Gene model for the ortholog of *Ilp4* in *Drosophila grimshawi*

Rachael Cowan[1], Alyssa Koehler[1], Clairine I. S. Larsen[2], Jeffrey S. Thompson[2], Andrew M Arsham[3], Lori Boies[4a]

1. The University of Alabama, Tuscaloosa, AL, USA
2. Denison University, Granville, OH, USA
3. Saint Mary's College of California, Moraga, CA, USA
4. St. Mary's University, San Antonio, TX, USA
    1. To whom correspondence should be addressed

Abstract

Gene Model for the *Insulin-like peptide 4 (Ilp4)* ortholog in the *D. grimshawi (dgriCAF1*) assembly (GeneBank Accession: GCA_000005155.1) of *Drosophila grimshawi*. The characterization of this ortholog was conducted as part of a broader, developing dataset aimed at investigating the evolution of the Insulin/insulin-like growth factor signaling (IIS) pathway across the genus *Drosophila*, using the Genomics Education Partnership gene annotation protocol within Course-based Undergraduate Research Experiences.

Introduction

> *This article reports a predicted gene model generated by undergraduate work using a structured gene model annotation protocol defined by the Genomics Education Partnership (GEP; thegep.org) for Course-based Undergraduate Research Experience (CURE). The following information in quotes may be repeated in other articles submitted by participants using the same GEP CURE protocol for annotating Drosophila species orthologs of Drosophila melanogaster genes in the insulin signaling pathway.*

The insulin signaling pathway is well characterized in *Drosophila* and is highly conserved across metazoans. Owing to this strong evolutionary conservation, *Drosophila* serves as a valuable model for studying insulin signaling in humans (Biglou, *et. al.* 2021). The *insulin-like peptide 4 (Ilp4)* gene has been shown to have high expression within the mesoderm of the embryo and the midgut in the larval stage (Brogiolo, *et. al.,* 2001). Of the eight insulin-like peptides within the *Drosophila* insulin signaling pathway, *Ilp4 1 - 7* are highly conserved across species, with *Ilp4 and Ilp7* being the most conserved (Grönke, *et. al.*, 2010). While highly conserved, there is scant evidence of the specific function of *Ilp4* within the scientific literature. The model presented here is the ortholog of *Ilp4* in the dgriCAF1 assembly of *D. grimshawi* (GCA_000005155.1 - Clark et al. 2007) and corresponds to the XP_032590595.1 predicted

model *D. grimshawi* (LOC6556940). This gene model is based on RNA-seq data from *D. grimshaw*i (SRP073087) in *D. melanogaster* from *FB2022_02* (Larkin *et al.,* 2021).

Drosophila grimshawi (NCBI:txid7222) belongs to the Picture Wing group of Hawaiian *Drosophila*, as defined by Kaneshiro and colleagues (1995). Phylogenetic studies indicate that the Hawaiian *Drosophila* form a monophyletic lineage that is sister to the *Scaptomyza* clade, with both lineages situated within the *Drosophila* subgenus of the genus *Drosophila* (Kambysellis et al., 1995; Baker & DeSalle, 1999). The group derives its name from the striking pigmentation patterns on their wings. *D. grimshawi*, first described by Oldenberg in 1914, inhabits high-elevation, cool tropical rainforests on the Maui Complex islands, where it reproduces on decaying plant material (Carson et al., 1970; Carson, 1983)

"In this GEP CURE protocol students use web-based tools to manually annotate genes in non-model Drosophila species based on orthology to genes in the well-annotated model organism, the fruit fly Drosophila melanogaster. This allows undergraduates to participate in course-based research by generating manual annotations of genes in non-model species (Rele et al., 2023). Computational-based gene predictions in any organism are often improved by careful manual annotation and curation, allowing for more accurate analyses of gene and genome evolution (Mudge and Harrow, 2016; Tello-Ruiz et al., 2019). These models of orthologous genes across species, such as the one presented here, then provide a reliable basis for further evolutionary genomic analyses when made available to the scientific community." (Myers et al., 2024).

