microPublication BIOLOGY

11/12/2024 - *Open Access*

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

Jhilam Dasgupta[1], Emile Moura Coelho da Silva[2], Gregory Sileo[3], Joyce Stamm[2], Thomas C. Giarla[3], Chinmay P. Rele[1§]

[1]University of Alabama, Tuscaloosa, Alabama, United States

[2]University of Evansville, Evansville, Indiana, United States

[3]Siena College, Albany, New York, United States

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

Abstract

Gene model for the ortholog of Thor (*Thor*) in the *D. yakuba* 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.

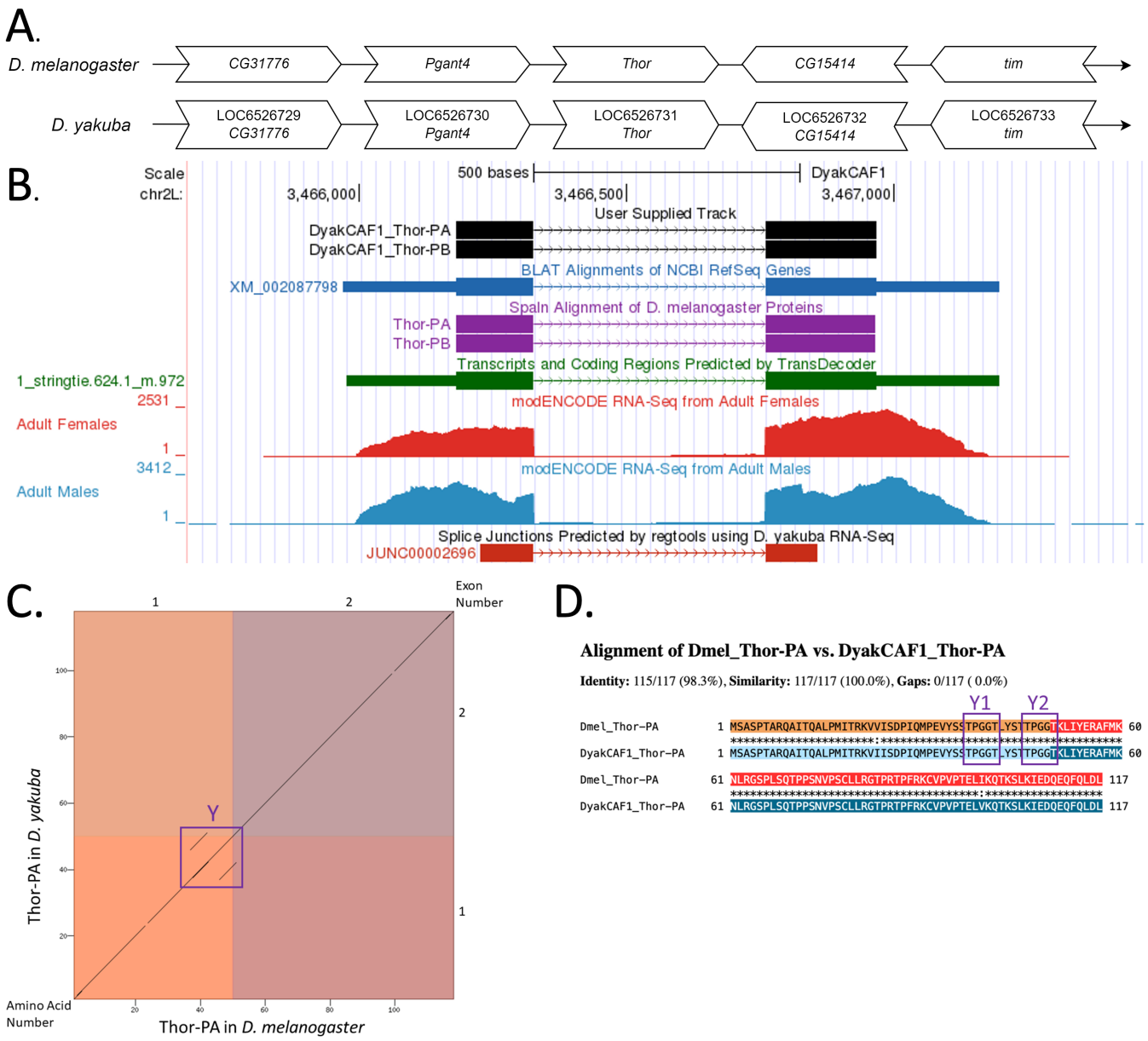

Figure 1.

