

## ***Drosophila kikkawai* ortholog of the *D. melanogaster* Muller D element *ash1***

Mia Mo<sup>1§</sup>, Destiny Sabb<sup>2</sup>, Larissa LoBello<sup>1</sup>, Kayla Chambers<sup>3</sup>, Kacie Kershaw<sup>4</sup>, Cameron Welles<sup>5</sup>, Joshua Kurucz<sup>5</sup>, Caleb Rhyne<sup>3</sup>, Angel Nichols<sup>3</sup>, John Stanga<sup>6</sup>, James E. J. Bedard<sup>7</sup>, Cindy Arrigo<sup>8</sup>

<sup>1</sup>Washington University in St. Louis, St. Louis, Missouri, United States

<sup>2</sup>New Jersey City University, Jersey City, New Jersey, United States

<sup>3</sup>Mercer University, Macon, Georgia, United States

<sup>4</sup>St. Joseph's University, Philadelphia, Pennsylvania, United States

<sup>5</sup>University of the Fraser Valley, Abbotsford, British Columbia, Canada

<sup>6</sup>Biology, Mercer University, Macon, Georgia, United States

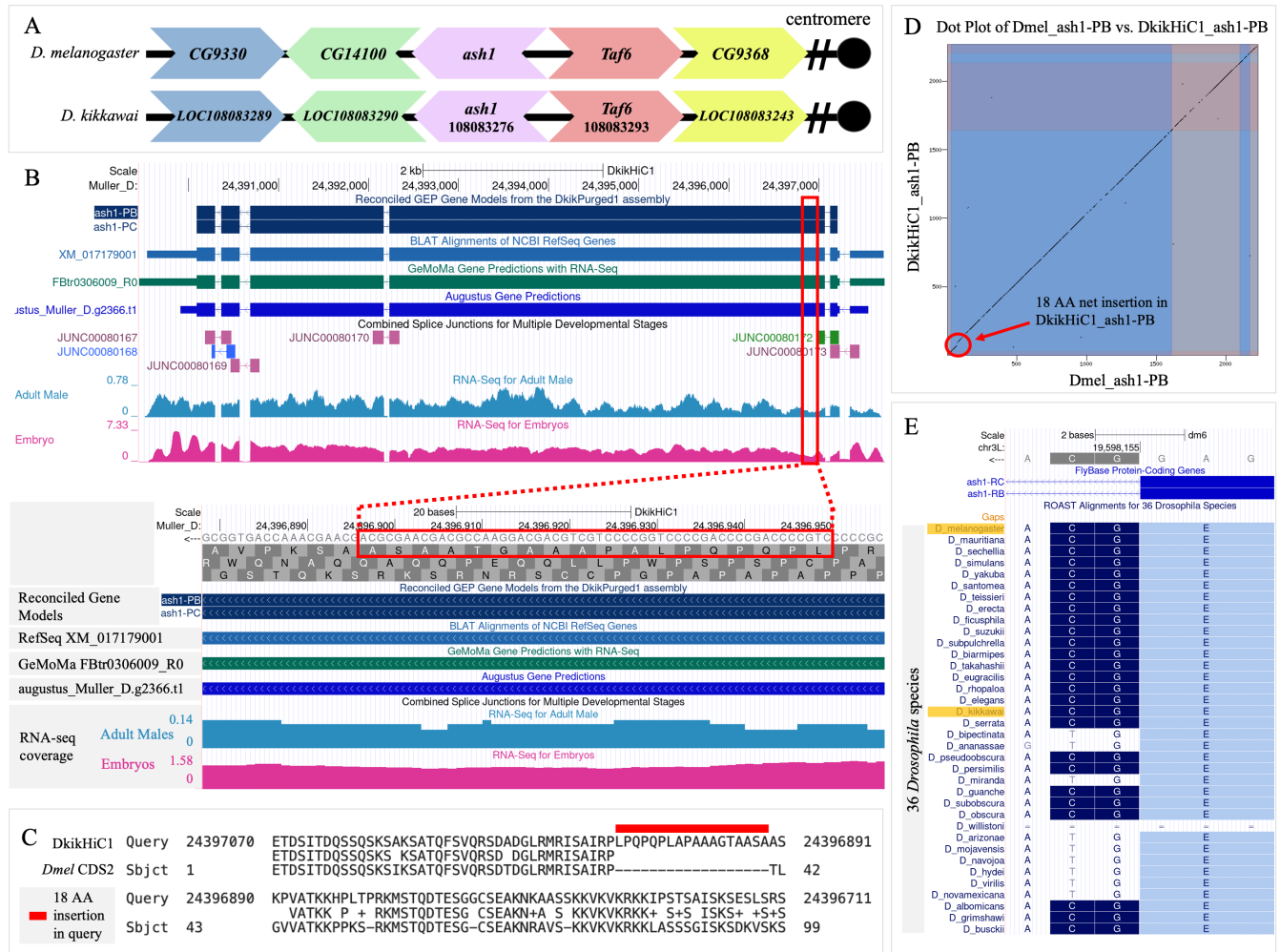
<sup>7</sup>Biology, University of the Fraser Valley, Abbotsford, British Columbia, Canada