Results

Synteny

Ilp4 occurs on *chromosome 3L* in *D. melanogaster* and is flanked by *I-2, CG43897* with *Ilp5* nested within*, CG32052* with *Ilp3* and *Ilp2* nested within*, Ilp4, Ilp1, Zasp67,* and *Ae2.* We determined that the putative ortholog of *Ilp4* is found on scaffold CH916366 in *D. grimshawi* at LOC6556940 (via *tblastn* search with an e-value of 8e-09 and percent identify of 37.04%), where it is surrounded by genes with the following locus IDs: *LOC6556854, LOC6556919* with *LOC6557001* nested within*, LOC6557056* with *LOC6556947, LOC6556893*, and *LOC6556922* nested within*, LOC6556876, LOC6556918,* and *LOC6556946.* which corresponds to *I-2, CG43897* with *Ilp5* nested within*, CG32052* with *Ilp3* and two *Ilp2* paralogs nested within*, Ilp1, Zasp67,* and *Ae2 in D. melanogaster* as determined by *blastp* (Figure 1A, Altschul et al., 1990). We believe this is the correct ortholog assignment for *Ilp4* in *D. grimshawi* because local synteny is conserved with the exception of there being a duplication event resulting in two paralogs of *Ilp2*.

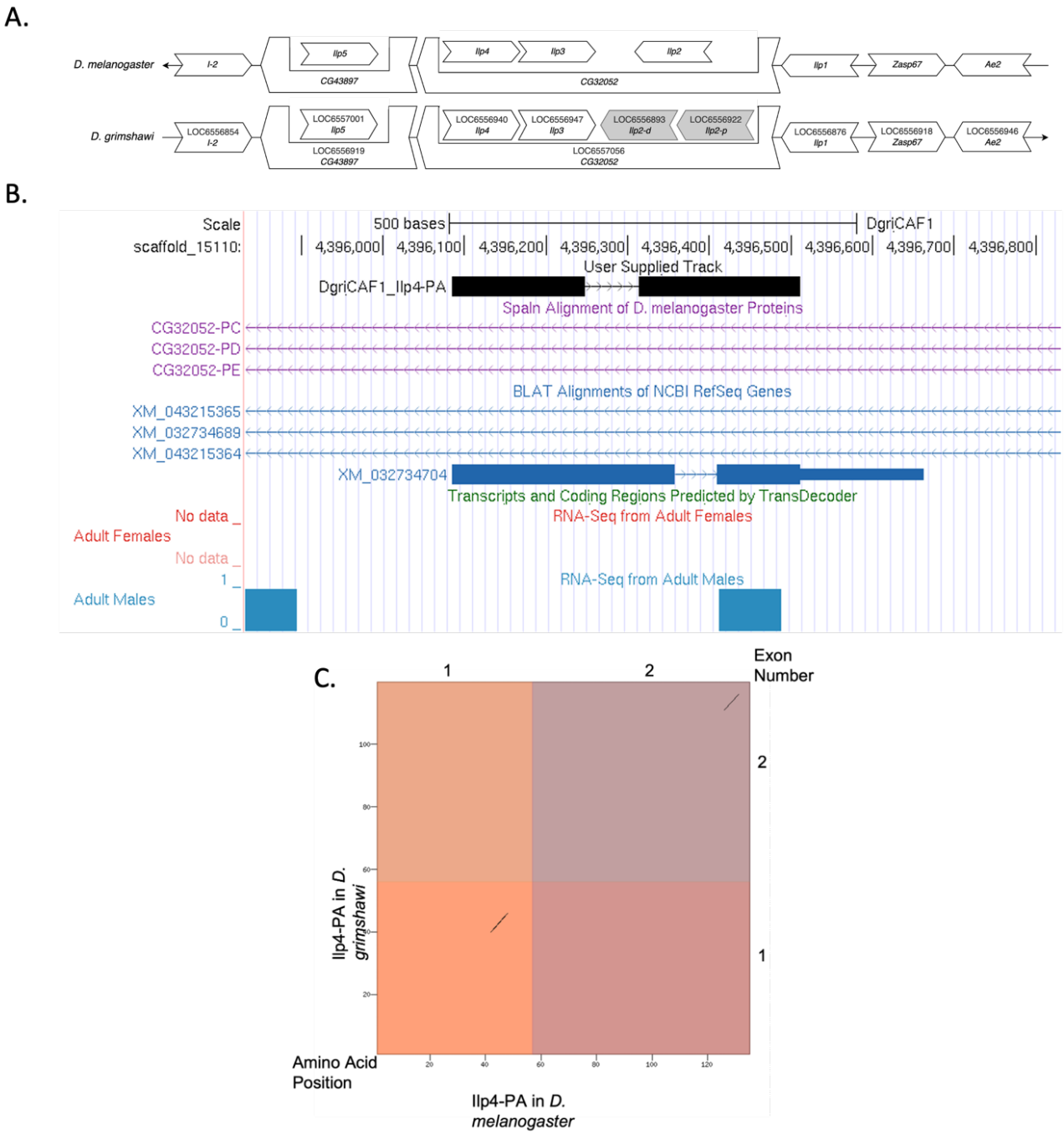


Figure 1: *Ilp4* gene model comparison between *Drosophila grimshawi* and *Drosophila melanogaster.*

(A) Local synteny of genomic neighborhood of *Ilp4* in *D. melanogaster* and *D. grimshawi*. Gene arrows pointing in the same direction as the target gene *(Ilp4)* in both *D. grimshawi* and *D. melanogaster* are on the same strand as the target gene; gene arrows pointing in the opposite direction are on the opposite strand. The thin underlying arrows pointing to the right indicate that *Ilp4* is on the + strand; arrows pointing to the

left indicate that *Ilp4* is on the – strand. White arrows in *D. grimshawi* indicate the locus ID and the orthology to the corresponding gene in *D. melanogaster*. Gray arrows indicate paralogs of a duplication of downstream gene *Ilp2* that is not present in *D. melanogaster*. The gene names given in the *D. gimshawi* gene arrows indicate the orthologous