microPublication BIOLOGY
11/30/2024 - *Open Access*

Gene model for the ortholog *Myc* in *Drosophila ananassae*

Abigail Myers[1], Alexa Hoffman[2], Mindy Natysin[3], Andrew M Arsham[3], Joyce Stamm[2], Jeffrey S. Thompson[4], Chinmay P. Rele[5], Laura K Reed[6§]
[1]University of Alabama, Tuscaloosa, Alabama, US
[2]University of Evansville, Evansville, Indiana, US
[3]Bemidji State University, Bemidji, Minnesota, US
[4]Denison University, Granville, Ohio, US
[5]The University of Alabama, Tuscaloosa, AL USA
[6]Biological Sciences, University of Alabama, Tuscaloosa, Alabama, US
[§]To whom correspondence should be addressed: lreed1@ua.edu

Abstract

Gene model for the ortholog of Myc (*Myc*) 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.

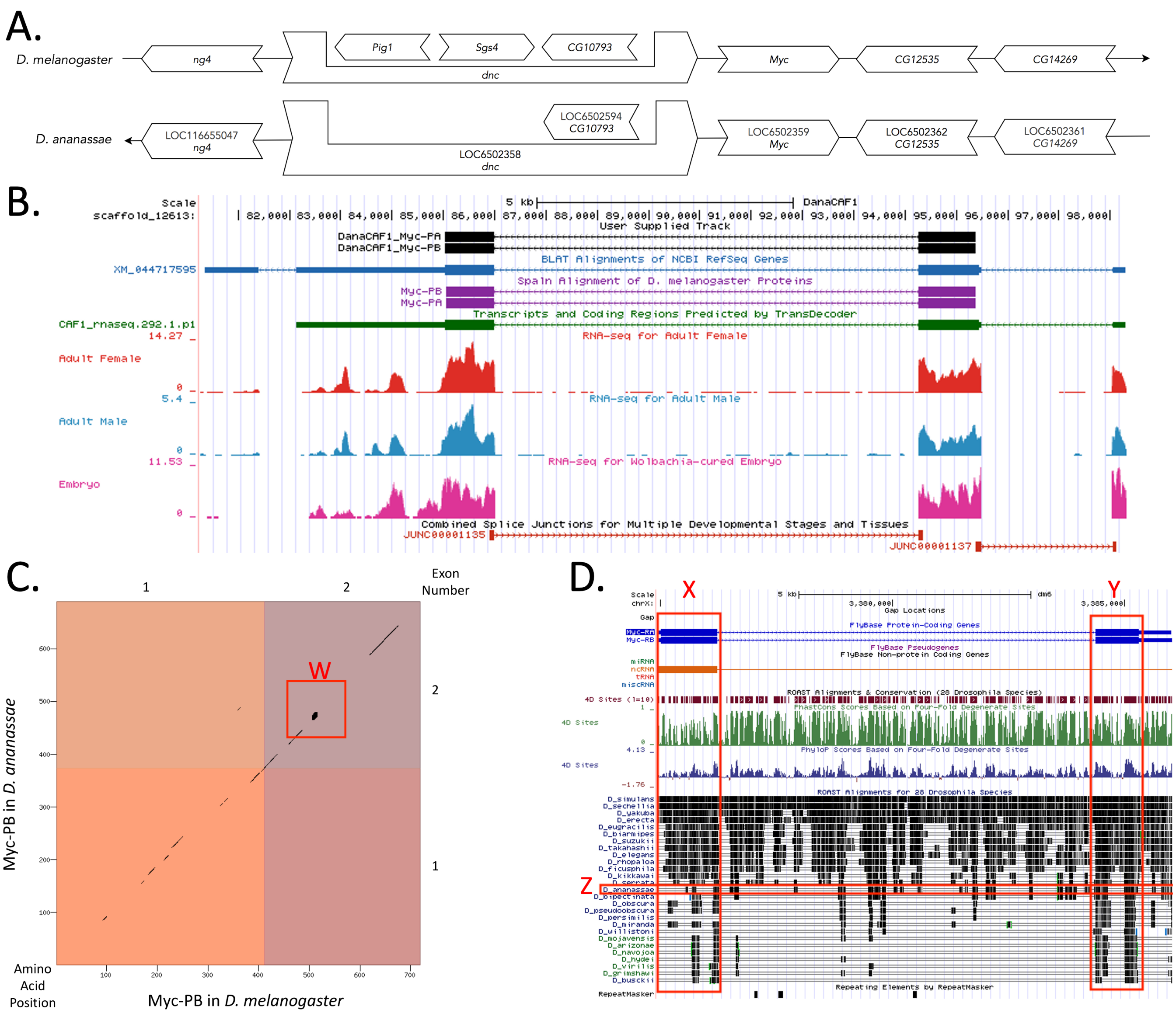

Figure 1. Genomic neighborhood and gene model for *Myc* in *Drosophila ananassae*:

