Gene model for the ortholog of *Pten* in *Drosophila ananassae*

Graham M. Jones[1], D'Andrew Harrington[2], Zachary Lill[3], Amy T. Hark[3], Chelsey McKenna[2], Chinmay P. Rele[1], Laura K Reed[1§]

[1]University of Alabama, Tuscaloosa, AL USA

[2]College of Southern Nevada, Henderson, NV USA

[3]Muhlenberg College, Allentown, PA USA

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

Abstract

Gene model for the ortholog of *Phosphatase and tensin homolog* (*Pten*) in the May 2011 (Agencourt dana_caf1/DanaCAF1) Genome Assembly (GenBank Accession: GCA_000005115.1) of *Drosophila ananassae*. 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.
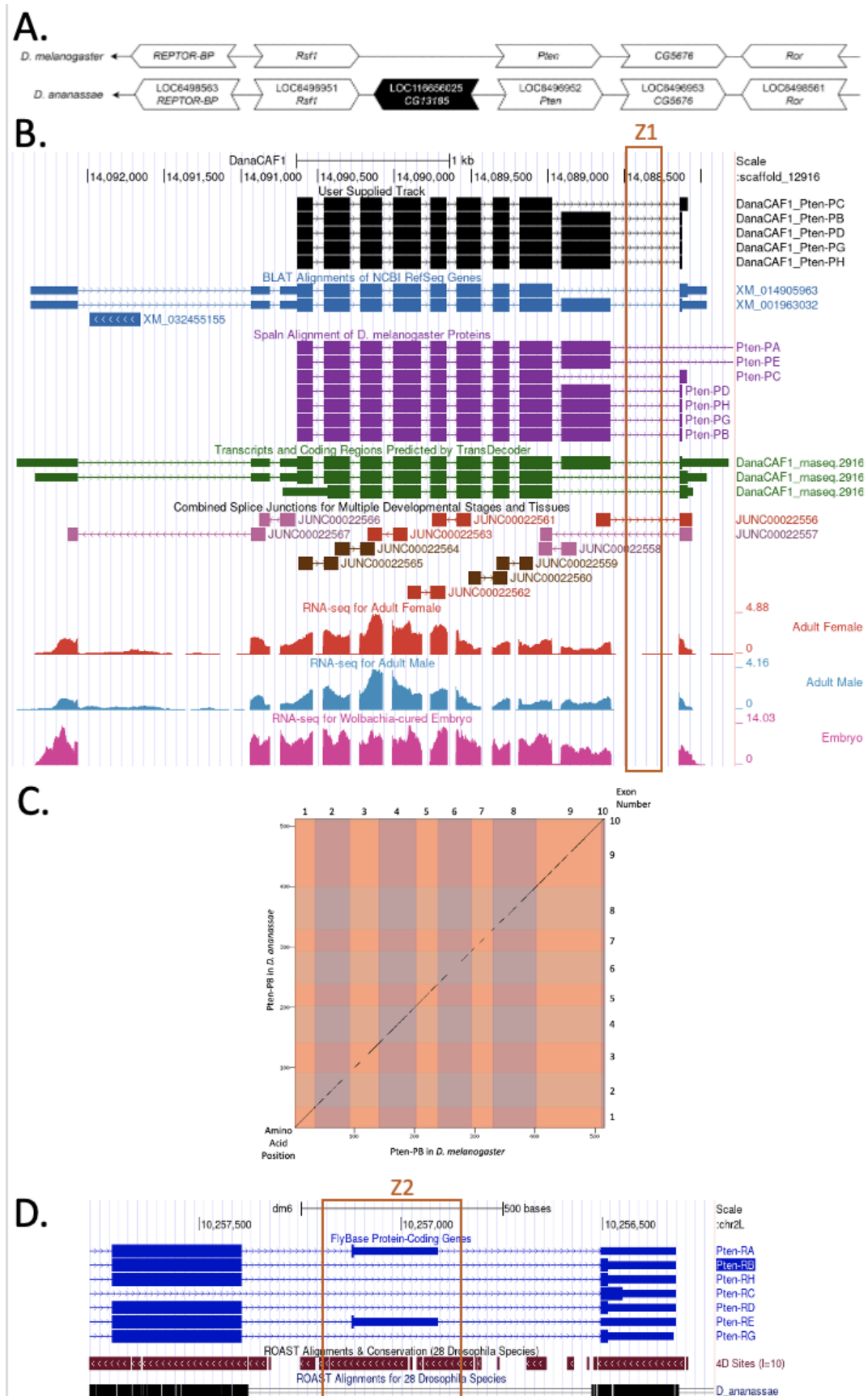
Figure 1. Genomic neighborhood and gene model for *Pten* in *Drosophila ananassae*:

(A) Synteny comparison of the genomic neighborhoods for _Pten_ in *Drosophila melanogaster* and *D. ananassae*. The thin arrows pointing to the left indicate that _Pten_ is on the negative (-) strand in *D. melanogaster* (top) and *D. ananassae*(bottom). The wide gene arrows pointing in the same direction as _Pten_ are on the same strand relative to the thin underlying arrows, while wide gene arrows pointing in the opposite direction of _Pten_ are on the opposite strand

relative to the thin underlying arrows. White gene arrows in *D. ananassae* indicate orthology to the corresponding gene in *D. melanogaster*, while black gene arrows indicate non-orthology. Gene symbols given in the *D. ananassae* gene arrows indicate the orthologous gene in *D. melanogaster*, while the locus identifiers are specific to *D. ananassae*. **(B) Gene Model in GEP UCSC Track Data Hub** (Raney et al., 2014). The coding-regions of *Pten* in *D. ananassae* are displayed in the User Supplied Track (black); CDSs 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. ananassae*), 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, Adult Males and Wolbachia-cured Embryos (red, light blue and pink, respectively); Splice Junctions Predicted by regtools using *D. ananassae* RNA-Seq, and alignment of Illumina RNA-Seq reads from *D. ananassae* (SRP006203, SRP007906; PRJNA257286, PRJNA388952). Splice junctions shown have a read-depth of 100-499, 500-999, >1000 supporting reads in pink, brown, and red, respectively. The region enclosed in orange, denoted Z1, highlights the region between the ninth and tenth CDS of *Rten-RB* (and all identical isoforms). **(C) Dot Plot of Pten-PB in *D. melanogaster* (*x*-axis) vs. the orthologous peptide in *D. ananassae* (*y*-axis).** Amino acid number is indicated along the left and bottom; CDS number is indicated along the top and right, and CDSs are also highlighted with alternating colors. Line breaks in the dot plot indicate mismatching amino acids at the specified location between species. **(D) Ninth and tenth CDSs *Pten* isoforms displayed in *D. melanogaster* UCSC Genome Browser with ROAST Alignments of 28 Drosophila Species track exhibiting conservation of sequence in *D. ananassae*.** Evidence tracks include FlyBase Protein-Coding Genes of *Pten* isoforms in *D. melanogaster* (blue), ROAST Alignments and Conservation (28 Drosophila Species) (burgundy) and conservation of sequence in *D. ananassae* (black). The final CDS of *Pten-RA* and *Pten-RE* is highlighted in orange, denoted Z2. The ROAST Alignment track (black) supports the absence of this CDS and its isoform (*Pten-RA*; *Pten-RE*) in *D. ananassae*.

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).

"The *Drosophila Phosphatase and tensin homolog* (*Pten* also known as *dPTEN*; FBgn0026379), identified due to is conservation to the human tumor suppressor gene acts as a protein and lipid phosphatase in the insulin signaling pathway (Goberdhan et al., 1999; Huang et al., 1999). Pten is known to affect cell number and size through the inhibition of the phosphoinositide 3-kinase (PI3K) and AKT kinase pathways (Goberdhan et al., 1999; Gao et al., 2000). Pten is involved in stabilizing cell junctions (Bardet et al., 2013) and regulates the cytoskeleton, controlling the localization and organization of actin (Goberdhan et al., 1999; von Stein et al., 2005)." (Lawson et al., 2024a).

"*D. ananassae* (NCBI:txid7217) 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 Doleschall (1858). *D. ananassae* is circumtropical (Markow and O'Grady 2005; https://www.taxodros.uzh.ch, accessed 1 Feb 2023), and often associated with human settlement (Singh 2010). It has been extensively studied as a model for its cytogenetic and genetic characteristics, and in experimental evolution (Kikkawa 1938; Singh and Yadav 2015)." (Lawson et al., 2024b).

We propose a gene model for the *D. ananassae* ortholog of the *D. melanogaster Phosphatase and tensin homolog* (*Pten*) gene. The genomic region of the ortholog corresponds to the uncharacterized protein XP_001963068.1 (Locus ID LOC6496952) in the *D. ananassae* May 2011 (Agencourt dana_caf1/DanaCAF1) Genome Assembly of *D. ananassae*

(GCA_000005115.1 - *Drosophila* 12 Genomes Consortium et al., 2007). This model is based on RNA-Seq data from *D. ananassae* (SRP006203, SRP007906; PRJNA257286, PRJNA388952 - Graveley et al., 2011) and *Pten* in *D. melanogaster* using FlyBase release FB2023_03 (GCA_000001215.4; Larkin et al., 2021; Gramates et al., 2022; Jenkins et al., 2022).