gene in *D. melanogaster*, while the locus identifiers are specific to *D. grimshawi*. **(B) Gene Model in UCSC Track Hub (Raney et al. 2014):** the gene model in *D. grimshawi* (black), Spaln of D. melanogaster Proteins (purple, alignment of refseq proteins from *D. melanogaster*), BLAT alignments of NCBI RefSeq Genes (blue, alignment of refseq genes for *D. grimshawi*), RNA-Seq from Adult males (blue) and adult females (red), alignment of Illumina RNAseq reads from *D. grimshawi*), and Transcripts (green) including coding regions predicted by TransDecoder and Splice Junctions Predicted by regtools using *D. grimshawi* RNA-Seq (SRP073087). The custom gene model (User Supplied Track) is indicated in black with exon depicted with wide boxes, intron with narrow lines (arrows indicate direction of transcription). **(C) Dot Plot of Ilp4-PA in *D. melanogaster* (*x*-axis) vs. the orthologous peptide in *D. grimshawi* (*y*-axis).** Amino acid number is indicated along the left and bottom; exon number is indicated along the top and right, and exons are also highlighted with alternating colors. There are large gaps in the dot plot which indicates dissimilarity in the amino acid sequence between Ilp4-PA in *D. melanogaster* and *D. grimshawi*.

Protein Model

Ilp4 in *D. grimshawi* has 1 protein coding isoform (Ilp4-PA) (Figure 1B). Ilp4-PA contains 2 protein coding exons. *Ilp4* in *D. melanogaster* also has one isoform (Ilp4-PA) which also has two protein coding exons. The sequence of Ilp4-PA in *D. grimshawi* has 46.6% identity with the *Ilp4* in *D. melanogaster* as determined by *blastp* (Figure 1C). These data are also available in Extended Data files below, which are archived in CaltechData.

Special characteristics of the protein model

Lack of Splice Junctions: There is no splice junction evidence track due to the relative lack of RNA-seq reads mapping to this region (Figure 1B). This may be because multiple *Ilp* paralogs with similar sequences impede accurate mapping of RNA-seq reads.

