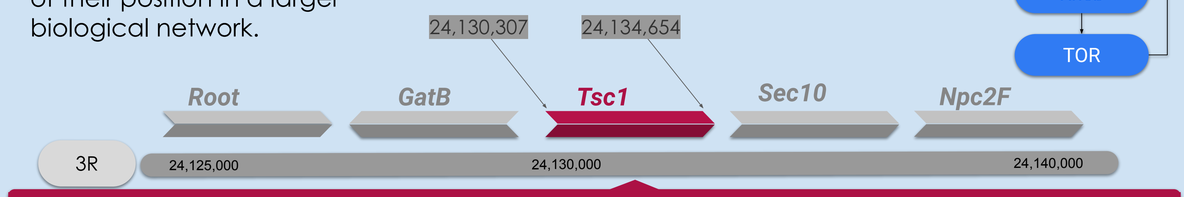
- In collaboration with GEP(Genomics Education Partnership), we have annotated the coding spans in two genomic regions (contigs) of the F element: one in *D. ananassae* and one in *D. bipectinata*.
- Goal is to identify precise intron/exon boundaries of the genes producing coherent coding region models (along with their transcript and peptide files).
- Annotations gathered by GEP will provide insight into the evolutionary impacts of chromosome and gene expansion, as well as how genes function within heterochromatin.

Project 2: Evolutionary Analysis of *Drosophila* Gene *Tsc1* in the Context of Insulin-Signaling Pathway

- The Pathways Project, conducted by GEP, aims to better understand the evolution of *Drosophila* genes in the context of their role within a biological pathway.
- Current focus is the insulin-signaling pathway of *Drosophila,* which is highly conserved among species and crucial to growth and metabolic homeostasis.
- This study focuses on structural evolution of a single gene within this pathway, *Tsc1,* across 9 different *Drosophila* species.
- Entire transcript spans (including the Transcription Start Sites, TSS) of the *Tsc1* gene have been annotated in this project.
- Annotations gathered by GEP will provide a better understanding of the regulatory region evolution in the context of their position in a larger biological network.

Summary of the signaling pathway downstream of the insulin-like hormone in *Drosophila Melanogaster.* Arrows indicate the direction of signal transduction between the genes. Gene of interest, *Tsc1,* is highlighted in red.





Tsc1 is a protein-coding gene in *D. melanogaster,* located on the 3rd right chromosome. It encodes the tumor-suppressor protein Tsc1 which forms a complex with Tsc2 (protein product of the gene *gig)*. Tsc1/Tsc2 complex controls cellular growth by antagonizing insulin signaling: it inhibits TOR, which is the central controller of cell growth.

Methods:

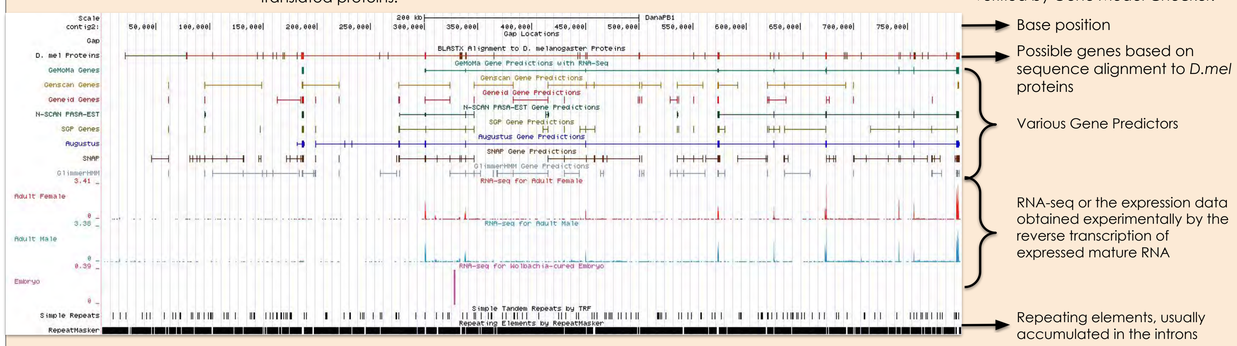
1. Inspecting contigs for possible genes
2. Finding the ortholog in *D. melanogaster*
3. Mapping CDS against the contig assembly
4. Refining coordinates

Gene Predictor and Protein Alignment tracks indicate the most likely genes, and are used to narrow down the search region.

