1/29/2025 - *Open Access*

Gene model for the ortholog of *eIF4E1* in *Drosophila yakuba*

Bailey Lose[1], Jeremy Girard[2], Josephine Hayes[3], Lane Weast[3], Natalie Minkovsky[4], Sarah Justice[5], Jack A. Vincent[6], James J. Youngblom[2], Lindsey J. Long[3], Chinmay P. Rele[1], Laura K Reed[1§]

[1]University of Alabama, Tuscaloosa, AL USA

[2]California State University Stanislaus, Turlock, CA USA

[3]Oklahoma Christian University, Edmond, OK USA

[4]Community College of Baltimore County, Baltimore, MD USA

[5]Taylor University, Upland, IN USA

[6]University of Washington - Tacoma, Tacoma, WA USA

[§]To whom correspondence should be addressed: lreed1@ua.edu

Abstract

Gene model for the ortholog of *eukaryotic translation initiation factor 4E1* (*eIF4E1*) in the Dyak_CAF1 Genome Assembly (GenBank Accession: GCA_000005975.1) of *Drosophila yakuba*. This ortholog was characterized as part of a developing dataset to study the evolution of the Insulin/insulin-like growth factor signaling pathway (IIS) across the genus *Drosophila* using the Genomics Education Partnership gene annotation protocol for Course-based Undergraduate Research Experiences.