<sup>8</sup>Biology, New Jersey City University, Jersey City, New Jersey, United States

<sup>§</sup>To whom correspondence should be addressed: m.mo@wustl.edu

### **Abstract**

The [Drosophila kikkawai](#) feature with Gene ID 108083276 was determined to be an ortholog of [Drosophila melanogaster](#) absent, small, or homeotic discs 1 ([ash1](#)). Two isoforms, ash1-PB and ash1-PC, were constructed on the [D. kikkawai](#) Muller D element using the GEP annotation protocol. The second coding exon of [D. kikkawai ash1](#) includes an insertion translated into 18 additional amino acids compared to the [D. melanogaster](#) protein and is supported by RNA-Seq coverage, the lack of splice junction predictions, and multiple gene predictors. The first intron in both isoforms of [D. kikkawai ash1](#) contains a well conserved non-canonical GC splice site.



**Figure 1.** *D. kikkawai* feature with Gene ID 108083276 was determined to be an ortholog of *D. melanogaster ash1* and possesses a net 18 AA insertion relative to the *D. melanogaster ash1* protein:

**(A) Gene neighborhood of *ash1* in *Drosophila melanogaster* (top) and *Drosophila kikkawai* (bottom).** The large chevron arrows are labeled with gene symbols. For *D. melanogaster*, these labels correspond to the FlyBase gene symbols. For *D. kikkawai*, the labels correspond to the gene symbols in the NCBI Gene database. For *D. kikkawai* genes with published gene symbols, the label begins with the gene symbol (e.g., *ash1*) rather than “LOC”, followed by the NCBI Gene ID (e.g., 108083276). The direction of the chevron arrows corresponds to the orientation of the gene relative to the centromere (black dot). The *ash1* gene is coded on the minus strand in *D. melanogaster* and *D. kikkawai*. Orthologous genes are denoted with the same color chevron arrow. **(B) Genome Browser image of *ash1* gene model(s) on the *D. kikkawai* DkikHiC1 assembly with evidence tracks.** The dark blue gene models at the top of the image represent the final gene models for *ash1*-PB and *ash1*-PC, which have identical coding regions. The provided evidence tracks include the BLAT alignment of the *D. kikkawai* RefSeq mRNA [XM\\_017179001](#), two gene predictions from GeMoMa and Augustus, mapped splice junctions, and overall RNA-Seq coverage from adult males and embryos. The inset shows the region in the second coding exon (CDS2) of the *D. kikkawai ash1* gene (red box) that contains an insertion of 18 amino acids (AA) compared to the *D. melanogaster* CDS2. Note the continuous RNA-Seq coverage, the absence of compatible splice donor and acceptor sites that satisfy the biological constraint for minimum intron size, and the absence of detectable RNA-Seq junction fragments in the combined splice junctions track. **(C) BLASTx alignment output comparing translated DNA sequence from CDS2 in *D. kikkawai* (Query; DkikHiC1) to the AA sequence of CDS2 in *D. melanogaster* (Subject; *Dmel* CDS2).** The alignment of the first 99 AA from CDS2 is shown where the match line in the middle denotes similarity between the query and subject input sequences. The dashes in the subject sequence correspond to gaps in the alignment, indicating an insertion in the query sequence. The bold red line above the alignment indicates the predicted 18 AA insertion in CDS2 of the *D. kikkawai* protein relative to the *D. melanogaster ash1* protein. **(D) Dot plot comparing AA sequences of *ash1*-PB in *D. melanogaster* (x-axis) and *D. kikkawai* (y-axis).** Black dots in the dot plot correspond to identical residues between the sequences on the x and y axes. Strong