(A) Synteny comparison of the genomic neighborhoods for *Thor* in *Drosophila melanogaster* and *D. yakuba*. Thin underlying arrows indicate the DNA strand within which the target gene–*Thor*–is located in *D. melanogaster* (top) and *D. yakuba* (bottom). The thin arrows pointing to the right indicate that *Thor* is on the positive (+) strand in *D. yakuba* 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. 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 *Thor* in *D. yakuba* 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. 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). The splice junction shown in red (JUNC00002696) has a read-depth score of 3487. **(C) Dot Plot of Thor-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, and CDSs are also highlighted with alternating colors. Line breaks in the dot plot indicate mismatching amino acids at the specified location between species. The line breaks shown are small and determined to be insignificant in the determination of the putative ortholog of *Thor* in *D. yakuba*. The purple box denoted Y encloses dots on either side of the CDS which indicates a repeating sequence in that region. **(D) Protein alignment of Thor-PA in *D. melanogaster* and the orthologous peptide in *D. yakuba*.** 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 two purple boxes Y1 and Y2, corresponding to the repeating sequence displayed in the dot plot.

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* (Taxonomic ID: 7245) 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 widespread in sub-Saharan Africa and Madagascar (Lemeunier et al., 1986; https://www.taxodros.uzh.ch, accessed 1 Feb 2023; Markow and O'Grady 2006) where figs served as a primary host along with other rotting fruits (Lachaise and Tsacas 1983)." (Koehler et al., 2024).

"*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 *tor* (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)." (Gruys et al., 2024).

We propose a gene model for the *D. yakuba* ortholog of the *D. melanogaster* Thor (*Thor*) gene. The genomic region of the ortholog corresponds to the uncharacterized protein LOC6526731 (RefSeq accession XP_002087834.1) in the Dyak_CAF1 Genome Assembly of *D. yakuba* (GenBank Accession: GCA_000005975.1; *Drosophila* 12 Genomes Consortium, 2007). This

model is based on RNA-Seq data from *D. yakuba* (SRP006203) 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).

Synteny

The target gene, *Thor*, occurs on chromosome 2L in *D. melanogaster* and is flanked upstream by *CG31776* and *Polypeptide N-Acetylgalactosaminyltransferase* (*Pgant4*) and downstream by *CG15414* and *timeless* (*tim*). The *tblastn* search of *D. melanogaster* Thor-PA (query) against the *D. yakuba* (GenBank Accession: GCA_000005975.1) Genome Assembly (database) placed the putative ortholog of *Thor* within scaffold chromosome 2L (CM000157.2) at locus LOC6526731 (XP_002087834.1)— with an E-value of 8e-40 and a percent identity of 98.55%. Furthermore, the putative ortholog is flanked upstream by LOC6526729 (XP_002087832.1) and LOC6526730 (XP_002087833.1), which correspond to *CG31776* and *Pgant4* in *D. melanogaster* (E-value: 0.0 and 0.0; identity: 76.41% and 91.31%, respectively, as determined by *blastp*; Figure 1A, Altschul et al., 1990). The putative ortholog of *Thor* is flanked downstream by LOC6526732 (XP_039226373.1) and LOC6526733 (XP_039226369.1), which correspond to *CG15414* and *tim* in *D. melanogaster* (E-value: 7e-136 and 0.0; identity: 95.94% and 95.39%, respectively, as determined by *blastp*). The putative ortholog assignment for *Thor* in *D. yakuba* 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, supported by e-values and percent identities, so we conclude that LOC6526731 is the correct ortholog of *Thor* in *D. yakuba* (Figure 1A).

Protein Model

Thor in *D. yakuba* has two mRNA isoforms (*Thor-RA*, *Thor-RB*; Figure 1B), both of which contain two CDSs. Relative to the ortholog in *D. melanogaster*, the RNA CDS number is conserved. The sequence of Thor-PA in *D. yakuba* has 98.29% identity (E-value: 7e-82) with the protein-coding isoform Thor-PA in *D. melanogaster*, as determined by *blastp* (Figure 1C). A small repeat sequence is present in the first CDS of the putative ortholog, shown in black (Box X) in the dot plot (Figure 1C) corresponding to the protein alignment in Figure 1D. Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (accession **BK059538** and **BK059539)**. These data are also archived in the CaltechDATA repository (see "Extended Data" section below).