(A) Synteny comparison of the genomic neighborhoods for *Myc* in *Drosophila melanogaster* and *D. ananassae*. Thin underlying arrows indicate the DNA strand within which the target gene–*Myc*–is located in *D. melanogaster* (top) and *D. ananassae* (bottom). Thin arrow(s) pointing to the right indicate(s) that *Myc* is on the positive (+) strand in *D. ananassae* and *D.melanogaster*. The wide gene arrows pointing in the same direction as *Myc* are on the same strand relative to the thin underlying arrows, while wide gene arrows pointing in the opposite direction of *Myc* 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 *Myc* 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; 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). Splice junctions shown have a read-depth of >1000 supporting reads in red. **(C) Dot Plot of Myc-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. Tandem repeats of serine are present in both sequences of the second CDS represented by the red box, Box W. (D) The Conservation Track of 28 *Drosophila* Species compared to CDSs one and two of *D. melanogaster* Myc-RA and Myc-RB contains many regions having lack of sequence similarity (vertical red boxes, Box X and Y; *D. ananassae* is highlighted in the horizontal red box, Box Z).

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 D. melanogaster genes in the insulin signaling pathway (ISP).

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* (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.

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).

"*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 2006; 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, submitted).

We propose a gene model for the *D. ananassae* ortholog of the *D. melanogaster* Myc (*Myc*) gene. The genomic region of the ortholog corresponds to the uncharacterized protein LOC6502359 (RefSeq accession XP_044573530.1) in the dana_caf1 Genome Assembly of *D. ananassae* (GenBank Accession: GCA_000005115.1; Drosophila 12 Genomes Consortium 2007). This model is based on RNA-Seq data from *D. ananassae* (SRP006203, SRP007906; PRJNA257286, PRJNA388952 - Graveley et al., 2011*)* and *Myc* in *D. melanogaster* using FlyBase release FB2022_04 (GCA_000001215.4; Larkin et al., 2021; Gramates et al., 2022; Jenkins et al., 2022).

Myc acts downstream of the insulin signaling pathway, with Myc protein accumulating in response to insulin through transcriptional and post-transcriptional mechanisms (Parisi et al., 2011), resulting in the activation of genes involved in anabolic processes that promote cell growth (Terakawa et al., 2022). *Myc* encodes a basic helix-loop-helix transcription factor in *Drosophila melanogaster* that is homologous to vertebrate *Myc* proto-oncogenes (Gallant et al., 1996). In *Drosophila melanogaster*, *Myc* transcriptionally regulates a wide range of genes, including those that influence cell growth and metabolism (Teleman et al., 2008; Gallant 2013).

Synteny

The reference gene, *Myc,* occurs on chromosome X in *D. melanogaster* and is flanked upstream by *new glue 4* (*ng4*), and *dunce* (*dnc*), which nests *Pre-intermoult gene 1* (*Pig1*), *salivary gland secretion 4* (*Sgs4*) and *CG10793*. *Myc* is flanked downstream by *CG12535* and *CG14269*. The *tblastn* search of *D. melanogaster* Myc-PB (query) against the *D. ananassae* (GenBank Accession: GCA_000005115.1) Genome Assembly (database) placed the putative ortholog of *Myc* within scaffold_12613 (CH902663.1) at locus LOC6502359 (XP_044573530.1)— with an E-value of 1e-42 and a percent identity of 36.30%. Furthermore, the putative ortholog is flanked upstream by LOC116655047 (XP_032307915.1) and LOC6502358 (XP_032307954.2) which nests LOC6502594 (XP_032307960.1) and correspond to *ng4*, *dnc*, and *CG10793* in *D. melanogaster* (E-value: 4e-16, 0.0, and 0.0; identity:71.43%, 84.49% and 80.13%, respectively, as determined by *blastp*; Figure 1A, Altschul et al., 1990). The putative ortholog of *Myc* is flanked downstream by LOC6502362 (XP_001967532.1) and LOC6502361 (XP_001967530.1), which correspond to *CG12535* and *CG14269* in *D. melanogaster* (E-value: 7e-45 and

1e-104; identity: 47.80% and 83.16%, respectively, as determined by *blastp*). The putative ortholog assignment for *Myc* in *D. ananassae* is supported by the following evidence: The genes surrounding the *Myc* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is nearly completely conserved, so we conclude that LOC6502359 is the correct ortholog of *Myc* in *D. ananassae* (Figure 1A).

