microPublication BIOLOGY

12/18/2024 - *Open Access*

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

Alyssa C. Koehler[1], Logan Cohen[2], Isaac Romo[3], Viet Le[4], James J. Youngblom[3], Amy T. Hark[4], Chinmay P. Rele[1], Laura K Reed[1§]

[1]University of Alabama, Tuscaloosa, AL US

[2]Worcester State University, Worcester MA, USA

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

[4]Muhlenberg College, Allentown, PA, USA

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

Abstract

Gene model for the ortholog of *glycogen synthase* (*Glys*) in the May 2011 (WUGSC dyak_caf1/DyakCAF1) 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.

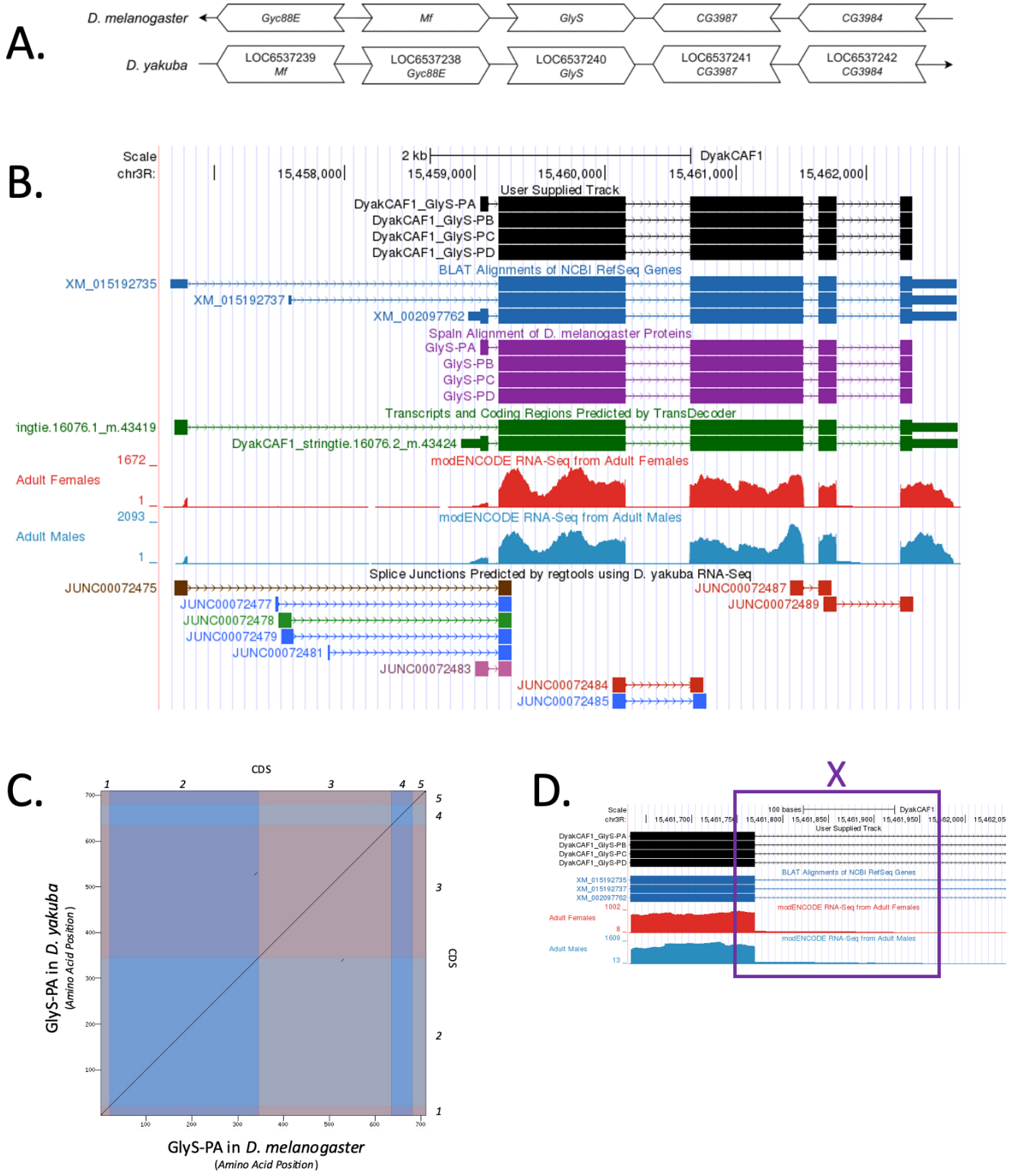

Figure 1. *Glys* ortholog gene model comparison between *D. yakuba* and *D. melanogaster*:

(A) Synteny comparison of the genomic neighborhoods for *Glys* in *Drosophila melanogaster* and *D. yakuba*. Thin underlying arrows indicate the DNA strand within which the target gene–*Glys*–is located in *D. melanogaster* (top) and *D. yakuba* (bottom). The thin arrow pointing to the right indicates that *Glys* is on the positive (+) strand in *D. yakuba*, and the thin arrow pointing to the left indicates that *Glys* is on the negative (-) strand in *D. melanogaster*. The wide gene arrows pointing in the same direction as *Glys* are on the same strand relative to the thin underlying arrows, while wide gene arrows pointing in the opposite direction of *Glys* 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 *Glys* 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). Splice junctions shown have a read-depth of 13-26, 60, 387, 810 and 1796-2030 supporting reads in blue, green, pink, brown, and red, respectively. **(C) Dot Plot of Glys-PA 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. CDS are also highlighted with alternating colors. **(D) UCSC Genome Browser displaying intronic RNA-Seq coverage within intron four of Glys-RA and intron three of Glys-RB, Glys-RC, and Glys-RD in *D. yakuba*.