similarity is indicated by a continuous black line. The alternating-colored boxes indicate exon boundaries. The red circle demarcates the discontinuity displayed in panels B and C. **(E) ROAST alignment of 36 *Drosophila* species shows level of conservation of the non-canonical GC splice donor site for the first intron.** The ROAST alignment depicts conservation of a non-canonical GC donor site across 26 of the 36 *Drosophila* species. The light blue boxes indicate the end of CDS1. The non-canonical GC site is highlighted in dark blue. The first intron of the *D. kikkawai* and *D. melanogaster ash1* genes both contain the GC donor site. *D. kikkawai* and *D. melanogaster* are highlighted in yellow.

## Description

The *Drosophila melanogaster* absent, small, or homeotic discs 1 (*ash1*) gene (FBgn0005386) is found on chromosome 3L (Muller D element) and encodes a [histone H3]-lysine(36) N-dimethyltransferase (EC 2.1.1.357) which is involved in regulating transcription (Goodliffe et al., 2007). The *ash1* gene has been assigned to the SET DOMAIN LYSINE METHYLTRANSFERASES (KMT) gene group by FlyBase (FBgg0000274; Jiang et al., 2017; Mohan et al., 2011). The group encodes proteins with a variety of proposed molecular functions including histone methyltransferase activity and chromatin binding (Huang et al., 2017). Specifically, the *ash1* protein enables histone H3K36 methyltransferase activity (GO:0046975), H3K36 dimethyltransferase activity (GO:0140954) and histone H3K4 methyltransferase activity (GO:0042800) (Byrd & Shearn, 2003; Dorigi & Tamkun, 2013; Huang et al., 2017; Tanaka et al., 2007). The product of *ash1* is a part of the TrxG (Trithorax) protein group. Proteins in this group aid in keeping homeotic genes active by preventing PcG (Polycomb Group) proteins from silencing homeotic genes (Dorafshan et al., 2019). According to FlyBase (release FB2022\_06), *ash1* is orthologous to ASH1 like histone lysine methyltransferase (*ASH1L*) in humans with a DIOPT score of 11/14 (Hu et al., 2011; Öztürk-Çolak et al., 2024).

