# Manual Annotation of Genes within *Drosophila* Species: the Genomics Education Partnership protocol

Chinmay P. Rele[1‡], Katie M. Sandlin[1], Wilson Leung[2], Laura K. Reed[1]

1. Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL 35487
2. Department of Biology, Washington University in St. Louis, St. Louis, MO 63130

‡ Corresponding author

Department of Biological Science, Box 870206

300 Hackberry Ln., University of Alabama

Tuscaloosa, AL 35487-0206

+1 (205) 348-0929

# Abstract

Annotating the genomes of multiple organisms allows us to study their genes as well as the evolution of those genes. While many eukaryotic genome assemblies already include computational gene predictions, these predictions can benefit from review and refinement through manual gene annotation. The Genomics Education Partnership (GEP; thegep.org) has developed an annotation protocol for protein-coding genes that enables undergraduate students and other researchers to create high-quality gene annotations that can be utilized in subsequent scientific investigations. For example, this protocol has been utilized by the GEP faculty to engage undergraduate students in the comparative annotation of genes involved in the insulin signaling pathway in 28 *Drosophila* species, using *D. melanogaster* as the informant genome. Students construct gene models using multiple lines of computational and experimental evidence including expression data (e.g., RNA-Seq), sequence similarity (e.g., BLAST, multiple sequence alignments), and computational gene predictions. For quality control, each gene is annotated by at least two students working independently, followed by reconciliation of the submitted gene models by a more experienced student. This article provides an overview of the annotation protocol and describes how discrepancies in student submitted gene models are resolved to produce a final, high-quality gene set suitable for subsequent analyses. This annotation protocol can be adapted to other scientific questions (e.g., expansion of the *Drosophila* Muller F element) and other species (e.g., parasitoid wasps) to provide additional opportunities for undergraduate students to participate in genomics research. These student annotation efforts can substantially improve the quality of gene annotations in publicly available genomic databases.

**Keywords**: comparative genomics, Course-based Undergraduate Research Experience, structural gene annotation

# Introduction

Genome annotation requires the assessment and integration of multiple lines of computational and experimental evidence. Several computational pipelines have been developed (e.g., BRAKER, MAKER) that can be used to construct an initial set of structural gene annotations for eukaryotic genomes (Holt and Yandell 2011; Hoff *et al.* 2019). Accuracy of the gene models produced by gene prediction algorithms depends on multiple biological factors (e.g., genome size, ploidy, repeat density, complexity of the transcriptome) and technical factors (e.g., quality of the genome assembly, evolutionary distance from the informant species, availability of transcriptome data) (Brůna *et al.* 2020). These factors can contribute to differences in the total number of gene predictions in closely related species (Figure 1).