Protein Model

Myc in *D. ananassae* has one unique protein-coding isoforms (Figure 1B), encoded by mRNA isoforms *Myc-RB* and *Myc-RA* that differ in their UTRs, and contain two CDSs. Relative to the ortholog in *D. melanogaster*, the RNA CDS number and protein isoform count is conserved. The sequence of Myc-PB in *D. ananassae* has 44.38% identity (E-value: 9e-77) with the protein-coding isoform Myc-PB in *D. melanogaster*, as determined by *blastp* (Figure 1C). Box W in red highlights tandem repeats of serine in CDS two, shown in Figure 1C. Coordinates of this curated gene model are stored by NCBI at GenBank/BankIt (BK064666 and BK064667). These data are also archived in the CaltechDATA repository (see “Extended Data” section below).

Special characteristics of the protein model

Regions of low conservation: Lack of sequence similarity in CDSs one and two of *Myc-RA* and *Myc-RB* is displayed in the 28 *Drosophila* Species Conservation track (Figure 1D) within the vertical red boxes (X and Y). The most obvious lack of sequence similarity primarily exists at the start of the first CDS and in the middle of CDS two in many *Drosophila* species including *D. ananassae* which is indicated by the horizontal red box (Figure 1D). This lack of sequence similarity is likely due to the divergence of the species *D. ananassae* from *D. melanogaster.*

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 (*D. ananassae* (GenBank Accession: GCA_000005115.1) 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. 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: We would like to thank Wilson Leung for developing and maintaining the technological infrastructure that was used to create this gene model and Laura K. Reed for overseeing the project. Also, thank you to Madeline Gruys and Logan Cohen for assistance in updating the manuscript to the current template. 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: GFF, FASTA, and PEP sequences for Myc in D ananassae. Resource Type: Dataset. File: DanaCAF1_Myc.zip. DOI: 10.22002/0c391-eeh07

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

Bock IR, Wheeler MR. 1972. The Drosophila melanogaster species-group. The University of Texas Publication. 7213: 1-102.

Doleschall CL. 1858. Derde bijdrage tot de kennis der dipterologische fauna van Nederlandsch Indië. Natuurkundig tijdschrift voor Nederlandsch Indië. 17: 73.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218. PubMed ID: 17994087

Gallant P. 2013. Myc Function in Drosophila. Cold Spring Harbor Perspectives in Medicine 3: a014324-a014324. PubMed ID: 24086064

Gallant P, Shiio Y, Cheng PF, Parkhurst SM, Eisenman RN. 1996. Myc and Max homologs in Drosophila. Science 274(5292): 1523-7. PubMed ID: 8929412

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

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

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 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 ofDrosophila ananassae. Genetica 20: 458-516. DOI: 10.1007/bf01531779

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., Rele, CP. 2024. Gene model for the ortholog of Ilp5 in Drosophila ananassae (submitted).

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

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. Nat Rev Genet 17(12): 758-772. PubMed ID: 27773922

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

Parisi F, Riccardo S, Daniel M, Saqcena M, Kundu N, Pession A, et al., Bellosta P. 2011. Drosophila insulin and target of rapamycin (TOR) pathways regulate GSK3 beta activity to control Myc stability and determine Myc expression in vivo. BMC Biol 9: 65. PubMed ID: 21951762

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

Teleman AA, Hietakangas V, Sayadian AC, Cohen SM. 2008. Nutritional control of protein biosynthetic capacity by insulin via Myc in Drosophila. Cell Metab 7(1): 21-32. PubMed ID: 18177722

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

Terakawa A, Hu Y, Kokaji T, Yugi K, Morita K, Ohno S, et al., Kuroda S. 2022. Trans-omics analysis of insulin action reveals a cell growth subnetwork which co-regulates anabolic processes. iScience 25(5): 104231. PubMed ID: 35494245

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 Reed. ,Supported by National Science Foundation (United States) 1915544 to Reed.

Author Contributions: Abigail Myers: formal analysis, validation, writing - original draft, writing - review editing. Alexa Hoffman: formal analysis, writing - review editing. Mindy Natysin: formal analysis, writing - review editing. Andrew M Arsham: supervision, writing - review editing. Joyce Stamm: supervision, writing - review editing. Jeffrey S. Thompson: writing - original draft. 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: David Molik, Sebastian Sorge, GEP Review Panel 2

Nomenclature Validated By: Anonymous

History: Received May 10, 2023 **Revision Received** December 22, 2023 **Accepted** November 28, 2024 **Published Online** November 30, 2024 **Indexed** December 14, 2024

Citation: 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*. microPublication Biology. 10.17912/micropub.biology.000856