To find the ortholog, Protein sequence of gene predictors in aligned to known *D.mel* protein-coding genes using *Flybase.* Gene structure (number and length of each coding exon or CDS) of the ortholog is obtained using *Gene Record Finder*, which also provides sequences to translated proteins.

NCBI blastx is used to align *D.mel* protein sequences against the contig nucleotide sequences, providing rough coordinates of exon-intron boundaries and the reading frame of translation.

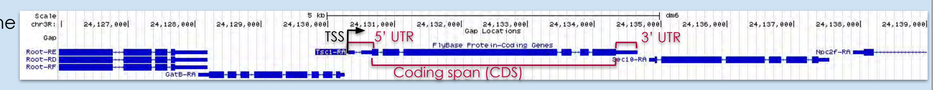
Coordinates are checked for splicing donor (GT) and acceptor sites (AG). It is made sure that there are no stop codons within coding spans. Gene models are verified by *Gene Model Checker.*



Default interface of *D. ananassae* contig 2 (800,000 bases) in GEP *Genome Browser,* Sep. 2019 (GEP/Muller F Element) Assembly To annotate coding spans along contigs experimental data (expression/RNA-seq) is combined with conservation data, or the sequence similarity to known *D. melanogaster* proteins. Gene models aim to resemble the ortholog (in protein length and exon number)as closely as possible.

Methods:

GEP genome browser view of *Tsc1* genomic neighborhood, chromosome 3R. Narrow bands are UTRs. Thick bands are CDS.



1. GEP genome browser: Examine the upstream and downstream genes (synteny) and obtain the protein sequence.

2. Tblastn: The protein sequence is aligned against the genome assembly of the target species. This provides the accession number (rough genomic region) of the *Tsc1* ortholog, which allows further analysis of the genomic neighborhood and gene structure.

Query: *Tsc1* protein sequence from *D.mel.* → **Subject:** Genome assembly of the target species

3. Blastp: Protein products of target and nearby genes are then aligned against *D. melanogaster ref_seq protein* database, to ensure that two upstream and downstream genes are orthologs of *D. mel* genes.

Query: protein sequence of target species' assembly → **Subject:** *D.mel* known protein database

4. Gene Record Finder: detailed information on exon/intron boundaries, direction of transcription, and peptide sequences of each of the six CDS are obtained from the *Gene Record Finder.* CDS sequences are mapped against the accession of the target species. Rough coordinates are obtained, which are fine-mapped using RNA-seq data and splice donor/acceptor searches.

5. TSS and UTR annotation:

- *D.mel* transcripts (*Gene Record Finder*) are aligned against species genome assembly.
- To find the Transcription Start Site, start location of RNA-seq and *blastn* alignment are inspected
- Core promoter motifs (short sequences) and RAMPAGE read density (experimental evidence), if present, could indicate exact coordinates.

CDS usage map:

Isoform	1_2219_0	2_2219_2	3_2219_0	4_2219_2	5_2219_0	6_2219_2
Tsc1-PA	1	2	3	4	5	6

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_2219_0	24,130,666	24,130,759	+	0	31
2_2219_2	24,130,820	24,131,076	+	2	85
3_2219_0	24,131,138	24,133,448	+	0	770
4_2219_2	24,133,506	24,133,663	+	2	52
5_2219_0	24,133,790	24,133,919	+	0	43
6_2219_2	24,133,977	24,134,329	+	2	117

Six protein-coding exons of *D.mel Tsc1*, their coordinates, strand of transcription/translation, phase, and polypeptide-product length, as shown on the *Gene Record Finder*.

Results:

Coding regions of Contig 2 in *D. ananassae* and Contig 12 in *D. bipectinata* have been fully annotated

Figure 1: Merged model of the coding spans of the genes annotated on *D. ananassae* contig2. Three gene models, colored in blue, have been produced: ***fd102C*** (one isoform), ***Ekar*** (three isoforms), and ***Gat*** (two isoforms).