***D. kikkawai* feature with Gene ID 108083276 is the putative ortholog of *ash1*.** The ortholog assignment of *ash1* is supported by local synteny with *D. melanogaster*, protein sequence similarity, and experimental RNA-Seq evidence. *D. melanogaster ash1* is surrounded by the genes *CG9330* (FBgn0036888) and *CG14100* (FBgn0036889) downstream, and *Taf6* (FBgn0010417) and *CG9368* (FBgn0036890) upstream. The putative ortholog for *ash1* was determined to correspond to a feature with Gene ID 108083276 on the *D. kikkawai* Muller D element with coordinates that are consistent with the top hit of a tBLASTn (v2.15.0+) search with the *D. melanogaster ash1*-PB protein against the entire *D. kikkawai* DkikHiC1 (GenBank Assembly Accession: [GCA\\_030179895.1](https://www.ncbi.nlm.nih.gov/assembly/GCA_030179895.1)) assembly (Camacho et al., 2009). The match with the lowest E-value mapped to scaffold CM058225.1 (assigned to the Muller D element) between *LOC108083290* and *Taf6* (Gene ID: 108083293) with an overall identity of 84.08%, a coverage of 99%, and an E-value of 0.0 (i.e., E-value less than  $1 \times 10^{-180}$ ). The second best hit maps to scaffold CM058228.1 (X chromosome) with a substantially lower identity of 40.10%, a coverage of 65%, and a substantially greater E-value of  $7 \times 10^{-36}$ . *D. kikkawai ash1* is surrounded by genes with high sequence similarity to the *D. melanogaster* genes *CG9330* (Gene ID: 108083289), *CG14100* (Gene ID: 108083290), *Taf6* (Gene ID: 108083293) and *CG9368* (Gene ID: 108083243) (Figure 1A). The consistent local synteny with *D. melanogaster* supports the assignment of the feature with Gene ID 108083276 as *ash1* in *D. kikkawai*. Orthology was further supported by the BLASTp (v2.15.0+) protein alignment between the *D. melanogaster ash1* protein and the protein product of the final *D. kikkawai* model, which shows a sequence similarity of 89.4% and a percent identity of 83.7%. Nucleotide sequence data reported are available in the Third-Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: [BK068291](https://www.ncbi.nlm.nih.gov/assembly/GCA_030179895.1) and [BK068292](https://www.ncbi.nlm.nih.gov/assembly/GCA_030179895.1).

**Characterizing the net 18 AA insertion in the *D. kikkawai ash1* protein.** The *ash1* gene in *D. melanogaster* has two isoforms, *ash1*-PB and *ash1*-PC, with identical polypeptide sequence. These isoforms have five Coding DNA Sequences (CDS). Comparing the *D. kikkawai* and *D. melanogaster* gene models, both have the same number of CDSs (Figure 1B). However, the dot plot of the protein alignment depicts a discontinuity consistent with an insertion in the second CDS of *D. kikkawai ash1*-PB (Figure 1D). The cumulative change in this region is an 18 AA insertion relative to the *D. melanogaster ash1* protein. This extension is one of the contributors to the difference in the total protein length from 2,226 AA in *D. melanogaster ash1* to 2,269 AA in *D. kikkawai ash1* (Figure 1C). Due to continuous RNA-Seq coverage across this discontinuity in the alignment, and the lack of compatible splice donor and acceptor sites, the 18 AA difference is not a novel intron. Further analysis would be required to determine if the presence of these new amino acids alters the protein's function in *D. kikkawai* compared to *D. melanogaster*. However, we note that the insertion does not disrupt any of the identified protein domains [AWS domain, SET domain, Bromo adjacent homology (BAH) domain].

**Characterizing the noncanonical GC splice donor site in the first intron of *D. kikkawai ash1*.** In *D. melanogaster*, both the B and C isoforms of *ash1* have a noncanonical GC splice donor site after the end of the first CDS. In *D. kikkawai*, the non-canonical splice site is supported by the substantial decrease in RNA-Seq read coverage and a splice junction prediction derived from spliced RNA-Seq reads (JUNC00080172; score = 79). Based on the ROAST alignment (March 2008 release),

this GC donor site appears to be conserved in 26 out of 36 *Drosophila* species that are part of the whole genome multiple sequence alignment, including in *D. kikkawai* (Figure 1E). While 10 of the 36 species exhibit the canonical GT donor site (e.g., *D. ananassae*, *D. bipectinata*, etc.), the conservation of the GC donor site in closely related and more evolutionarily distant *Drosophila* species suggests that the common ancestor of *D. melanogaster* and *D. kikkawai* possessed the non-canonical GC splice donor site.

