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

Madeline L. Gruys[1], Jhilam Dasgupta[1], Joshua Williams[2], Emile Moura Coelho da Silva[3], Jacqueline Wittke-Thompson[2], Chinmay P. Rele[1], Laura K. Reed[1§]

[1]The University of Alabama, Tuscaloosa, AL USA

[2]University of St. Francis, Joliet, IL USA

[3]University of Evansville, Evansville, IN USA

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

Abstract

Here is presented a gene model for the ortholog of Thor (*Thor*) in the *D. ananassae* 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.

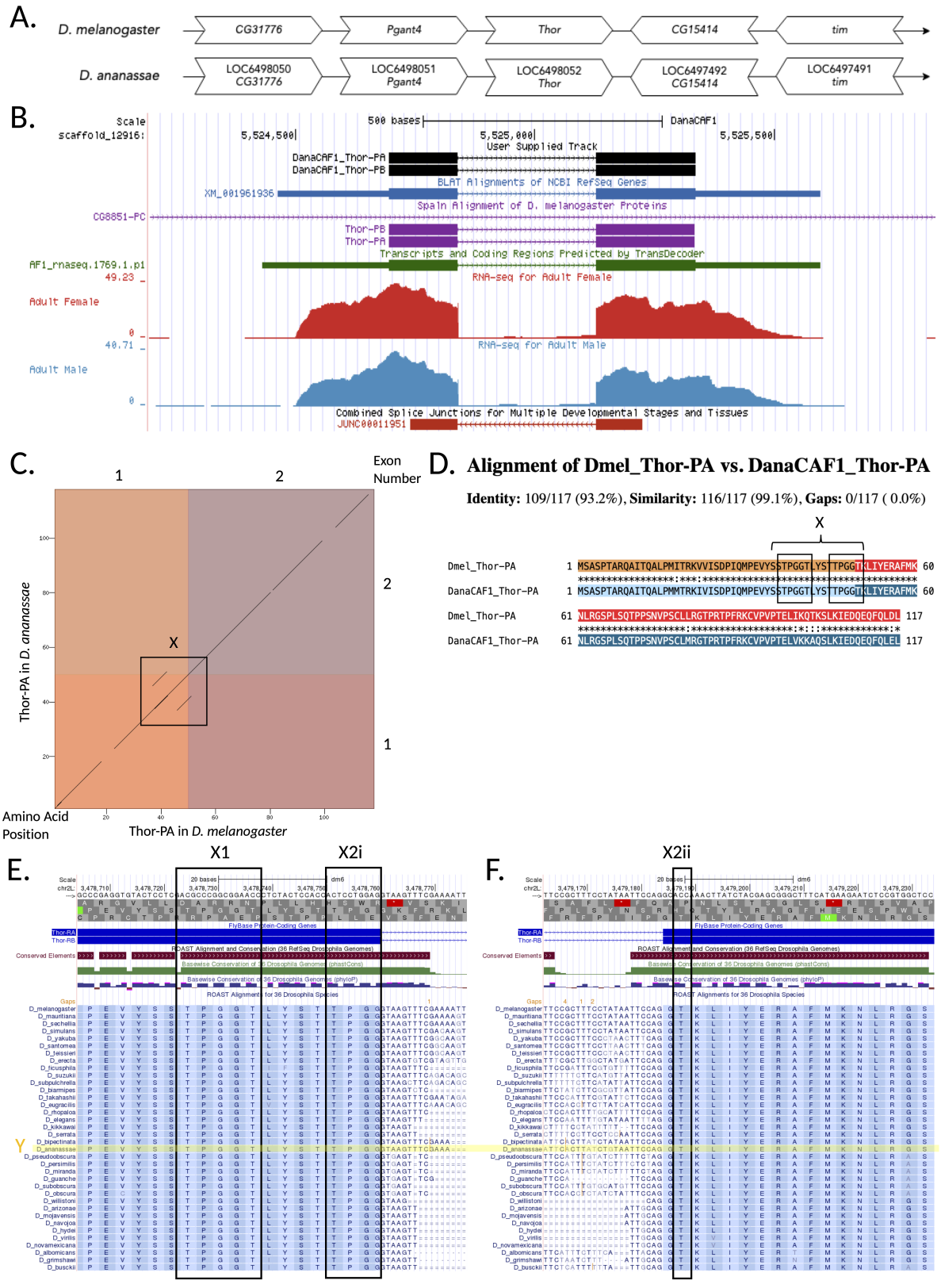

Figure 1. Genomic neighborhood and gene model for *Thor* in *Drosophila ananassae*:

(A) Synteny comparison of the genomic neighborhoods for *Thor* in *Drosophila melanogaster* and *D. ananassae*. Thin underlying arrows indicate the DNA strand within which the target gene–*Thor*–is located in *D. melanogaster* (top) and *D. ananassae* (bottom). Thin arrows pointing to the right indicate that *Thor* is on the positive (+) strand in both *D. ananassae* and *D. melanogaster*. The wide gene arrows pointing in the same direction as *Thor* are on the same strand relative to the

thin underlying arrows, while wide gene arrows pointing in the opposite direction of *Thor* 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*. 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 *Thor* in *D. ananassae* are displayed in the User Supplied Track (black); coding 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 and Adult Males (red and light blue, respectively; alignment of Illumina RNA-Seq reads from *D. ananassae*), and Splice Junctions Predicted by regtools using *D. ananassae* RNA-Seq (Graveley *et al.*, 2011; SRP006203, SRP007906; PRJNA257286, PRJNA388952). The splice junction shown in red (JUNC00011951) has a read-depth of 4677. **(C) Dot Plot of Thor-PA in *D. melanogaster* (*x*-axis) vs. the orthologous peptide in *D. ananassae* (*y*-axis).** Amino acid number is indicated along the left and bottom; CDSs number is indicated along the top and right, and CDSs are also highlighted with alternating colors. The black box X highlights a repeating amino acid sequence in that region. **(D) Protein Alignment Thor-PA in *D. melanogaster* and the orthologous peptide in *D. ananassae*.** The alternating colored rectangles represent adjacent CDSs. The symbols in the match line denote the level of similarity between the aligned residues. An asterisk (*) indicates that the aligned residues are identical. A colon (:) indicates the aligned residues have highly similar chemical properties—roughly equivalent to scoring > 0.5 in the Gonnet PAM 250 matrix (Gonnet *et al.*, 1992). A period (.) indicates that the aligned residues have weakly similar chemically properties—roughly equivalent to scoring > 0 and ≤ 0.5 in the Gonnet PAM 250 matrix. A space indicates a gap or mismatch when the aligned residues have a complete lack of similarity—roughly equivalent to scoring ≤ 0 in the Gonnet PAM 250 matrix. The amino acid sequence shows there is a small repeat at the end of CDS one (TPGGT) as shown in the 2 boxes, corresponding to Box X in the dot plot. **(E) End of first CDS of *Thor* in *D. melanogaster* displaying conservation of TPGGT repeats.** Evidence tracks include base positions shown in black at the top of the browser, amino acid sequence (grey blocks), FlyBase Protein-Coding Genes (blue), ROAST Alignment and Conservation (36 RefSeq *Drosophila* Genomes) (burgundy), Basewise Conservation of 36 *Drosophila* Genomes generated by phastCons and phyloP shown in green and dark blue, respectively; ROAST Alignments of 36 *Drosophila* species with species listed on the left in dark blue and amino acid sequence in pale blue. Boxes in black denoted X1 and X2i highlight the conservation of the amino acid repeat (TPGGT) at the end of the first CDS of *Thor* across 36 *Drosophila* species. The yellow highlighted box labeled Y on the left encloses the sequence in *D. ananassae*, with X2i encompassing only a portion of the TPGGT repeat, (TPGG). **(F) Beginning of second CDS of *Thor* in *D. melanogaster* displaying conservation of amino acid sequence.** Evidence tracks are identical to those in Figure 1E. The box in black denoted X2ii highlights the conservation of the amino acid repeat (TPGGT) at the beginning of the second CDS of *Thor* in *D. melanogaster*, featuring the remaining amino acids of the repeat (T). The continuation of the yellow highlighted box Y encloses the sequence 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 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. 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., 2024).

