microPublication BIOLOGY

12/18/2024 - *Open Access*

Gene model for the ortholog of *S6k* in *Drosophila yakuba*

Grace Keirn[1], Cole A. Kiser[1], Leon F. Laskowski[2], Jordan Hensley[3], Rachel Mortan[1], Thuy Nguyen[4], Shallee T. Page[5], Charles Du[6], Jeroen T. F. Gillard[7], Anya Goodman[4], James J. Youngblom[3], 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]California State University Stanislaus, Turlock, CA USA

[4]California Polytechnic State University, San Luis Obispo, CA USA

[5]Franklin Pierce University, Rindge, MA, USA

[6]Montclair State University, Montclair, NJ USA

[7]California State University, Bakersfield CA USA

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

Abstract

Gene model for the ortholog of *Ribosomal protein S6 kinase* (*S6k*) in the Dyak_CAF1 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.

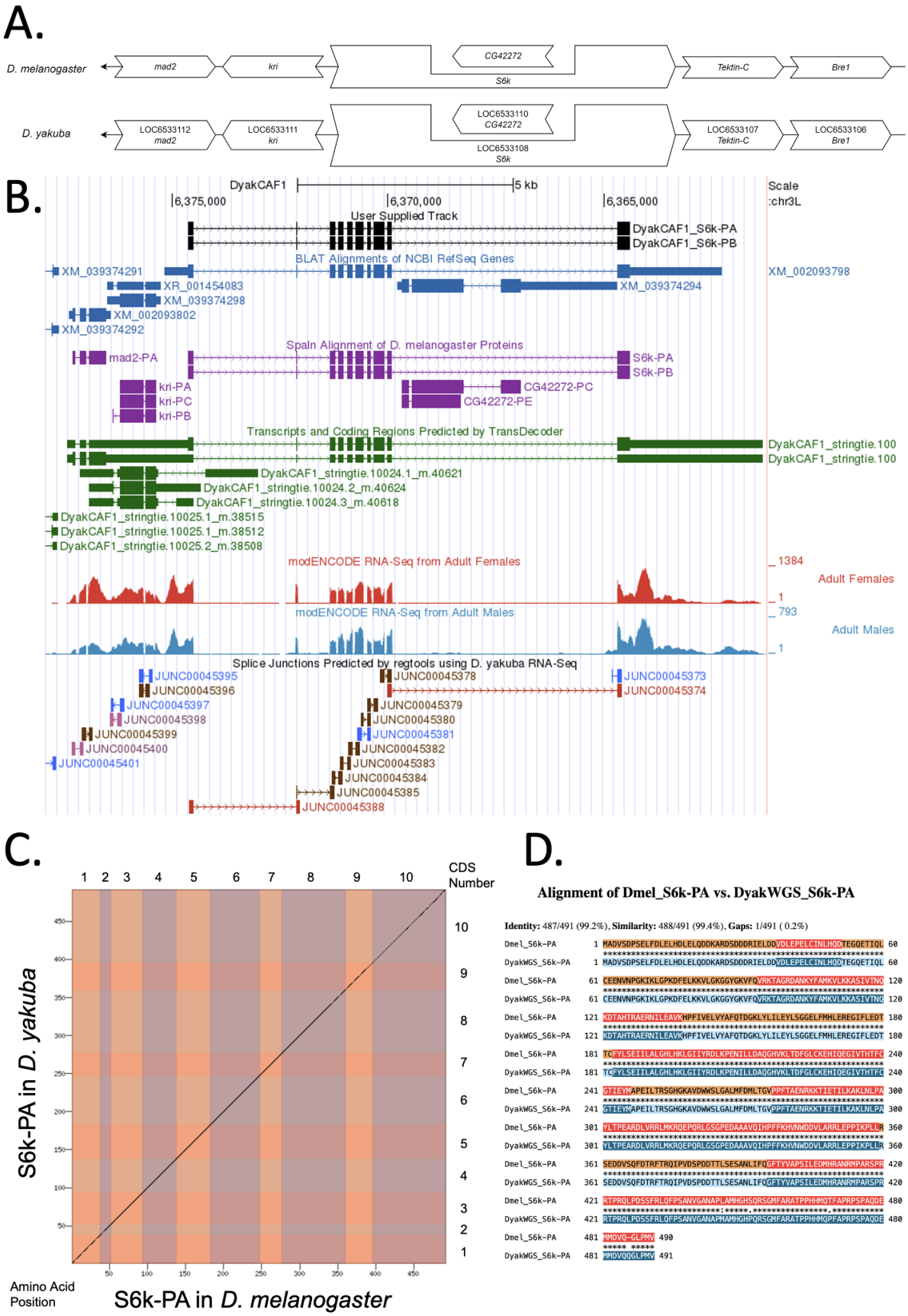

Figure 1. Genomic neighborhood and gene model for *S6k* in *Drosophila yakuba*.:

(A) Synteny comparison of the genomic neighborhoods for *S6k* in *Drosophila melanogaster* and *D. yakuba*. Thin underlying arrows indicate the DNA strand within which the target gene–*S6k*–is located in *D. melanogaster* (top) and *D. yakuba* (bottom) genomes. Thin arrows pointing to the left indicate that *S6k* is on the negative (-) strand in *D. yakuba* and *D. melanogaster*. The wide gene arrows pointing in the same direction as *S6k* are on the same strand relative to the thin underlying arrows, while wide gene arrows pointing in the opposite direction of *S6k* 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 *S6k* 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). Splice junctions shown have a read-depth of 10-49, 100-499, 500-999, >1000 supporting reads in blue, pink, brown, and red, respectively. **(C) Dot Plot of S6k-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 CDS are also highlighted with alternating colors. **(D) Protein alignment between *D. melanogaster* S6k-PA and its putative ortholog 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.

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 wide-spread in sub-Saharan Africa and Madagascar (Lemeunier et al., 1986; https://www.taxodros.uzh.ch, accessed 1 Feb 2023; Markow and O'Grady 2005) where figs served as a primary host along with other rotting fruits (Lachaise and Tsacas 1983)." (Koehler et al., 2024).

We propose a gene model for the *D. yakuba* ortholog of the *D. melanogaster Ribosomal protein S6 kinase* (*S6k*) gene. The genomic region of the ortholog corresponds to the uncharacterized protein LOC6533108 (RefSeq accession XP_002093834.1) in the Dyak_CAF1 Genome Assembly of *D. yakuba* (GenBank Accession: GCA_000005975.1). This model is based on RNA-

Seq data from *D. yakuba* (SRP006203) and *S6k* in *D. melanogaster* using FlyBase release FB2022_04 (GCA_000001215.4; Larkin et al., 2021).

Ribosomal protein S6 kinase (*S6k* aka p70S6K, FBgn0283472) is part of the insulin signaling pathway downstream of the target of rapamycin (*dTOR*) (Toker 2000), homologous to mammalian p70S6k (Watson et al., 1996). S6k is a serine/threonine kinase in *Drosophila melanogaster* and acts as a regulator of cell size (Montagne et al., 1999), as well as innate immunity and senescence (Fabian et al., 2021). The species summary can be found in the box above.

Synteny

The target gene, *S6k*, occurs on chromosome 3L in *D. melanogaster* and the gene *CG42272* (*CG42272*) nests within it. *S6k* is flanked upstream by *mad2* (*mad2*) and *krishah* (*kri*) and downstream by *Tektin C* (*Tektin-C*) and *Bre1* (*Bre1*). The *tblastn* search of *D. melanogaster* S6k-PA (query) against the *D. yakuba* (GenBank Accession: GCA_000005975.1) Genome Assembly (database) placed the putative ortholog of *S6k* within scaffold chromosome 3L (CM000159.2) at locus LOC6533108 (XP_002093834.1)— with an E-value of 3e-69 and a percent identity of 43.04%. Furthermore, the putative ortholog of *CG42272* (LOC6533110; XP_039230228.1), nests within LOC6533108 (E-value: 0.0; identity: 87.79%, as determined by *blastp*; Figure 1A, Altschul et al., 1990) and is flanked upstream by LOC6533112 (XP_002093838.1) and LOC6533111 (XP_039230232.1), which correspond to *mad2* and *kri* in *D. melanogaster* (E-value: 6e-153 and 0.0; identity: 96.62% and 98.08%, respectively, as determined by *blastp*). The putative ortholog of *S6k* is flanked downstream by LOC6533107 (XP_002093833.1) and LOC6533106 (XP_002093832.1), which correspond to *Tektin-C* and *Bre1* in *D. melanogaster* (E-value: 0.0 and 0.0; identity: 100.00% and 98.28%, respectively, as determined by *blastp*). The putative ortholog assignment for *S6k* in *D. yakuba* is supported by the following evidence: The genes surrounding the *S6k* ortholog are orthologous to the genes at the same locus in *D. melanogaster* and local synteny is completely conserved, supported by results generated from *blastp*, so we conclude that LOC6533108 is the correct ortholog of *S6k* in *D. yakuba* (Figure 1A).