## Methods

The protocol used to annotate and reconcile the *ash1* gene model and neighboring gene models is described in Rele et al., 2023. The annotations are based on the annotated gene models for FlyBase release FB2022\_06 (*D. melanogaster* release 6.49) in the release 6 assembly (Hoskins et al., 2015). A mirror of the UCSC Genome Browser (v435) (Kent et al., 2002; Navarro Gonzalez et al., 2021) is maintained by the Genomics Education Partnership (GEP) at <https://gander.wustl.edu/>. Within the *D. kikkawai* Hi-C genome browser, evidence tracks that display the results of experimental data (e.g., RNA-Seq) and computational tools including tBLASTn, gene predictors, and BLAT were used to support the assignment of the *ash1* ortholog and annotate a gene model. The *D. kikkawai* RNA-Seq data was generated by the modENCODE project (Chen et al., 2014). The tBLASTn (v2.13.0+) results report the region of the genome with the highest similarity to *D. melanogaster* protein-coding sequences. The NCBI BLAST+ web interface uses v2.15.0+. Gene predictors, such as Augustus (v3.5.0) and GeMoMa (v1.9), provide computational predictions for the structures and locations of putative genes. BLAT (v37x1) alignments report the region of the genome with the highest similarity to *D. melanogaster* transcripts.

**Acknowledgements:** We would like to thank Wilson Leung for developing and maintaining the technological infrastructure that was used to create this gene model. We would also like to thank Dr. Christopher Shaffer for supervising Mia Mo in the reconciliation process and for his feedback on the manuscript.

## Extended Data

Description: Transcript, peptide, and generic feature format version 3 (GFF3) files for all isoforms (B and C) of *ash1* for DkikHiC1 assembly. Resource Type: Dataset. File: [DkikHiC1\\_ash1.zip](#). DOI: [10.22002/sdh1e-5wn19](https://doi.org/10.22002/sdh1e-5wn19)

## References

- Byrd KN, Shearn A. 2003. ASH1, a *Drosophila* trithorax group protein, is required for methylation of lysine 4 residues on histone H3. *Proceedings of the National Academy of Sciences* 100: 11535-11540. DOI: [10.1073/pnas.1933593100](https://doi.org/10.1073/pnas.1933593100)
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 10.1186/1471-2105-10-421. DOI: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421)
- Chen ZX, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al., Richards. 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Research* 24: 1209-1223. DOI: [10.1101/gr.159384.113](https://doi.org/10.1101/gr.159384.113)
- Dorafshan E, Kahn TG, Glotov A, Savitsky M, Walther M, Reuter G, Schwartz YB. 2019. Ash1 counteracts Polycomb repression independent of histone H3 lysine 36 methylation. *EMBO reports* 20: 10.15252/embr.201846762. DOI: [10.15252/embr.201846762](https://doi.org/10.15252/embr.201846762)
- Dorigi KM, Tamkun JW. 2013. The trithorax group proteins Kismet and ASH1 promote H3K36 dimethylation to counteract Polycomb group repression in *Drosophila*. *Development* 140: 4182-4192. DOI: [10.1242/dev.095786](https://doi.org/10.1242/dev.095786)
- Goodliffe JM, Cole MD, Wieschaus E. 2007. Coordinated regulation of Myc trans-activation targets by Polycomb and the Trithorax group protein Ash1. *BMC Molecular Biology* 8: 40. DOI: [10.1186/1471-2199-8-40](https://doi.org/10.1186/1471-2199-8-40)
- Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al., Celniker. 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Research* 25: 445-458. DOI: [10.1101/gr.185579.114](https://doi.org/10.1101/gr.185579.114)
- Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 12: 10.1186/1471-2105-12-357. DOI: [10.1186/1471-2105-12-357](https://doi.org/10.1186/1471-2105-12-357)
- Huang C, Yang F, Zhang Z, Zhang J, Cai G, Li L, et al., Zhu. 2017. Mrg15 stimulates Ash1 H3K36 methyltransferase activity and facilitates Ash1 Trithorax group protein function in *Drosophila*. *Nature Communications* 8: 10.1038/s41467-017-01897-3. DOI: [10.1038/s41467-017-01897-3](https://doi.org/10.1038/s41467-017-01897-3)
- Jiang F, Liu Q, Wang Y, Zhang J, Wang H, Song T, et al., Kang. 2017. Comparative genomic analysis of SET domain family reveals the origin, expansion, and putative function of the arthropod-specific *SmydA* genes as histone modifiers in insects.