Low Sequence Similarity: There is low sequence similarity between *Ilp4* in *D. melanogaster* and *D. grimshawi* shown by *blastp* (46.6% identity), *tblastn* (37.04% identity) and the dot plot showing the protein alignment (Figure 1C). This is likely because *D. grimshawi* is highly diverged from *D. melanogaster.*

Methods

"Detailed methods including algorithms, database versions, and citations for the complete annotation process can be found in Rele et al. (2023). Briefly, students use the GEP instance of the UCSC Genome Browser v.435 (https://gander.wustl.edu; Kent WJ et al., 2002; Navarro Gonzalez et al., 2021) to examine the genomic neighborhood of their reference IIS gene in the *D. melanogaster* genome assembly (Aug. 2014; BDGP Release 6 + ISO1 MT/dm6). Students then retrieve the protein sequence for the *D. melanogaster* reference gene for a given isoform and run it using *tblastn* against their target *Drosophila* species genome assembly on the NCBI BLAST server (https://blast.ncbi.nlm.nih.gov/Blast.cgi; Altschul et al., 1990) to identify potential orthologs. To validate the potential ortholog, students compare the local genomic neighborhood of their potential ortholog with the genomic neighborhood of their reference gene in *D. melanogaster*. This local synteny analysis includes at minimum the two upstream and downstream genes relative to their putative ortholog. They also explore other sets of genomic evidence using multiple alignment tracks in the Genome Browser, including BLAT alignments of RefSeq Genes, Spaln alignment of *D. melanogaster* proteins, multiple gene prediction tracks (e.g., GeMoMa, Geneid, Augustus), and modENCODE RNA-Seq from the target species. Detailed explanation of how these lines of genomic evidenced are leveraged by students in gene model development are described in Rele et al. (2023). Genomic structure information (e.g., CDSs, intron-exon number and boundaries, number of isoforms) for the *D. melanogaster* reference gene is retrieved through the Gene Record Finder (https://gander.wustl.edu/~wilson/dmelgenerecord/index.html; Rele et al*.,* 2023). Approximate splice sites within the target gene are determined using *tblastn* using the CDSs from the *D. melanogaste*r reference gene. Coordinates of CDSs are then refined by examining aligned modENCODE RNA-Seq data, and by applying paradigms of molecular biology such as identifying canonical splice site sequences and ensuring the maintenance of an open reading frame across hypothesized splice sites. Students then confirm the biological validity of their target gene model using the Gene Model Checker (https://gander.wustl.edu/~wilson/dmelgenerecord/index.html; Rele et al., 2023), which compares the structure and translated sequence from their hypothesized target gene model against the *D. melanogaster* reference gene model. At least two independent models for a gene are generated by students under mentorship of their faculty course instructors. Those models are then reconciled by a third independent researcher mentored by the project leaders to produce the final model. Note: comparison of 5' and 3' UTR sequence information is not included in this GEP CURE protocol." (Gruys et al, 2025)

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol*. 215:403-410. PMID: 2231712

Baker RH, DeSalle R. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Systems Biology* 46(4): 654-73.

Biglou, S.G., W.G. Bendena, and I. Chin-Sang, An overview of the insulin signaling pathway in model organisms Drosophila melanogaster and Caenorhabditis elegans. *Peptides*, 2021. 145: p. 170640. PMID: 34450203

Brogiolo, W., Stocker, H., Ikeya, T., Rintelen, F., Fernandez, R. and Hafen, E., 2001. An evolutionarily conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Current Biology* 11, 213-221. PMID: 11250149

Carson HL. 1983. Chromosomal sequences and interisland colonizations in hawaiian Drosophila. *Genetics* 103(3): 465-82.

Carson HL, Hardy DE, Spieth HT, Stone WS. 1970. The Evolutionary Biology of the Hawaiian Drosophilidae. Essays in Evolution and Genetics in Honor of Theodosius Dobzhansky : 437-543

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuente, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., …

MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature, 450(7167), 203–218.
https://doi.org/10.1038/nature06341. PMID: 17994087

Grönke, S., Clarke, D.-F., Broughton, S., Andrews, T.D. and Partridge, L., 2010. Molecular Evolution and Functional Characterization of *Drosophila* Insulin-Like Peptides. *PLOS Genetics* 6, e1000857. PMID: 20195512

Gruys ML, Sharp MA, Lill Z, Xiong C, Hark AT, Youngblom JJ, Rele CP, Reed LK. 2025. Gene model for the ortholog of glys in drosophila simulans. *MicroPubl Biol.* 025:10.17912/micropub.biology.001168. eCollection 2025. doi: 10.17912/micropub.biology.001168. PMID: 39845267.