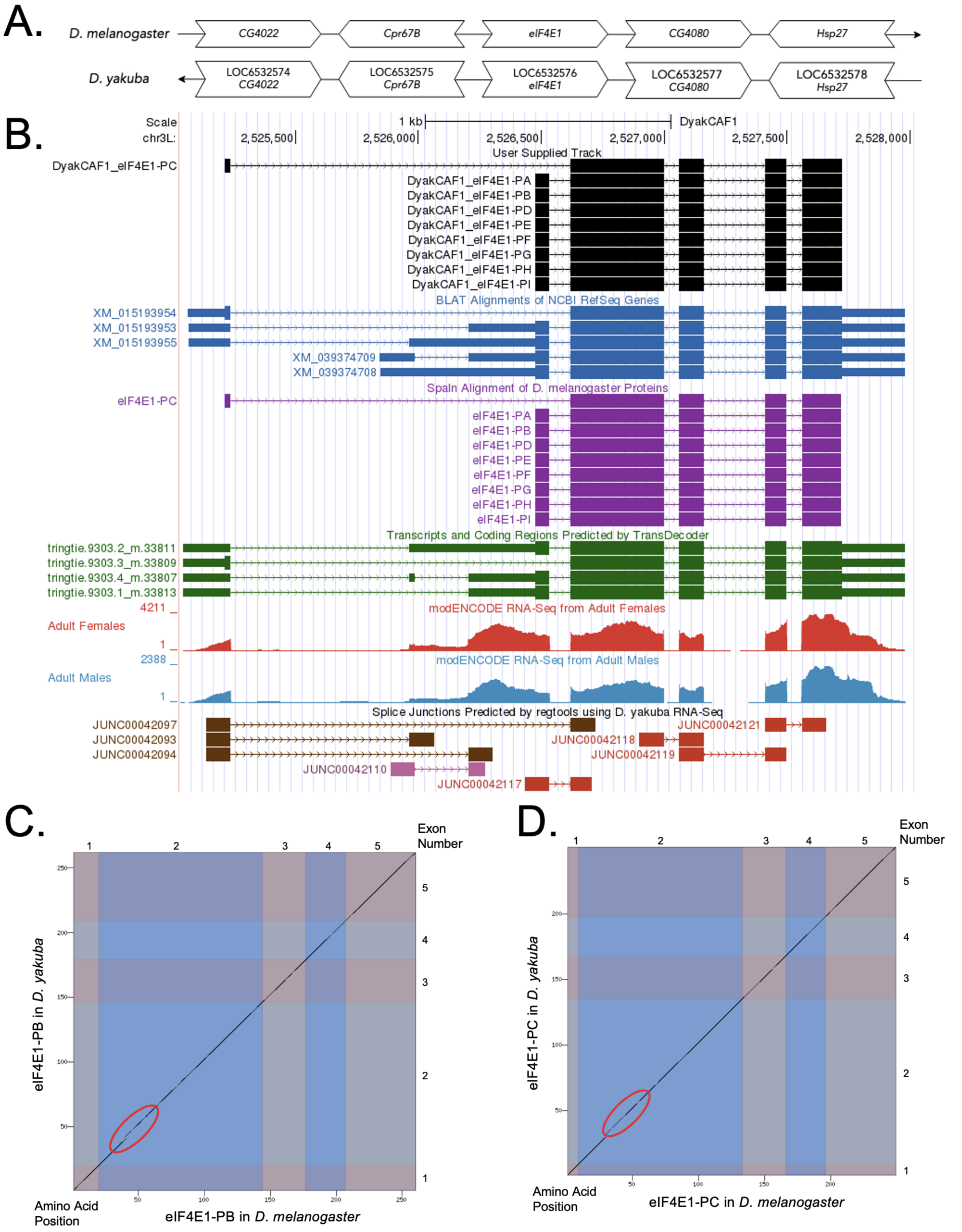

Figure 1. Genomic neighborhood and gene model for *eIF4E1* in *D. yakuba*:

(A) Synteny comparison of the genomic neighborhoods for *eIF4E1* in *Drosophila melanogaster* and *D. yakuba*. Thin underlying arrows indicate the DNA strand within which the reference gene–*eIF4E1*–is located in *D. melanogaster* (top) and *D. yakuba* (bottom) genomes. The thin arrow pointing to the right indicates that *eIF4E1* is on the positive (+) strand in *D. melanogaster*, and the thin arrow pointing to the left indicates that *eIF4E1* is on the negative (-) strand in *D. yakuba* The wide gene arrows pointing in the same direction as *eIF4E1* are on the same strand relative to the thin underlying arrows, while wide gene arrows pointing in the opposite direction of *eIF4E1* are on the opposite strand relative to the thin underlying arrows. White gene arrows in *D. yakuba* indicate orthology to the corresponding gene in *D. melanogaster*. Gene symbols given in the *D. yakuba* gene arrows indicate the orthologous gene in *D. melanogaster*, while the locus identifiers are specific to *D. yakuba*. **(B) Gene Model in GEP UCSC Track Data Hub (Raney et al., 2014).** The coding-regions of *eIF4E1* in *D. yakuba* are displayed in the User Supplied Track (black); CDS are depicted by thick rectangles and introns by thin lines with arrows indicating the direction of transcription. Subsequent evidence tracks include BLAT Alignments of NCBI RefSeq Genes (dark blue, alignment of Ref-Seq genes for *D. yakuba*), Spaln of *D. melanogaster* Proteins (purple, alignment of Ref-Seq proteins from *D. melanogaster*), Transcripts and Coding Regions Predicted by TransDecoder (dark green), RNA-Seq from Adult Females and Adult Males (red and light blue, respectively; alignment of Illumina RNA-Seq reads from *D. yakuba*), and Splice Junctions Predicted by regtools using *D. yakuba* RNA-Seq (SRP006203 - Graveley et al, 2010). Splice junctions shown have a read-depth of 232, 500-999 and >1000 supporting reads in pink, brown and red, respectively. **(C) Dot Plot of eIF4E1-PB in *D. melanogaster* (*x*-axis) vs. the orthologous peptide in *D. yakuba* (*y*-axis).** Amino acid number is indicated along the left and bottom; CDS number is indicated along the top and right, and CDS are also highlighted with alternating colors. A region with a reduced sequence similarity is circled in red. **(D) Dot Plot of eIF4E1-PC in *D. melanogaster* (x-axis) vs. the orthologous peptide in *D. yakuba* (y-axis).** A region with reduced sequence similarity is circled in red.

Description

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 this box 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.

"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 fruitfly *Drosophila melanogaster*. The GEP uses web-based tools to allow 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).