GigaScience 6: 10.1093/gigascience/gix031. DOI: [10.1093/gigascience/gix031](https://doi.org/10.1093/gigascience/gix031)

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The Human Genome Browser at UCSC. *Genome Research* 12: 996-1006. DOI: [10.1101/gr.229102](https://doi.org/10.1101/gr.229102)

Mohan M, Herz HM, Smith ER, Zhang Y, Jackson J, Washburn MP, et al., Shilatifard. 2011. The COMPASS Family of H3K4 Methylases in *Drosophila*. *Molecular and Cellular Biology* 31: 4310-4318. DOI: [10.1128/MCB.06092-11](https://doi.org/10.1128/MCB.06092-11)

Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al., Kent. 2020. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research* 49: D1046-D1057. DOI: [10.1093/nar/gkaa1070](https://doi.org/10.1093/nar/gkaa1070)

Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al., Lovato. 2024. FlyBase: updates to the *Drosophila* genes and genomes database. *GENETICS* 227: 10.1093/genetics/iyad211. DOI: [10.1093/genetics/iyad211](https://doi.org/10.1093/genetics/iyad211)

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.126839.3](https://doi.org/10.12688/f1000research.126839.3)

Tanaka Y, Katagiri Zi, Kawahashi K, Kioussis D, Kitajima S. 2007. Trithorax-group protein ASH1 methylates histone H3 lysine 36. *Gene* 397: 161-168. DOI: [10.1016/j.gene.2007.04.027](https://doi.org/10.1016/j.gene.2007.04.027)

**Funding:** This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2114661 to Dr. Cindy Arrigo. The Genomics Education Partnership (GEP) is supported by the NSF under Grant No. 1915544 and National Institute of General Medical Sciences of the National Institutes of Health under award number R25GM130517 to the Genomics Education Partnership (<https://thegep.org/>). 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) 2114661 to Cindy Arrigo. ,Supported by National Science Foundation (United States) 1915544 to Laura Reed. ,Supported by National Institutes of Health (United States) R25GM130517 to Laura Reed.

**Author Contributions:** Mia Mo: formal analysis, writing - original draft, writing - review editing. Destiny Sabb: formal analysis, writing - review editing, writing - original draft. Larissa LoBello: writing - review editing, formal analysis. Kayla Chambers: formal analysis, writing - review editing. Kacie Kershaw: formal analysis, writing - review editing. Cameron Welles: formal analysis, writing - review editing. Joshua Kurucz: formal analysis, writing - review editing. Caleb Rhyne: formal analysis, writing - review editing. Angel Nichols: formal analysis, writing - review editing. John Stanga: supervision, writing - review editing. James E. J. Bedard: writing - review editing, supervision. Cindy Arrigo: writing - review editing, supervision, funding acquisition, project administration, conceptualization.

**Reviewed By:** Anonymous

**Nomenclature Validated By:** Anonymous

**History:** Received August 6, 2024 **Revision Received** November 22, 2024 **Accepted** November 21, 2024 **Published Online** November 22, 2024 **Indexed** December 6, 2024

**Copyright:** © 2024 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Citation:** Mo, M; Sabb, D; LoBello, L; Chambers, K; Kershaw, K; Welles, C; et al.; Arrigo, C (2024). *Drosophila kikkawai* ortholog of the *D. melanogaster* Muller D element *ash1*. *microPublication Biology*. [10.17912/micropub.biology.001322](https://doi.org/10.17912/micropub.biology.001322)