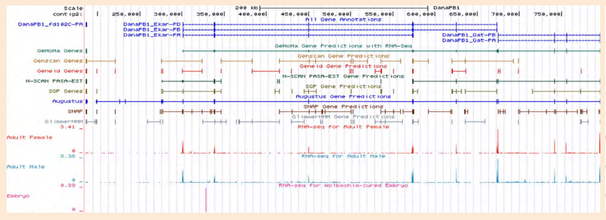
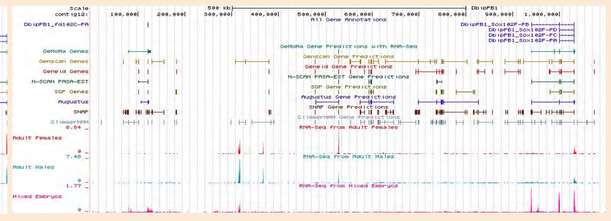
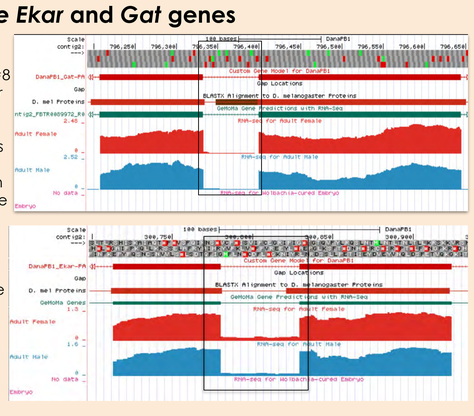


Figure 2: Merged model of the coding spans of the genes annotated on *D. bipectinata* contig12. Two gene models, colored in blue, have been produced: ***fd102C*** (one isoform), and ***Sox102F*** (four isoforms).



Additional introns have been found in *D. ananassae Ekar* and *Gat* genes

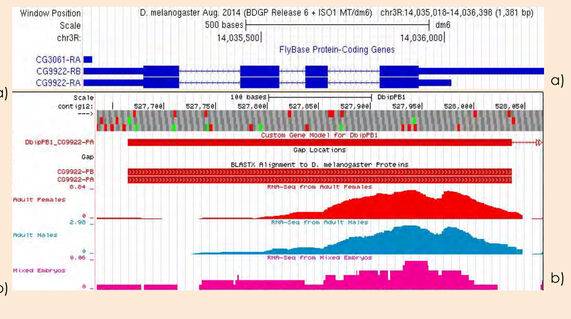
Figure 3(a) *D.ana Gat* exons #7 and #8 and **Figure 3(b)** *Ekar* exons #2 and #3. RNA-seq data is shown below, peaks corresponding to exons. Sharp drop in expression and gene predictor tracks are in favor of an additional intron in both genes. *BLAST* alignment shows the that the peptide sequence is preserved nonetheless.



- There are two coding sequences (Fig. 2(a)) within the *D. ana. Gat* gene in place of a single 7th CDS that produce a protein sequence with high percent similarity to the ortholog when joined during processing. Length of exon 7 and 8 of the predicted gene model is 125 and 245 bps, respectively, together producing a protein sequence of 123 aa, identical to the length of *D.mel.* exon 7 translated protein sequence of 123 aa.
- Similarly, an extra intron is found in the coding span of 2nd CDS of *Ekar* (Fig.2(b)), preserving the ortholog peptide length: exon 2 and 3 are 67 and 95 bps long, yielding a protein sequence of 54 aa, only one aa longer than the peptide product of the 2nd CDS in *D.mel.* (53 aa).

Retrotransposed pseudogene of *CG9922* has been found in *D. bipectinata*

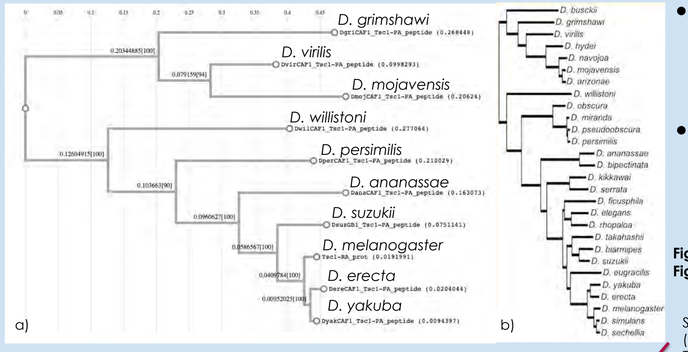
Figure 4(a): Gene structure of *CG9922 in D.mel (Gene Record Finder).*
Figure 4(b): Putative gene model of *CG9922* in *D. bip*, based on sequence similarity to *D.mel* protein and RNA-seq data. *(GEP Genome Browser).*



