microPublication BIOLOGY

8/3/2024 - *Open Access*

Gene model for the ortholog of *Myc* in *Drosophila eugracilis*

Megan E. Lawson[1], Alexa Hoffman[2], Isabel G. Wellik[3], Jeffrey S. Thompson[3], Joyce Stamm[2], Chinmay P. Rele[1§]

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

[2]University of Evansville, Evansville, IN USA

[3]Denison University, Granville, OH USA

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

Abstract

Gene model for the ortholog of Myc (*Myc*) in the *D. eugracilis* Apr. 2013 (BCM-HGSC/Deug_2.0) (DeugGB2) Genome Assembly (GenBank Accession: GCA_000236325.2) of *Drosophila eugracilis*. 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.

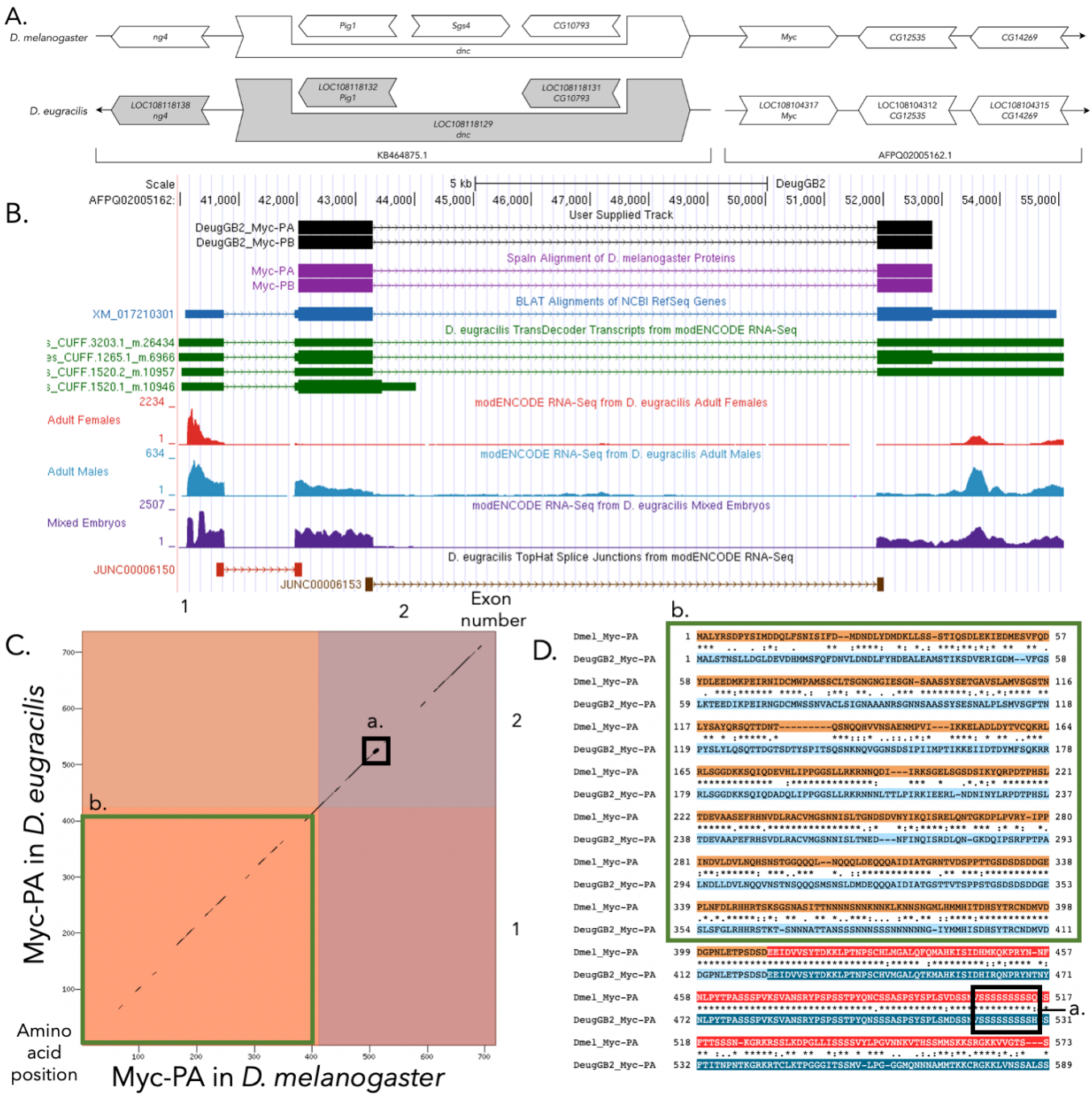

Figure 1.

(A) **Synteny of genomic neighborhood of *Myc* in *D. melanogaster* and *D. eugracilis*.** Gene arrows pointing in the same direction as target gene in both *D. eugracilis* and *D. melanogaster* are on the same strand as the target gene; gene arrows pointing in the opposite direction are on the opposite strand. The thin underlying arrows pointing to the right indicate that *Myc* and the downstream genes are on the + strand; the arrow pointing to the left indicates that the upstream genes are on the – strand in *D. eugracilis*. White arrows in *D. eugracilis* indicate the locus ID and the orthology to the corresponding gene in *D. melanogaster*, and gray arrows indicate that the orthologous genes upstream of *Myc* in *D. melanogaster* are found on a different scaffold in *D. eugracilis* than *Myc*. The brackets beneath the local synteny diagram for *D. eugracilis* show which scaffold each gene is found on. The gene names given in the *D. eugracilis* gene arrows indicate the orthologous gene in *D. melanogaster*, while the locus identifiers are specific to *D. eugracilis*. (B) **Gene Model in UCSC Track Hub** (Raney et al. 2014): the gene model in *D. eugracilis* (black), Spaln of D. melanogaster Proteins (purple, alignment of RefSeq proteins from *D. melanogaster*), BLAT alignments of NCBI RefSeq Genes (blue, alignment of RefSeq genes for *D. eugracilis*), RNA-seq