Kambysellis MP, Ho KF, Craddock EM, Piano F, Parisi M, Cohen J. 1995. Pattern of ecological shifts in the diversification of Hawaiian Drosophila inferred from a molecular phylogeny. *Current Biology* 5(10): 1129-39.

Kaneshiro, KY, RG Gillespie, and HL Carson. "Chromosomes and male genitalia of Hawaiian Drosophila: tools for interpreting phylogeny and geography." *Hawaiian biogeography: evolution on a hot spot archipelago* (1995): 57-71

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. TheHuman Genome Browser at UCSC. *Genome Res*. 12:996-1006. DOI: 10.1101/gr.229102

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* 17:758-772. DOI: 10.1038/nrg.2016.119

Myers A, Hoffman A, Natysin M, Arsham AM, Stamm J, Thompson JS, Rele CP, Reed LK. Gene model for the ortholog Myc in Drosophila ananassae. 2024. MicroPubl Biol. DOI: 10.17912/micropub.biology.000856

Navarro Gonzalez J, Zweig AS, Speir, ML, Schmelter D, Rosenbloom KR, Raney BJ, et al., 2021. The UCSC Genome Browser database: 2021 update. Nucl Acids Res 49:1046-1057. DOI: 10.1093/nar/gkaa107

Oldenberg, L. 1914. Beitrag zur Kenntnis der europäische n Drosophiliden (Dipt.). Archiv für Naturgeschichte (A) 80 (2): 1-42. [1914.05.18]

Rele CP, Sandlin KM, Leung W, Reed LK. 2023. Manual annotation of Drosophila genes: a Genomics Education Partnership protocol. *F1000Research* 11: 1579. DOI: 10.12688/f1000research.126839.2

Tello-Ruiz MK, Marco CF, Hsu FM, Khangura RS, Qiao P, Sapkota S, et al. 2019. Double triage to identify poorly annotated genes in maize: The missing link in community curation. *PLoS One* 14:e0224086. DOI: 10.1371/journal.pone.0224086

Supplemental Material

1. Zip file containing FASTA, PEP, GFF files for the gene model
2. Figure 1 in high resolution

Metadata: Bioinformatics, Genomics, *Drosophila*, Genotype Data, New Finding

Acknowledgements

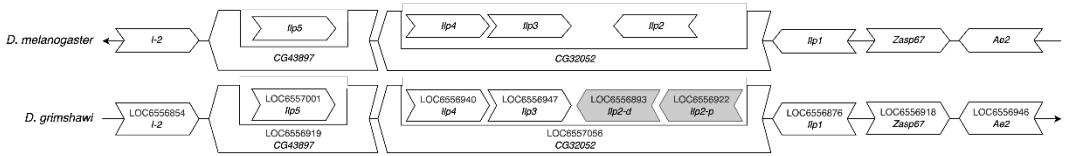
We thank Wilson Leung (Washington University, St. Louis) for developing and maintaining the technological infrastructure that supported the creation of this gene model, as well as Chinmay Rele and Laura K. Reed (University of Alabama) for their guidance and encouragement throughout the project. We are also grateful to FlyBase for providing the authoritative database for *Drosophila melanogaster* gene models. FlyBase is supported by grants NHGRI U41HG000739

and U24HG010859, UK Medical Research Council MR/W024233/1, NSF 2035515 and 2039324, BBSRC BB/T014008/1, and Wellcome Trust PLM13398.