Protein Model

S6k in *D. yakuba* has two mRNA isoforms (*S6k-RA; S6k-RB*) that encode one unique protein-coding isoform (S6k-PA and S6k-PB; Figure 1B). The gene contains ten CDSs. Relative to the ortholog in *D. melanogaster*, the RNA CDS number and protein isoform count are conserved. The sequence of S6k-PA in *D. yakuba* has 99.19% identity (E-value: 0.0) with the protein-coding isoform S6k-PA in *D. melanogaster*, as determined by *blastp* (Figure 1D). Coordinates of this curated gene model of S6k-PA and S6k-PB are stored by NCBI at GenBank/BankIt (accession **BK064475 and BK064476**, respectively**)**. These data are also archived in the CaltechDATA repository (see “Extended Data” section below).

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 (GenBank Accession: 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. 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. 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. 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.

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. Univ. Texas Publs Stud. Genet. 7(7213): 1-102. FBrf0024428.

Burla H. 1954. Zur Kenntnis der Drosophiliden der Elfenbeinkuste (Franzosisch West-Afrika). Revue suisse Zool. 61(Suppl.): 1-218. FBrf0009861.

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

Fabian DK, Fuentealba M, Dönertaş HM, Partridge L, Thornton JM. 2021. Functional conservation in genes and pathways linking ageing and immunity. Immun Ageing 18(1): 23. PubMed ID: 33990202

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, Savraj S 2009. Insulin/TOR signaling in growth and homeostasis: a view from the fly world. Int. J. Biochem. Cell Biol.. 41: 1006-1010. PubMed ID: 18992839

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

Koehler AC, Romo I, Le V, Romo I, Youngblom JJ, Hark AT, Rele CP, Reed LK. 2024. Gene model for the ortholog of *Glys* in *Drosophila yakuba*, *microPublication Biology*. DOI: 10.17912/micropub.biology.000983

Lachaise D, Tsacas L. 1983. Breeding-sites of tropical African Drosophilids. Ashburner, Carson, Thompson, 1981-1986. 3d: 21-332. FBrf0038884.

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 J, Tsacas L. Ashburner M. 1986. The melanogaster species group. Ashburner, Carson, Thompson, 1981-1986. e: 147-256. FBrf0043749.

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

Montagne J, Stewart MJ, Stocker H, Hafen E, Kozma SC, Thomas G. 1999. Drosophila S6 kinase: a regulator of cell size. Science 285(5436): 2126-9. PubMed ID: 10497130

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

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

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

Toker A. 2000. Protein kinases as mediators of phosphoinositide 3-kinase signaling. Mol Pharmacol 57(4): 652-8. PubMed ID: 10727509

Watson KL, Chou MM, Blenis J, Gelbart WM, Erikson RL. 1996. A Drosophila gene structurally and functionally homologous to the mammalian 70-kDa s6 kinase gene. Proc Natl Acad Sci U S A 93(24): 13694-8. PubMed ID: 8942996

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: Grace Keirn: formal analysis, validation, writing - original draft, writing - review editing. Cole A. Kiser: formal analysis, validation, writing - original draft, writing - review editing. Leon F. Laskowski: formal analysis, validation, writing - original draft, writing - review editing. Jordan Hensley: formal analysis, writing - review editing. Rachel Mortan: formal analysis, writing - review editing. Thuy Nguyen: formal analysis, writing - review editing. Shallee T. Page: writing - original draft, writing - review editing. Charles Du: writing - original draft, writing - review editing. Jeroen T. F. Gillard: writing - original draft, writing - review editing. Anya Goodman: supervision, writing - review editing. James J. Youngblom: 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, writing - review editing.

Reviewed By: John Stanga, Anonymous

Nomenclature Validated By: Anonymous

History: Received October 4, 2023 **Revision Received** December 10, 2024 **Accepted** December 11, 2024 **Published Online** December 18, 2024 **Indexed** January 1, 2025

Citation: Keirn, G; Kiser, CA; Laskowski, LF; Hensley, J; Mortan, R; Nguyen, T; et al.; Reed, LK (2024). Gene model for the ortholog of *S6k* in *Drosophila yakuba*. microPublication Biology. 10.17912/micropub.biology.001018