from female (red), male (blue), and mixed embryos (purple) (alignment of Illumina RNA-seq reads from *D. eugracilis*), and Transcripts (green) including coding regions predicted by TransDecoder and Splice Junctions Predicted by regtools using D. eugracilis RNA-seq (Chen *et al.*, 2014; PRJNA63469). Note that there is no measured expression of the first CDS of *Myc* (Flybase ID: 1_12880_0) in females (Flybase IDs from FB2022_03; Larkin *et al.*, 2021). Splice junctions shown have a minimum read-depth of 702 with 500-999 and >1000 supporting reads in brown and red respectively. The custom gene model (User Supplied Track) is indicated in black with CDSs depicted by boxes and introns by narrow lines (arrows indicate direction of transcription). (C) **Dot Plot of Myc-PA in *D. melanogaster* (*x*-axis) vs. the orthologous peptide in *D. eugracilis* (*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. The gaps in the dot plot indicate regions with low sequence similarity. The region within the black box (box a) contains a tandem repeat in CDS 2 that is conserved across both *D. melanogaster and D. eugracilis*. The green box (box b) highlights low sequence similarity in CDS one (D) **Idiosyncrasies in the protein alignment.** CDS one, which is boxed in green (box b), has many segments with low amino acid sequence similarity between *D. melanogaster* and *D. eugracilis*. The black box (box a) indicates a tandem repeat in the second CDS.

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).

"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)." (Myers et al., 2024).

"*D. eugracilis* (NCBI:txid29029) is part of the *melanogaste*r species group within the subgenus *Sophophora* of the genus *Drosophila* (Pélandakis et al., 1993). It was first described as *Tanygastrella gracilis* by Duda (1924) and revised to *Drosophila eugracilis* by Bock and Wheeler (1972). *D. eugracilis* is found in humid tropical and subtropical forests across southeast Asia (https://www.taxodros.uzh.ch, accessed 1 Feb 2023)." (Morgan et al., 2022).