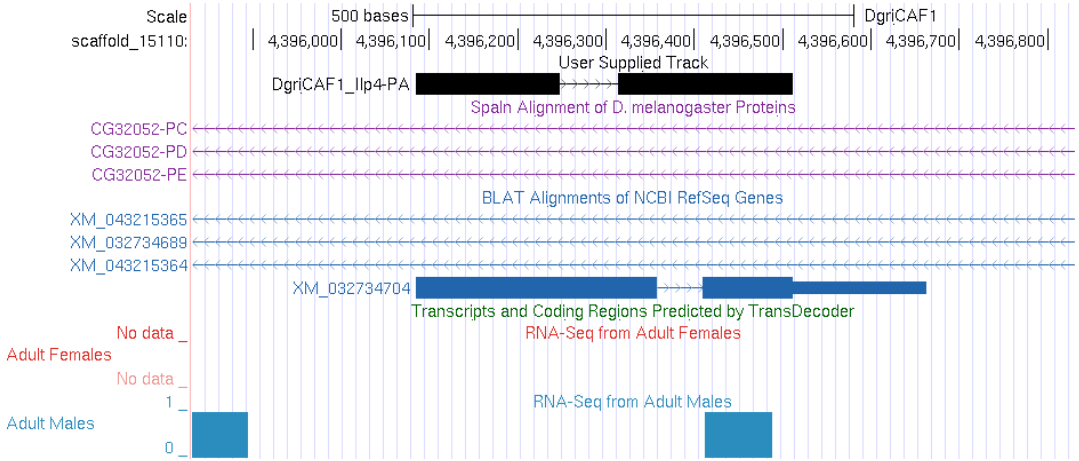
Funding

This material is based upon work supported by the National Science Foundation (1915544) and the National Institute of General Medical Sciences of the National Institutes of Health (R25GM130517) to the Genomics Education Partnership (GEP; https://thegep.org/; PI-Laura K. Reed). Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the author(s) and do not necessarily reflect the official views of the National Science Foundation nor the National Institutes of Health.

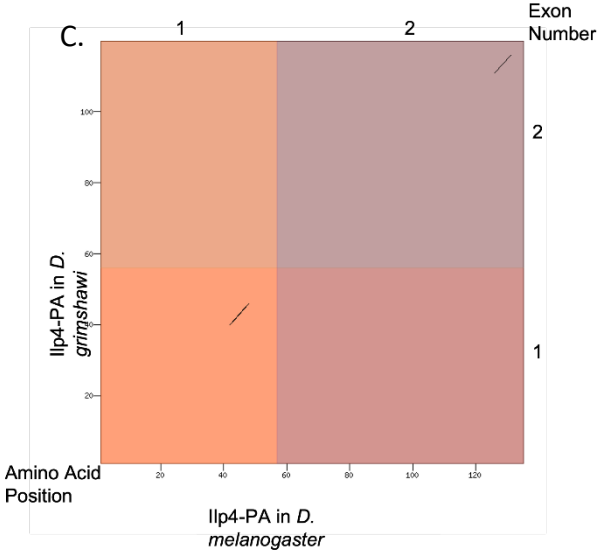
A.


D. melanogaster
I-2
Ilp5
CG43897
Ilp4
Ilp3
Ilp2
CG32052
Ilp1
Zasp67
Ae2
D. grimshawi
LOC6556854
I-2
LOC6557001
Ilp5
LOC6556919
CG43897
LOC6556940
Ilp4
LOC6556947
Ilp3
LOC6556893
Ilp2-d
LOC6556922
Ilp2-p
LOC6557056
CG32052
LOC6556876
Ilp1
LOC6556918
Zasp67
LOC6556946
Ae2

B.


Scale
500 bases
DgriCAF1
scaffold_15110:
4,396,000
4,396,100
4,396,200
4,396,300
4,396,400
4,396,500
4,396,600
4,396,700
4,396,800
User Supplied Track
DgriCAF1_Ilp4-PA
Spaln Alignment of D. melanogaster Proteins
CG32052-PC
CG32052-PD
CG32052-PE
BLAT Alignments of NCBI RefSeq Genes
XM_043215365
XM_032734689
XM_043215364
XM_032734704
Transcripts and Coding Regions Predicted by TransDecoder
No data
Adult Females
RNA-Seq from Adult Females
No data
1
Adult Males
RNA-Seq from Adult Males
0

C.


1
2
Exon Number
2
1
Ilp4-PA in D. grimshawi
Amino Acid Position
Ilp4-PA in D. melanogaster