** The coding-regions of *Glys* 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*) and RNA-Seq from Adult Females and Adult Males (red and light blue, respectively; alignment of Illumina RNA-Seq reads from *D. yakuba*). The purple box denoted X highlights the higher coverage than expected in intron four for *Glys-RA* and intron three for *Glys-RB*, *Glys-RC*, and *Glys-RD*.

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).

"*Glycogen synthase* (*Glys*; aka. *GS*, *GlyS*) is a gene within the Insulin-signaling pathway in *Drosophila* and encodes a glycosyltransferase that catalyzes linkage of glucose monomers into glycogen. Glys activity is regulated allosterically by glucose 6-phosphate and phosphorylation/dephosphorylation allowing for control of cellular glycogen levels (Plyte et al., 1992; Roach et al., 2012). Null *Glys* mutants exhibit growth defects and reduced larval viability in *Drosophila* (Yamada et al., 2019)." (Backlund et al., 2024).

We propose a gene model for the *D. yakuba* ortholog of the *D. melanogaster Glycogen synthase* (*Glys*) gene, the summary of which can be found in the box above. The genomic region of the ortholog corresponds to the uncharacterized protein XP_002097798.1 (Locus ID LOC6537240) in the Dyak_CAF1 Genome Assembly of *D. yakuba* (GenBank Accession:

GCA_000005975.1 - Graveley et al., 2011). This model is based on RNA-Seq data from *D. yakuba* (SRP006203) and *Glys* in *D. melanogaster* using FlyBase release FB2022_04 (GCA_000001215.4; Larkin et al., 2021).

*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).