The model presented here is the ortholog of *Myc* in the Apr. 2013 (BCM-HGSC/Deug_2.0) assembly of *D. eugracilis* (GCA_000236325.2) and corresponds to the *Gnomon Peptide ID (XP_017065790.1)* predicted model in *D. eugracilis (LOC108104317)*. This gene model is based on RNA-seq data from *D. eugracilis* (Chen et al., 2014; PRJNA63469) and the *Myc* (Drosophila 12 Genomes Consortium et al., 2007; *GCA_000001215.4*) in *D. melanogaster* from FB2023_03 (GCA_000001215.4; Larkin et al., 2021; Gramates et al., 2022).

Gene and species details can be found in the description above.

Synteny

Myc occurs on chromosome X in *D. melanogaster* and is flanked by *ng4* and *dnc* upstream and *CG12535* and *CG14269* downstream. The upstream *dnc* gene in *D. melanogaster* nests *Pig1*, *Sgs4*, and *CG10793*. We determined that the putative ortholog of *Myc* is found on the AFPQ02005162.1 scaffold in *D. eugracilis* (GB2 assembly GCA_000236325.2) with LOC108104317 (XP_017065790.1) (via *tblastn* search with an e-value of 0.0 and percent identity of 68.18%). It is flanked downstream by LOC108104312 (XP_017065785.1) and LOC108104315 (XP_017065789.1), which correspond to *CG12535* and *CG14269* *in D. melanogaster* with e-values of 7e-72 and 6e-110 respectively, and percent identities of 55.83% and 79.47% respectively, as determined by *blastp* (Figure 1A, Altschul et al., 1990). *Myc* is the first gene on the AFPQ02005162.1 scaffold in *D. eugracilis*, so there are no upstream genes to analyze for local synteny. However, *blastp* results indicated that the orthologs of genes upstream of *Myc* in *D. melanogaster* are located on scaffold KB464875.1 in *D. eugracilis* with LOC108118138 (XP_017086187.1), LOC108118129 (XP_041674349.1), LOC108118132 (XP_017086181.1), and LOC108118131 (XP_017086180.1), which correspond to *ng4*, *dnc*, *Pig1*, and *CG10793* with e-values of 9e-22, 0.0, 2e-77, and 0.0, respectively, and percent identities of 82.76%, 95.42%, 58.33%, and 87.81% respectively, as determined by *blastp* (Figure 1A, Altschul et al., 1990). Local synteny was conserved within this part of the neighborhood in *D. eugracilis* as well, with the exception of *Sgs4*, for which an ortholog in *D. eugracilis* could not be located. We believe this is the correct ortholog assignment for *Myc* in *D. eugracilis* because all of the *BLAST* hits had very low e-values and were the best *BLAST* result by a wide margin, and because local synteny is well-conserved throughout the genomic neighborhood.

Protein Model

Myc in *D. eugracilis* has one unique protein coding isoform encoded by mRNAs Myc-RA and Myc-RB, which differ in their UTRs (Figure 1B). Myc-PA/Myc-PB contain two protein coding CDSs. This is the same relative to the ortholog in *D. melanogaster.* The sequence of Myc in *D. eugracilis* has 67.0% identity with *Myc* in *D. melanogaster* as determined by *blastp* (Figure 1C). This is more divergence between the sequence than one would expect, considering how closely related *D. melanogaster* and *D. eugracilis* are. The coordinates of the curated gene models can be found in NCBI at GenBank/BankIt using the accessions BK063012 and BK063013. These data are also available in Extended Data files below, which are archived in CaltechData.

Special characteristics of the protein model

Tandem repeat in CDS two

There is a tandem repeat present in CDS two in *Myc*, which is shown in figures C and D in black box a. Specifically, the repeat is made up of nine uninterrupted Serine amino acids present in the protein sequence.

Low sequence similarity in CDS one

There are many regions in CDS one of *Myc* that have very low conservation of their amino acid sequences between the two species. This is pictured in the green boxes (box b) in figures C and D.