Special characteristics of the protein model

There is a small repeat at the end of CDS one consisting of five amino acids (TPGGT) as shown in purple (Box Y in Figure 1C and Box Y1 and Y2 in Figure 1D). This small repeat is conserved in many species of *Drosophila*, such as *D. eugracilis*, *D. simulans*, and *D. willistoni* (Gruys et al., 2023).

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)) 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. melanogaster* 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:

We would like to thank Wilson Leung for developing and maintaining the technological infrastructure that was used to create this gene model, Madeline L. Gruys for retrofitting this model and Laura K. Reed for overseeing the project. 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. This article was prepared while Joyce Stamm was employed at the University of Evansville. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Extended Data

Description: Amino acid, nucleotide, and GFF files describing the gene model. Resource Type: Dataset. File: DyakCAF1_Thor 2.zip. DOI: 10.22002/wz8w1-tq831

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

Bernal A, Kimbrell DA. 2000. *Drosophila Thor* participates in host immune defense and connects a translational regulator with innate immunity. Proceedings of the National Academy of Sciences 97: 6019-6024. DOI: 10.1073/pnas.100391597

Bock IR, Wheeler MK. 1972. The *Drosophila melanogaster* species group. *University of Texas Publications*, **7213**, 1–102.

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

Gruys ML, Dasgupta J, Williams J, Moura Coelho Da Silva E, Stamm J, Wittke-Thompson J, Rele CP. 2024. Gene model for the ortholog of *Thor* in *Drosophila ananassae*. (submitted)

Hietakangas V, Cohen SM. 2009. Regulation of Tissue Growth through Nutrient Sensing. Annual Review of Genetics 43: 389-410. DOI: 10.1146/annurev-genet-102108-134815

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

Koehler AC, Romo I, Le V, Romo I, Youngblom JJ, Hark AT, Rele CP. 2024. Gene model for the ortholog of *GlyS* in *Drosophila yakuba*. microPublication Biology. (submitted)

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

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

Lemeunier F, David JR, Tsacas L, Ashburner MA. 1986. The melanogaster species group, pp. 147–256 in The Genetics and Biology of Drosophila, Vol. 3e, edited by M. A. Ashburner, H. L. Carson and J. N. Thompson. Academic Press, London.

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

Marr MT, D'Alessio JA, Puig O, Tjian R. 2007. IRES-mediated functional coupling of transcription and translation amplifies insulin receptor feedback. Genes & Development 21: 175-183. DOI: 10.1101/gad.1506407

Miron M, Verdú J, Lachance PED, 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. Nature Cell Biology 3: 596-601. DOI: 10.1038/35078571

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, Hoffmann A, Natysin M, Arsham AM, Stamm J, Thompson JS, Rele CP. 2024. Gene model for the ortholog of *Myc* in *Drosophila ananassae*. (submitted).

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

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 & Development 19: 1840-1843. DOI: 10.1101/gad.1311805

Sturtevant AH. 1939. On the Subdivision of the Genus Drosophila. Proc Natl Acad Sci U S A 25(3): 137-41. PubMed ID: 16577879

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: Jhilam Dasgupta: formal analysis, validation, writing - original draft, writing - review editing. Emile Moura Coelho da Silva: formal analysis, writing - review editing. Gregory Sileo: formal analysis, writing - review editing. Joyce Stamm: supervision, writing - review editing. Thomas C. Giarla: supervision, writing - review editing. Chinmay P. Rele: data curation, formal analysis, methodology, project administration, software, supervision, validation, visualization, writing - review editing.

Reviewed By: David Molik

History: Received October 11, 2023 **Revision Received** July 8, 2024 **Accepted** October 18, 2024 **Published Online** November 12, 2024 **Indexed** November 26, 2024

Citation: Dasgupta, J; Moura Coelho da Silva, E; Sileo, G; Stamm, J; Giarla, TC; Rele, CP (2024). Gene model for the ortholog of *Thor* in *Drosophila yakuba*. microPublication Biology. 10.17912/micropub.biology.001029