Synteny

The target gene, *Pten*, occurs on chromosome 2L in *D. melanogaster* and is flanked upstream by *REPTOR-binding partner* (*REPTOR-BP*) and *Repressor splicing factor 1* (*Rsf1*) and downstream by *CG5676* and *Ror* (*Ror*). The *tblastn* search of *D. melanogaster* Pten-PB (query) against the *D. ananassae* (GenBank Accession: GCA_000005115.1) Genome Assembly (database) placed the putative ortholog of *Pten* within scaffold scaffold_12916 (CH902620.1) at locus LOC6496952 (XP_001963068.1), with an E-value of 2e-82 and a percent identity of 92.78%. Furthermore, the putative ortholog is flanked upstream by LOC6498563 (XP_001963071.2), LOC6496951 (XP_032311424.1) and LOC116656025 (XP_032311046.1), which correspond to *REPTOR-BP*, *Rsf1* and *CG13185* in *D. melanogaster* (E-value: 1e-67, 2e-80 and 6e-32; identity: 83.90%, 82.50% and 67.39%, respectively, as determined by *blastp*; Figure 1A; Altschul et al., 1990). The putative ortholog of *Pten* is flanked downstream by LOC6496953 (XP_001963067.1) and XP_001963067.1 (XP_001963066.2), which correspond to *CG5676* and *Ror* in *D. melanogaster* (E-value: 1e-104 and 0.0; identity: 91.67% and 93.00%, respectively, as determined by *blastp*). The putative ortholog assignment for *Pten* in *D. ananassae* is supported by the following evidence: The genes surrounding the *Pten* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is nearly completely conserved (aside from the insertion of *CG13185*), supported by results generated from *blastp*, so we conclude that LOC6496952 is the correct ortholog of *Pten* in *D. ananassae* (Figure 1A).

Protein Model

Pten in *D. ananassae* has two unique protein-coding isoforms: Pten-PB (identical to Pten-PH, Pten-PG and Pten-PD) and Pten-PC (Figure 1B, black). mRNA isoforms *Pten-RB*, *Pten-RH*, *Pten-RG* and *Pten-RD* contain ten CDSs and mRNA isoform *Pten-RC* contains nine CDSs. The two isoforms differ due to alternative splice junctions JUNC 00022557 and 00022558. Relative to the ortholog in *D. melanogaster*, the CDS number and protein isoform count are not-conserved. In *D. melanogaster, Pten-RB* (*Pten-RH*, *Pten-RG* and *Pten-RD*) contains nine CDSs and *Pten-RC* contains eight CDSs. In addition, *Pten* in *D. melanogaster* contains an additional unique isoform, *Pten-RA* (identical to *Pten-RE*), which contains ten CDSs. The sequence of Pten-PB in *D. ananassae* has 82.72% identity (E-value: 0.0) with the protein-coding isoform Pten-PB in *D. melanogaster*, as determined by *blastp* (Figure 1C). Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (accession BK064655, BK064656, BK064657, BK064658, BK064659). These data are also archived in the CaltechDATA repository (see "Extended Data" section below).

Special characteristics of the protein model

Putative absence of *Pten-RA* and *Pten-RE* in *D. ananassae*:
D. melanogaster contains three unique protein isoforms: Pten-PB (Pten-PH, Pten-PG and Pten-PD), Pten-PC, and Pten-PA (Pten-PE). Only two of these unique isoforms have been confirmed in *D. ananassae*: Pten-PB (Pten-PH, Pten-PG and Pten-PD) and Pten-PC (Figure 1B, black). In *D. melanogaster*, *Pten-RA* (*Pten-RE*) uses an alternative splice acceptor site for its final CDS, located between the ninth and tenth CDS of *Pten-RB* (and identical isoforms). Lack of RNA-Seq expression and predicted splice junctions in the expected region of the final CDS of *Pten-RA* (*Pten-RE*), highlighted in orange (Z1), support the absence of this unique isoform in *D. ananassae* (Figure 1B). Furthermore, conservation of sequence of *D. ananassae* by ROAST Alignments and Conservation (28 Drosophila Species) displays a lack of conservation in the region of the expected final CDS of *Pten-RA* (*Pten-RE*), indicating the loss of Pten-PA and Pten-PE in *D. ananassae* (Figure 1D, black).

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).