"The particular gene ortholog described here was characterized as part of a developing dataset to study the evolution of the Insulin/insulin-like growth factor signaling pathway (IIS) across the genus *Drosophila*. The Insulin/insulin-like growth factor signaling pathway (IIS) is a highly conserved signaling pathway in animals and is central to mediating organismal responses to nutrients (Hietakangas and Cohen 2009; Grewal 2009)." (Myers et al., 2024).

"*D. yakuba (*NCBI:txid7245) is part of the *melanogaster* species group within the subgenus *Sophophora* of the genus *Drosophila* (Sturtevant 1939; Bock and Wheeler 1972). It was first described by Burla (1954). *D. yakuba* is wide-spread in sub-Saharan Africa and Madagascar (Lemeunier et al., 1986; https://www.taxodros.uzh.ch, accessed 1 Feb 2023; Markow and O'Grady 2005) where figs served as a primary host along with other rotting fruits (Lachaise and Tsacas 1983)." (Koehler et al., 2024).

We propose a gene model for the *D. yakuba* ortholog of the *D. melanogaster* eukaryotic translation initiation factor 4E1 (*eIF4E1*) gene. The genomic region of the ortholog corresponds to the uncharacterized protein LOC6532576 (RefSeq accession XP_015049441.1) in the Dyak_CAF1 Genome Assembly of *D. yakuba* (GenBank Accession: GCA_000005975.1; *Drosophila* 12 Genomes Consortium et al., 2007). This model is based on RNA-Seq data from *D. yakuba* (SRP006203; Graveley et al., 2011) and *eIF4E1* in *D. melanogaster* using FlyBase release FB2022_04 (GCA_000001215.4; Larkin et al., 2021; Gramates et al., 2022; Jenkins et al., 2022).

eukaryotic translation initiation factor 4E1 (*eIF4E1*) encodes eIF4F cap-binding complex essential cap-dependent translation of mRNA, and binds the 7-methyl-guanosine cap structure of mRNA in *Drosophila* (Lachance et al., 2002; Lavoie et al., 1996). The protein product of *eIF4E-3*, a paralog of *eIF4E1*, is specifically required during spermatogenesis in *Drosophila* (Hernendez et al., 2012).

Synteny

The reference gene, *eIF4E1*, occurs on chromosome 3L in *D. melanogaster* and is flanked upstream by *CG4022* and *Cuticular protein 67B* (*Cpr67b*) and downstream by *CG4080* and *Heat shock protein 27* (*Hsp27*). The *tblastn* search of *D. melanogaster* eIF4E1-PB (query) against the *D. yakuba* (GenBank Accession: GCA_000005975.1) Genome Assembly (database) placed the putative ortholog of *eIF4E1* within scaffold chromosome 3L (CM000159.2) at locus LOC6532576 (XP_015049441.1) with an E-value of 1e-77 and a percent identity of 65.56%. Furthermore, the putative ortholog is flanked upstream by LOC6532574 (XP_015049438.1) and LOC6532575 (XP_002093319.1) which correspond to *CG4022* and *Cpr67b* in *D. melanogaster* (E-value: 0.0 and 7e-170; identity: 90.08% and 98.46%, respectively, as determined by *blastp*; Figure 1A, Altschul et al., 1990). The putative ortholog *eIF4E1* is flanked downstream by LOC6532577 (XP_015049442.1) and LOC6532578 (XP_002093322.1) which correspond to *CG4080* and *Hsp27* in *D. melanogaster* (E-value: 0.0 and 2e-132; identity: 96.63% and 89.72%, respectively, as determined by *blastp*). The putative ortholog assignment for *eIF4E1* in *D. yakuba* is supported by the following evidence: The genes surrounding the *eIF4E1* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is completely conserved, supported by e-values and percent identities, so we conclude that LOC6532576 is the correct ortholog of *eIF4E1* in *D. yakuba* (Figure 1A).