We propose a gene model for the *D. ananassae* ortholog of the *D. melanogaster* Thor *(Thor)* gene. The genomic region of the ortholog corresponds to the uncharacterized protein LOC6498052 (RefSeq accession XP_001961972.1) in the dana_caf1 Genome Assembly of *D. ananassae* (GenBank Accession: GCA_000005115.1, Drosophila 12 Genomes Consortium et al., 2007). This model is based on RNA-Seq data from *D. ananassae* (Graveley et al., 2011; SRP006203, SRP007906; PRJNA257286, PRJNA388952) and *Thor* in *D. melanogaster* using FlyBase release FB2022_04, (GCA_000001215.4; Larkin et al., 2021; Gramates et al., 2022; Jenkins et al., 2022).

Thor (*Thor*; also known as *4E-BP*), a core component of the insulin signaling pathway, encodes a eukaryotic translation initiation factor 4E binding protein that is controlled by the product of *mTor* (Bernal and Kimbrell 2000; Marr II et al., 2007). The *Drosophila* forkhead transcription factor (*dFOXO*) activates *Thor* transcription and contributes to translation regulation, response to environmental stress, and cell growth regulation (Tettweiler et al., 2005; Miron et al., 2001). *Thor* is an effector of *PI(3)K/Akt* signaling and cell growth in *Drosophila* (Miron et al., 2001) and participates in host immune defense by connecting a translational regulator with innate immunity (Bernal and Kimbrell 2000).

Synteny

The reference gene, *Thor*, is found on chromosome 2L in *D. melanogaster* and is flanked upstream by *CG31776* and Polypeptide N-Acetylgalactosaminyltransferase 4 *(Pgant4)* and downstream by *CG15414* and timeless *(tim)*. The *tblastn* search of *D. melanogaster* Thor-PA (query) against the *D. ananassae* (GenBank Accession:GCA_000005115.1) Genome Assembly (database) placed the putative ortholog of *Thor* within scaffold_12916 (CH902620.1) at locus LOC6498052 (XP_001961972.1)— with an E-value of 1e-78 and a percent identity of 93.16%. Furthermore, the putative ortholog is flanked upstream by LOC6498050 (XP_001961970.1) and LOC6498051 (XP_001961971.1), which correspond to *CG31776* and *Pgant4* in *D. melanogaster* (E-value: 0.0 and 0.0; identity: 47.45% and 74.40%, respectively, as determined by *blastp*; Figure 1A, Altschul et al., 1990). The putative ortholog of *Thor* is flanked downstream by LOC6497492 (XP_044572519.1) and LOC6497491 (XP_032307461.1), which correspond to *CG15414* and *tim* in *D. melanogaster* (E-value: 2e-108 and 0.0; identity: 83.33% and 85.86%, respectively, as determined by *blastp*). The putative ortholog assignment for *Thor* in *D. ananassae* is supported by the following evidence: The genes surrounding the *Thor* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is completely conserved, verified by e-values and percent identities, so we conclude that LOC6498052 is the correct ortholog of *Thor* in *D. ananassae* (Figure 1A).

Protein Model

Thor in *D. ananassae* has two mRNA transcripts (Thor-RA, Thor-RB), which encode identical protein-coding isoforms (Thor-PA; Thor-PB; Figure 1B). These isoforms (Thor-PA, Thor-PB) contain two CDSs. Relative to the ortholog in *D. melanogaster*, the CDS number is conserved. The sequence of Thor-PA in *D. ananassae* has 93.16% identity (E-value: 1e-78) with the protein-coding isoform Thor-PA in *D. melanogaster*, as determined by *blastp* (Figure 1C). Consistent with

the *blastp* search result which shows 93.16% identity between *D. melanogaster* Thor-PA and the *D. ananassae* protein model as well as the low sensitivity parameters used to generate the dot plot (i.e., word size = 3; neighborhood threshold = 11), the dot plot of the two protein sequences contain multiple small gaps along the diagonal (Figure 1C). The dots on either side of the diagonal at the end of CDS one in the dot pot (Box X, Figure 1C) indicate the presence of two identical copies of the sequence TPGGT, as shown in the two boxes labeled B in the protein alignment (Box X, Figure 1D). Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (accession BK064668, BK064669). These data are also archived in the CaltechDATA repository (see “Extended Data” section below).