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. eugracilis* (GCA_000236325.2)) 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 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 and Logan Cohen 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.

Extended Data

Description: GFF, FASTA, and PEP of the model. Resource Type: Model. File: Deug_myc.zip. DOI: 10.22002/wtccf-p2596

Description: response to editorial pre-review. Resource Type: Text. File: DeugGB2_Myc_Review.docx. DOI: 10.22002/c7079-pkh57

References

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

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

Chen ZX, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al., Richards S. 2014. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res 24: 1209-23. PubMed ID: 24985915

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: 203-18. PubMed ID: 17994087

Duda, O. (1924). Revision der Europaischen Arten der Gattung Drosophila Fallen (Dipteren). *Ent. Medd.* **14**: 246--313. PubMed ID: null

Gallant P. 2013. Myc function in Drosophila. Cold Spring Harb Perspect Med 3: a014324. PubMed ID: 24086064

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

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. 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: 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

Johnston LA, Prober DA, Edgar BA, Eisenman RN, Gallant P. 1999. Drosophila myc regulates cellular growth during development. Cell 98(6): 779-90. PubMed ID: 10499795

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

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: D899-D907. PubMed ID: 33219682

Leone G, DeGregori J, Sears R, Jakoi L, Nevins JR. 1997. Myc and Ras collaborate in inducing accumulation of active cyclin E/Cdk2 and E2F. Nature 387: 422-6. PubMed ID: 9163430

Morgan A, Kiser CA, Bronson I, Lin H, Guillette N, McMahon R, et al., Rele CP. 2022. Drosophila eugracilis - Akt. MicroPubl Biol 2022. PubMed ID: 35856017

8/3/2024 - *Open Access*

Myers A., Hoffmann A., Natysin M., Arsham A.M, Stamm J., Thompson J.S., Rele C.P. 2024. Gene model for the ortholog *Myc* in *Drosophila ananassae*, *microPublication Biology*, submitted. PubMed ID: null

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: 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

Pélandakis M, Solignac M. 1993. Molecular phylogeny of Drosophila based on ribosomal RNA sequences. J Mol Evol 37: 525-43. PubMed ID: 8283482

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: 1003-5. PubMed ID: 24227676

Rele CP, Sandlin KM, Leung W, Reed LK. 2022. Manual annotation of Drosophila genes: a Genomics Education Partnership protocol. F1000Research 11: 1579. PubMed ID: null

Steiger D, Furrer M, Schwinkendorf D, Gallant P. 2008. Max-independent functions of Myc in Drosophila melanogaster. Nat Genet 40: 1084-91. PubMed ID: 19165923

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

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: 104231. PubMed ID: 35494245

Funding: This material is based upon work supported by the National Science Foundation under Grant No. IUSE-1915544 to LKR and the National Institute of General Medical Sciences of the National Institutes of Health Award R25GM130517 to LKR. The Genomics Education Partnership is fully financed by Federal moneys. The content is solely the responsibility of the authors and does not necessarily represent the official views of 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: Megan E. Lawson: formal analysis, validation, writing - original draft, writing - review editing. Alexa Hoffman: formal analysis, writing - review editing. Isabel G. Wellik: formal analysis, writing - review editing. Jeffrey S. Thompson: supervision, writing - review editing. Joyce Stamm: supervision, writing - review editing. Chinmay P. Rele: data curation, formal analysis, methodology, project administration, software, supervision, validation, visualization, writing - review editing.

Reviewed By: Sebastian Sorge

Nomenclature Validated By: Anonymous

History: Received July 1, 2023 **Revision Received** July 22, 2024 **Accepted** July 29, 2024 **Published Online** August 3, 2024 **Indexed** August 17, 2024

Citation: Lawson, ME; Hoffman, A; Wellik, IG; Thompson, JS; Stamm, J; Rele, CP (2024). Gene model for the ortholog of *Myc* in *Drosophila eugracilis*. microPublication Biology. 10.17912/micropub.biology.000912