Protein Model

eIF4E1 in *D. yakuba* has two unique protein-coding isoforms eIF4E1-PB (identical to eIF4E1-PA, eIF4E1-PD, eIF4E1-PE, eIF4E1-PF, eIF4E1-PG, eIF4E1-PH, eIF4E1-PI) and eIF4E1-PC (Figure 1B). mRNA isoforms *eIF4E1-RB* (*eIF4E1-RA*, *eIF4E1-RD*, *eIF4E1-RE*, *eIF4E1-RF*, *eIF4E1-RG*, *eIF4E1-RH*, *eIF4E1-RI*) and *eIF4E1-RC* contain five CDSs. Relative to the ortholog in *D. melanogaster*, the RNA CDS number and protein isoform count is conserved. The sequence of eIF4E1-PB in *D. yakuba* has 93.49% identity (E-value: 1e-142) with the protein-coding isoform eIF4E1-PB in *D. melanogaster*, as determined by *blastp* (Figure 1C). Minor gaps in the dot plots of eIF4E1-PB (Figure 1C) and eIF4E1-PC (Figure 1D) represent a region of lower sequence similarity, highlighted by red circles, including a short indel of two amino acids in the second exon of both isoforms. Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (accession **BK059542**, **BK059543**, **BK059544**, **BK059545**, **BK059546**, **BK059547**, **BK059548**, **BK059549** and **BK059550**). These data are also archived in the CaltechDATA repository (see "Extended Data" section below).

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 this gene were generated by students under mentorship of their faculty course instructors. These models were then reconciled by a third independent researcher mentored by the project leaders to produce the final model presented here. Note: comparison of 5' and 3' UTR sequence information is not included in this GEP CURE protocol.

Acknowledgements: This publication is dedicated to the memory of Dr. James J. Youngblom. We would like to thank Wilson Leung for developing and maintaining the technological infrastructure that was used to create this gene model, Cole A. Kiser and Leon F. Laskowski for their contributions in reviewing this model, Madeline L. Gruys for retrofitting this model. Thank you to FlyBase for providing the definitive 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.

Extended Data

Description: A GFF, FASTA, and PEP of the model. Resource Type: Model. File: DyakCAF1_eIF4E1.zip. DOI: 10.22002/hx6n6-een12

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. J Mol Biol. 215: 403. PubMed ID: 2231712

Bock IR, Wheeler MR. 1972. The *Drosophila melanogaster* species group. Univ. Texas Publs Stud. Genet. 7(7213): 1.

Burla H. 1954. Zur Kenntnis der Drosophiliden der Elfenbeinkuste (Franzosisch West-Afrika). Revue suisse Zool. 61(Suppl.): 1.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al., MacCallum I. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450(7167): 203-18. PubMed ID: 17994087

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, et al., the FlyBase Consortium. 2022. FlyBase: a guided tour of highlighted features. Genetics 220(4). PubMed ID: 35266522

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al., Celniker SE. 2011. The developmental transcriptome of *Drosophila melanogaster.* Nature 471(7339): 473-9. PubMed ID: 21179090

Grewal SS 2009. Insulin/TOR signaling in growth and homeostasis: a view from the fly world. Int. J. Biochem. Cell Biol.. 41: 1006. PubMed ID: 18992839

Hernández G, Han H, Gandin V, Fabian L, Ferreira T, Zuberek J, et al., Lasko P. 2012. Eukaryotic initiation factor 4E-3 is essential for meiotic chromosome segregation, cytokinesis and male fertility in *Drosophila*. Development 139(17): 3211-20. PubMed ID: 22833128

Hietakangas V, Cohen SM. 2009. Regulation of tissue growth through nutrient sensing. Annu Rev Genet 43: 389-410. PubMed ID: 19694515

Jenkins VK, Larkin A, Thurmond J, FlyBase Consortium. 2022. Using FlyBase: A Database of *Drosophila* Genes and Genetics. Methods Mol Biol 2540: 1-34. PubMed ID: 35980571

Kent, W James, Sugnet, Charles W, Furey, Terrence S, Roskin, Krishna M, Pringle, Tom H, Zahler, Alan M, Haussler, David 2002. The human genome browser at UCSC. Genome Res.. 12: 996. PubMed ID: 12045153

Koehler AC, Cohen L, Romo I, Le V, Youngblom JJ, Hark AT, Rele CP, Reed LK. 2024. Gene model for the ortholog of Glys in *Drosophila yakuba*. MicroPubl Biol 2024. PubMed ID: 39758580

Lachaise D, Tsacas L. 1983. Breeding-sites of tropical African *Drosophilids*. Ashburner, Carson, Thompson, 1981-1986. 3d: 21.