Synteny

The target gene, *Glys*, occurs on chromosome 3R in *D. melanogaster* and is flanked upstream by *Guanylyl cyclase at 88E* (*Gyc88E*) and *Myofilin* (*Mf*) and downstream by *CG3987* and *CG3984*. The *tblastn* search of *D. melanogaster* Glys-PA (query) against the *D. yakuba* (GenBank Accession: GCA_000005975.1) Genome Assembly (database) placed the putative ortholog of *Glys* within scaffold chromosome 3R (CM000160.2) at locus LOC6537240 (XP_002097798.1)— with an E-value of 0.0 and a percent identity of 99.09%. Furthermore, the putative ortholog is flanked upstream by LOC6537238 (XP_039232370.1) and LOC6537239 (XP_015048219.1), which correspond to *Gyc88E* and *Mf* in *D. melanogaster* (E-value: 0.0 and 0.0; identity: 95.56% and 99.18%, respectively, as determined by *blastp*; Figure 1A, Altschul et al., 1990). The putative ortholog of *Glys* is flanked downstream by LOC6537241 (XP_002097799.1) and LOC6537242 (XP_002097800.1), which correspond to *CG3987* and *CG3984* in *D. melanogaster* (E-value: 5e-157 and 4e-110; identity: 74.33% and 78.05%, respectively, as determined by *blastp*). The putative ortholog assignment for *Glys* in *D. yakuba* is supported by the following evidence: The genes surrounding the *Glys* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is completely conserved, supported by results generated from *blastp*; we conclude that LOC6537240 is the correct ortholog of *Glys* in *D. yakuba* (Figure 1A).

Protein Model

Glys in *D. yakuba* has two unique protein-coding isoforms Glys-PA and Glys-PC (identical to Glys-PB and Glys-PD; Figure 1B). mRNA isoform (Glys-RA) contains five CDSs. mRNA isoforms Glys-RC, Glys-RB and Glys-RD, which differ in their 5' UTRs, contain four CDSs. Relative to the ortholog in *D. melanogaster*, the RNA CDS number and protein isoform count are conserved. The sequence of Glys-PA in *D. yakuba* has 100.00% identity (E-value: 0.0) with the protein-coding isoform Glys-PA in *D. melanogaster*, as determined by *blastp* and illustrated by dot plot in *Figure 1C*. There appears to be high RNA-Seq coverage in the fourth intron and third intron of Glys-PA and Glys-PB, displayed in Figure 1D. Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (accession BK064679, BK064680, BK064681, BK064682). These data are also archived in the CaltechDATA repository (see “Extended Data” section below).

Special characteristics of the protein model

Intronic RNA-Seq Coverage in intron four of *Glys-RA* and intron three of *Glys-RC*: There is higher coverage than expected in intron four for *Glys-RA* and intron three for *Glys-RB*, *Glys-RC*, and *Glys-RD*, highlighted by the purple box denoted X (Figure 1D). One possible explanation is that RNA-Seq coverage could be from another location in the genome that has high sequence similarity, causing incorrect mapping of the RNA-Seq reads. To test this hypothesis, a *blastn* search was performed using the intronic DNA as the query sequence and the *D. yakuba* genome as the subject. This showed no results other than the *Glys* ortholog, suggesting the RNA-Seq coverage was mapped to the correct region. It is also possible that the high levels of intronic RNA-coverage could represent a novel isoform. High RNA-Seq coverage in this region has been observed in other species including *D. eugracilis*, *D. elegans*, and *D. ficusphila*. While this evidence supports a novel isoform, we do not have the data to definitively conclude the presence of a novel isoform in *D. yakuba*.

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* target gene for a given isoform and run it using *tblastn* against their target *Drosophila* species genome assembly (*Drosophila yakuba* (GCA_000005975.1)- Graveley et al., 2010)) 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. Genomic structure information (e.g., CDSs, CDS 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 each gene are generated by students under mentorship of their faculty course instructors. These models are then reconciled by a third independent researcher mentored by the project leaders to produce a final model like the one 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 mode, and Madeline 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 DyakCAF1_Glys. Resource Type: Model. File: DyakCAF1_GlyS.zip. DOI: 10.22002/171sf-ny118

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403-410. PubMed ID: 2231712

Backlund AE, White J, Grillo L, Ianniello T, Swedrowski A, Yang M, Jemc J, Rele CP 2024. Gene model for the ortholog of GlyS in *Drosophila busckii. microPublication Biology.* submitted.

Bock IR, Wheeler MR. 1972. The Drosophila melanogaster species group. Univ. Texas Publs Stud. Genet. 7(7213): 1-102. FBrf0024428.