Extended Data

Description: GFF, FASTA, and PEP of Pten in Drosophila ananassae. Resource Type: Model. File: DanaCAF1_Pten.zip. DOI: 10.22002/p0pst-n6h15

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3): 403-10. PubMed ID: 2231712

Bardet PL, Guirao B, Paoletti C, Serman F, Léopold V, Bosveld F, et al., Bellaïche Y. 2013. PTEN controls junction lengthening and stability during cell rearrangement in epithelial tissue. Dev Cell 25(5): 534-46. PubMed ID: 23707736

Bock, I.R., Wheeler, M.R. (1972). The Drosophila melanogaster species group. Univ. Texas Publs Stud. Genet. 7(7213): 1-102. FBrf0024428 PubMed ID: null

Doleschall, C.L. 1858. Derde bijdrage tot de kennis der Dipteren fauna van nederlandsch indie. Natuurk. Tijd. Ned.-Indie 17: 73-128. FBrf0000091 PubMed ID: null

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

Gao X, Neufeld TP, Pan D. 2000. Drosophila PTEN regulates cell growth and proliferation through PI3K-dependent and -independent pathways. Dev Biol 221(2): 404-18. PubMed ID: 10790335

Goberdhan DC, Paricio N, Goodman EC, Mlodzik M, Wilson C. 1999. Drosophila tumor suppressor PTEN controls cell size and number by antagonizing the Chico/PI3-kinase signaling pathway. Genes Dev 13(24): 3244-58. PubMed ID: 10617573

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, et al., undefined. 2022. FlyBase: a guided tour of highlighted features. Genetics 220: 10.1093/genetics/iyac035. DOI: 10.1093/genetics/iyac035

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(5): 1006-10. PubMed ID: 18992839

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 2025: 10.17912/micropub.biology.001168. PubMed ID: 39845267

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

Huang H, Potter CJ, Tao W, Li DM, Brogiolo W, Hafen E, Sun H, Xu T. 1999. PTEN affects cell size, cell proliferation and apoptosis during Drosophila eye development. Development 126(23): 5365-72. PubMed ID: 10556061

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 WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6): 996-1006. PubMed ID: 12045153

Kikkawa H. 1938 Studies on the genetics and cytology of Drosophila ananassae. Genetica20, 458–516. https://doi.org/10.1007/BF01531779 PubMed ID: null

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

Lawson ME, Dela Cruz M, Harrington DL, Vincent JA, McKenna C, Goodman A, Barnard D, Rele CP. 2024a. Gene model for the ortholog of Pten in Drosophila miranda. MicroPubl Biol 2024: 10.17912/micropub.biology.000986. PubMed ID: 39381638

Lawson ME, McAbee M, Lucas RA, Tanner S, Wittke-Thompson J, Pelletier TA, Ozsoy Z, Sterne-Marr R, Rele CP, Reed LK. 2024b. Gene model for the ortholog of *Ilp5* in *Drosophila ananassae*. MicroPubl Biol 2024: 10.17912/micropub.biology.000782. PubMed ID: 39717145

Markow, T.A. and O'Grady, P. (2005) Drosophila: A guide to species identification and use. London: Academic Press. ISBN: 978-0-12-473052-6 PubMed ID: null

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. Nat Rev Genet 17(12): 758-772. 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: 10.17912/micropub.biology.000856. 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. F1000Research 11: 1579. DOI: 10.12688/f1000research.126839.2

Singh BN. 2010. Drosophila ananassae: a good model species for genetical, behavioural and evolutionary studies. Indian J Exp Biol 48(4): 333-45. PubMed ID: 20726331

Singh BN, Yadav JP. 2015. Status of research on Drosophila ananassae at global level. J Genet 94(4): 785-92. PubMed ID: 26690536

Sturtevant AH. 1939. On the Subdivision of the Genus Drosophila. Proc Natl Acad Sci U S A 25(3): 137-41. 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

von Stein W, Ramrath A, Grimm A, Müller-Borg M, Wodarz A. 2005. Direct association of Bazooka/PAR-3 with the lipid phosphatase PTEN reveals a link between the PAR/aPKC complex and phosphoinositide signaling. Development 132(7): 1675-86. PubMed ID: 15743877

Author Contributions: Graham M. Jones: formal analysis, validation, writing - original draft, writing - review editing. D'Andrew Harrington: formal analysis, writing - review editing. Zachary Lill: formal analysis, writing - review editing. Amy T. Hark: supervision, writing - review editing. Chelsey McKenna: 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, methodology, writing - review editing.

Reviewed By: Anonymous, GEP Review Panel 2

Nomenclature Validated By: Anonymous