Lachance PE, Miron M, Raught B, Sonenberg N, Lasko P. 2002. Phosphorylation of eukaryotic translation initiation factor 4E is critical for growth. Mol Cell Biol 22(6): 1656-63. PubMed ID: 11865045

Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, et al., FlyBase Consortium. 2021. FlyBase: updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Res 49(D1): D899-D907. PubMed ID: 33219682

Lavoie CA, Lachance PE, Sonenberg N, Lasko P. 1996. Alternatively spliced transcripts from the *Drosophila* eIF4E gene produce two different Cap-binding proteins. J. Biol. Chem.. 271: 16393. PubMed ID: 8663200

Lemeunier F, David J, Tsacas L, Ashburner M. 1986. The *melanogaster* species group. Ashburner, Carson, Thompson, 1981-1986. e: 147.

Markow TA, O'Grady P. 2005. *Drosophila*: A guide to species identification and use. Academic Press 978-0-12-473052-6

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. Nat. Rev. Genet. 17: 758. PubMed ID: 27773922

Myers A, Hoffman A, Natysin M, Arsham AM, Stamm J, Thompson JS, Rele CP, Reed LK. 2024. Gene model for the ortholog Myc in *Drosophila ananassae*. MicroPubl Biol 2024. PubMed ID: 39677519

Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al., Kent WJ. 2021. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res 49(D1): D1046-D1057. PubMed ID: 33221922

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al., Kent WJ. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30(7): 1003-5. PubMed ID: 24227676

Rele CP, Sandlin KM, Leung W, Reed LK. 2023. Manual annotation of *Drosophila* genes: a Genomics Education Partnership protocol [version 2; peer review: 2 approved with reservations]. F1000Research. 11 DOI: 10.12688/f1000research.126839.2

Sturtevant AH. 1939. On the Subdivision of the Genus *Drosophila*. Proc Natl Acad Sci U S A. 1939 Mar;25(3): 137. PubMed ID: 16577879

Tello-Ruiz MK, Marco CF, Hsu FM, Khangura RS, Qiao P, Sapkota S, et al., Micklos DA. 2019. Double triage to identify poorly annotated genes in maize: The missing link in community curation. PLoS One 14(10): e0224086. PubMed ID: 31658277

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-LKR). 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. Supported by National Science Foundation (United States) 1915544 to LK Reed. ,Supported by National Institutes of Health (United States) R25GM130517 to LK Reed.

Author Contributions: Bailey Lose: formal analysis, validation, writing - original draft, writing - review editing. Jeremy Girard: formal analysis, writing - review editing. Josephine Hayes: formal analysis, writing - review editing. Lane Weast: formal analysis, writing - review editing. Natalie Minkovsky: writing - original draft, writing - review editing. Sarah Justice: writing - original draft, writing - review editing. Jack A. Vincent: writing - original draft, writing - review editing. James J. Youngblom: supervision, writing - review editing. Lindsey J. Long: supervision, writing - review editing. Chinmay P. Rele: data curation, formal analysis, methodology, project administration, software, supervision, validation, visualization, writing - review editing. Laura K Reed: supervision, funding acquisition, conceptualization, project administration, writing - review editing.

Reviewed By: Anonymous

Nomenclature Validated By: Anonymous

History: Received October 10, 2023 **Revision Received** December 22, 2024 **Accepted** January 7, 2025 **Published Online** January 29, 2025 **Indexed** February 12, 2025

Citation: Lose, B; Girard, J; Hayes, J; Weast, L; Minkovsky, N; Justice, S; et al.; Reed, LK (2025). Gene model for the ortholog of *eIF4E1* in *Drosophila yakuba*. microPublication Biology. 10.17912/micropub.biology.001020