- *CG9922* is a multi-exon gene in *D. mel. (*4 CDS in both isoforms (Fig.4 (a)). However, *BLASTX* alignment track to *D. mel.* proteins shows no introns, but one continuous sequence alignment to the *D.bip.* genome assembly.
- Presence of a continuous exon in place of multiple is an indicator of a retrotransposed pseudogene, which has been reverse transcribed and inserted in the genomic region of contig12.
- The actual *CG9922* in *D.bip* is found on scaffold KB464125, April 2013 Assembly.

Results:

Full TSS and transcript annotations of *Tsc1* have been produced across nine *Drosophila* species: Synteny and gene structure are highly conserved



- Synteny has shown to be fully conserved in 8 of the 9 annotated species: *blast* search against D.mel proteins has found matches to *Sec10* and *Npc2F* downstream, and *GatB* and *Root* upstream of *Tsc1* ortholog (*D. suz.* shows only partial conservation).
- Gene structure is also highly-conserved: in 8 species (except *D. suz.*) all six CDS are supported by expression data and have been fine-mapped. All species show evidence supported TSS and 5' UTR as well.
- Phylogenetic tree based on the translated protein sequence of *Tsc1* across 9 species has been produced by *CLUSTALW*: evolutionary relationship based on the coding region reflects what has been hypothesized by GEP.

Figure 5(a): *CLUSTALW* phylogenetic tree based on *Tsc1* peptide similarity.
Figure 5(b): Phylogenetic tree used by GEP.

D. suzukii is likely missing the last CDS and 3'UTR

- There is lack of expression data available downstream of the 5th CDS of *Tsc1* ortholog. Significant match to *D.mel* protein has not been found around this region either, hence the last CDS is absent in *D.suz.*
- Significant sequence similarity and RNA-seq for the previous exons still suggests that gene is expressed and is able to function without the 6th CDS and 3' UTR.

Shown in blue is the last (5th) CDS of putative *Tsc1* ortholog

RNA-seq data decreases gradually. No new peaks or splice junctions indicate at the 6th CDS

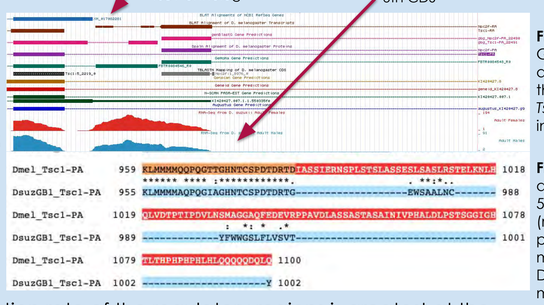


Figure 6: Genomic region downstream of the 5th CDS of *Tsc1-PA* ortholog in *D. suz.*

Figure 7: Protein alignment of the 5th CDS of D. mel (red) against the putative gene model (blue). Dashes indicate missing amino acids.

Conclusion: *Pathways Project* predicts that the evolution rate of the regulatory regions is greatest at the top of the insulin pathway, while genes with many interacting partners are more constrained. *Tsc1* is under high evolutionary constraint, due to its significance within a pathway, which is reflected in gene structure and synteny conservation among 9 *Drosophila* species.

References:

- Leung W, Shaffer CD, Reed LK, et al. *Drosophila* Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution. *G3 (Bethesda)*. 2015;5(5):719-740. Published 2015 Mar 4. doi:10.1534/g3.114.015966
- Reed LK, *Pathways Project: Analyzing evolution of metabolic and signaling pathway genes.* Genomics Education Partnership. May 7, 2021. https://www.youtube.com/watch?v=gLXoT2oXEz8&t=2s
- GEP website: https://thegep.org/

Acknowledgements:
I would like to express my many thanks to Professor Kennell for her support and guidance throughout this project. Despite being fully remote, she managed to bring in enthusiasm and sense of teamwork in our weekly lab meetings. I would also like to thank Trina Chou and Veronica Gomes for being such great lab partners who helped me through challenges of annotation. Finally, I would like to express my gratitude to Mrs. Florence A. Davis for funding this research opportunity and allowing me to pursue my interest in scientific investigation.