Special characteristics of the protein model

TPGGT Repeat: The dots on either side of the diagonal at the end of CDS one in the dot pot (Figure 1C) indicates the presence of two identical copies of the sequence TPGGT, as shown in the two boxes labeled X1 and X2 in the protein alignment (Figure 1D). This small repeat is conserved in many *Drosophila* species as shown by the black boxes denoted X1, X2i and X2ii, with *D. ananassae* highlighted in the yellow box, Y (Figure 1E and Figure 1F).

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/genechecker/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).

Acknowledgements: This publication is dedicated to the memory of Joshua Williams. We would like to thank Wilson Leung for developing and maintaining the technological infrastructure that was used to create this gene model. Thank you to FlyBase for providing the definitive database for *Drosophila melanogaster* gene models.

Extended Data

Description: GFF, FASTA, PEP. Resource Type: Model. File: DanaCAF1_Thor.zip. DOI: 10.22002/7z36m-eek13

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

Bernal A, Kimbrell DA. 2000. Drosophila Thor participates in host immune defense and connects a translational regulator with innate immunity. Proc Natl Acad Sci U S A 97(11): 6019-24. PubMed ID: 10811906

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

Doleschall CL. 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

Gonnet GH, Cohen MA, Benner SA. 1992. Exhaustive matching of the entire protein sequence database. Science 256(5062): 1443-5. PubMed ID: 1604319

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

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. Genetica 20: 458-516. 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, McAbee M, Lucas RA, Tanner S, Wittke-Thompson J, Pelletier TA, et al., Reed LK. 2024. Gene model for the ortholog of Ilp5 in Drosophila ananassae. MicroPubl Biol 2024. PubMed ID: 39717145

Markow TA 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

Marr MT 2nd, D'Alessio JA, Puig O, Tjian R. 2007. IRES-mediated functional coupling of transcription and translation amplifies insulin receptor feedback. Genes Dev 21(2): 175-83. PubMed ID: 17234883

Miron M, Verdú J, Lachance PE, Birnbaum MJ, Lasko PF, Sonenberg N. 2001. The translational inhibitor 4E-BP is an effector of PI(3)K/Akt signalling and cell growth in Drosophila. Nat Cell Biol 3(6): 596-601. PubMed ID: 11389445

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. (2020). Manual Annotation of Genes within Drosophila Species: the Genomics Education Partnership protocol. bioRxiv 2020.12.10.420521 doi: 10.1101/2020.12.12.420521 PubMed ID: null

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

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

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

Tettweiler G, Miron M, Jenkins M, Sonenberg N, Lasko PF. 2005. Starvation and oxidative stress resistance in Drosophila are mediated through the eIF4E-binding protein, d4E-BP. Genes Dev 19(16): 1840-3. PubMed ID: 16055649

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.

Conflicts of Interest: The authors declare that there are no conflicts of interest present.

Author Contributions: Madeline L. Gruys: formal analysis, validation. Jhilam Dasgupta: formal analysis, validation, writing - original draft, writing - review editing. Joshua Williams: formal analysis, writing - review editing. Emile Moura Coelho da Silva: formal analysis, writing - review editing. Jacqueline Wittke-Thompson: supervision, writing - review editing. Chinmay P. Rele: data curation, formal analysis, methodology, project administration, software, supervision, validation, visualization, writing - review editing. Laura K. Reed: conceptualization, funding acquisition, project administration, methodology, supervision, writing - review editing.

Reviewed By: Anonymous

Nomenclature Validated By: Anonymous

History: Received May 10, 2023 **Revision Received** November 29, 2023 **Accepted** July 3, 2026 **Published Online** July 6, 2026 **Indexed** July 20, 2026

Citation: Gruys ML, Dasgupta J, Williams J, Moura Coelho da Silva E, Wittke-Thompson J, Rele CP, Reed LK. 2026. Gene model for the ortholog of *Thor* in *Drosophila ananassae*. microPublication Biology. 10.17912/micropub.biology.000854