Burla H. 1954. Zur Kenntnis der Drosophiliden der Elfenbeinkuste (Franzosisch West-Afrika). Revue suisse Zool. 61(Suppl.): 1-218. FBrf0009861.

Drosophila 12 Genomes Consortium, Clark, Andrew G, Eisen, Michael B, Smith, Douglas R, Bergman, Casey M, Oliver, Brian, et al., MacCallum, Iain. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450: 203-218. 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): 6546290. PubMed ID: 35266522

Grewal SS. 2009. Insulin/TOR signaling in growth and homeostasis: A view from the fly world. The International Journal of Biochemistry & Cell Biology 41: 1006-1010. PubMed ID: 18992839

Hietakangas V, Cohen SM. 2009. Regulation of Tissue Growth through Nutrient Sensing. Annual Review of Genetics 43: 389-410. PubMed ID: 19694515

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The Human Genome Browser at UCSC. Genome Research 12: 996-1006. PubMed ID: 12045153

Lachaise D, Tsacas L. 1983. Breeding-sites of tropical African Drosophilids. Ashburner, Carson, Thompson, 1981-1986. 3d: 21--332. FBrf0038884.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, et al., Lovato. 2020. FlyBase: updates to the*Drosophila melanogaster*knowledge base. Nucleic Acids Research 49: D899-D907. PubMed ID: 33219682

Lemeunier F, David J, Tsacas L, Ashburner M. 1986. The melanogaster species group. Ashburner, Carson, Thompson, 1981-1986. e: 147--256. FBrf0043749.

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. Nature Reviews Genetics 17: 758-772. PubMed ID: 27773922

Myers A, Hoffmann A, Natysin M, Arsham AM, Stamm J, Thompson JS, Rele CP. 2024. Gene model for the ortholog of *Myc* in *Drosophila ananassae*. *microPublication Biology* DOI: 10.22002/0c391-eeh07

Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al., Kent. 2020. The UCSC Genome Browser database: 2021 update. Nucleic Acids Research 49: D1046-D1057. PubMed ID: 33221922

Plyte S. 1992. Glycogen synthase kinase-3: functions in oncogenesis and development. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1114: 147-162. PubMed ID: 1333807

Prakash S. 1972. Origin of reproductive isolation in the absence of apparent genic differentiation in a geographic isolate of Drosophila pseudoobscura. Genetics 72(1): 143-55. PubMed ID: 5073854

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al., Kent. 2013. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30: 1003-1005. PubMed ID: 24227676

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

Roach PJ, Depaoli-Roach AA, Hurley TD, Tagliabracci VS. 2012. Glycogen and its metabolism: some new developments and old themes. Biochemical Journal 441: 763-787. PubMed ID: 22248338

Sturtevant AH. 1939. On the Subdivision of the Genus Drosophila. Proceedings of the National Academy of Sciences 25: 137-141. PubMed ID: 16577879

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

Yamada T, Habara O, Yoshii Y, Matsushita R, Kubo H, Nojima Y, Nishimura T. 2019. Role of glycogen in development and adult fitness in *Drosophila*. Development : 10.1242/dev.176149. PubMed ID: 30918052

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 Institutes of Health (United States) R25GM130517 to L.K. Reed. ,Supported by U.S. National Science Foundation (United States) 1915544 to L.K. Reed.

Author Contributions: Alyssa C. Koehler: formal analysis, validation, writing - original draft, writing - review editing. Logan Cohen: validation, visualization, writing - review editing. Isaac Romo: formal analysis, writing - review editing. Viet Le: formal analysis, writing - review editing. James J. Youngblom: supervision, writing - review editing. Amy T. Hark: 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 August 30, 2023 **Revision Received** November 20, 2024 **Accepted** December 5, 2024 **Published Online** December 18, 2024 **Indexed** January 1, 2025

Citation: 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*. microPublication Biology. 10.17912/micropub.biology.000983