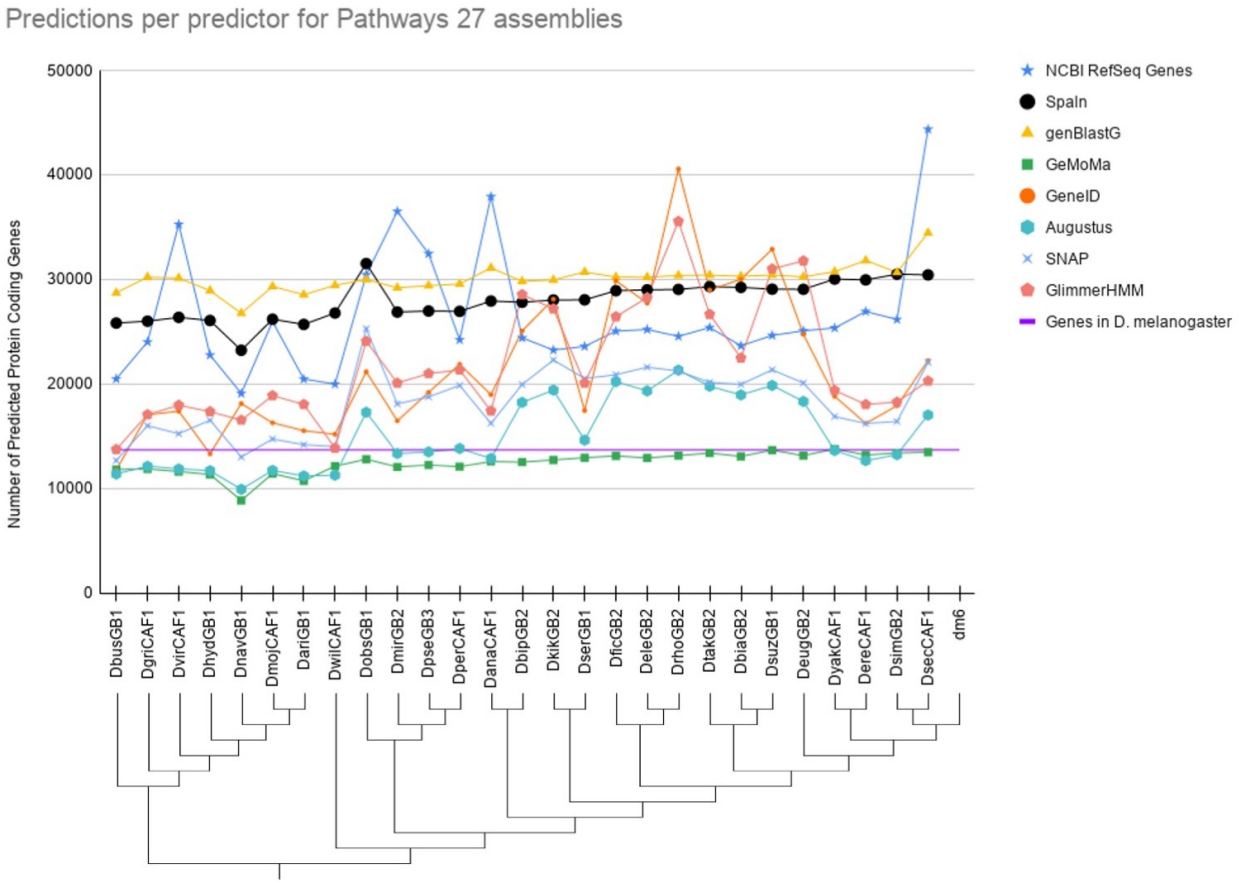


*Figure 1: Variable Number of Gene Predictions*
Variable number of gene predictions per gene predictor algorithm for each genome assembly for species in our clade (*Drosophila*). The number of genes predicted by different gene prediction algorithms differs depending on divergence from *D. melanogaster* for those predictors that use homology as their primary tool. The number of gene predictions varies drastically across algorithms and species. Some algorithms appear to consistently either underpredict or overpredict the number of genes relative to the well curated set of genes in *D. melanogaster* as maintained by FlyBase (purple line). The genome assemblies indicated in cladogram correspond to those mentioned in Table 1.

| Assembly | NCBI RefSeq Accession | Species Name |
|---|---|---|
| DbusGB1 | GCA_001277935.1 | *D. buskii* |
| DgriCAF1 | GCA_000005155.1 | *D. grimshawi* |
| DvirCAF1 | GCA_000005245.1 | *D. virilis* |
| DhydGB1 | GCA_002780465.1 | *D. hydei* |
| DnavGB1 | GCA_001654015.1 | *D. navajoa* |
| DmojCAF1 | GCA_000005175.1 | *D. mojavensis* |
| DariGB1 | GCA_001654025.1 | *D. arizonae* |
| DwilCAF1 | GCA_000005925.1 | *D. willistoni* |
| DobsGB1 | GCA_002217835.1 | *D. obscura* |
| DmirGB2 | GCA_000269505.2 | *D. miranda* |
| DpseGB3 | GCA_000001765.2 | *D. pseudoobscura* |
| DperCAF1 | GCA_000005195.1 | *D. persimilis* |
| DanaCAF1 | GCA_000005115.1 | *D. ananassae* |
| DbipGB2 | GCA_000236285.2 | *D. bipectinata* |
| DkikGB2 | GCA_000224215.2 | *D. kikkawai* |
| DserGB1 | GCA_002093755.1 | *D. serrata* |
| DficGB2 | GCA_000220665.2 | *D. ficusphila* |
| DeleGB2 | GCA_000224195.2 | *D. elegans* |
| DrhoGB2 | GCA_000236305.2 | *D. rhopaloa* |
| DtakGB2 | GCA_000224235.2 | *D. takahashii* |
| DbiaGB2 | GCA_000233415.2 | *D. biarmipes* |
| DsuzGB1 | GCA_000472105.1 | *D. suzukii* |
| DeugGB2 | GCA_000236325.2 | *D. eugracilis* |
| DyakCAF1 | GCA_000005975.1 | *D. yakuba* |
| DereCAF1 | GCA_000005135.1 | *D. erecta* |
| DsimGB2 | GCA_000754195.3 | *D. simulans* |
| DsecCAF1 | GCA_000005215.1 | *D. sechellia* |
| dm6 | GCA_000001215.4 | *D. melanogaster* |

*Table 1: Assemblies used for Insulin Signaling Pathway*
Assemblies used for the Insulin Signaling Pathway project along with their NCBI RefSeq Accession Number and species names

Evidence-based gene prediction algorithms (e.g., *Gnomon* (Souvorov *et al.* 2010)) use data, such as RNA-Seq derived from the same species, to predict genes within that assembly. The advent of

high-throughput RNA sequencing (RNA-Seq) data has led to substantial improvement in the quality of gene annotations (Hoff *et al.* 2016; Keilwagen *et al.* 2018), particularly for species that lack high-quality gene annotations from a closely related informant species. However, the efficacy of assembling transcripts from RNA-Seq data depends on the expression level of the transcript in the developmental stages and tissues sampled by RNA-Seq (Shao and Kingsford 2017). Furthermore, long-read RNA sequencing, which is needed to resolve mRNA isoforms, has several challenges pertaining to RNA extraction methods, bias toward short transcripts, sequencing throughput, and read accuracy (reviewed in (Byrne et al. 2019)). Past studies have also shown that transcriptome constructed from long-read RNA-Seq data (e.g., from Iso-Seq by Pacific Biosciences and Direct RNA sequencing by Oxford Nanopore) has high sensitivity but low precision (Kovaka *et al.* 2019).

Consequently, despite recent advances in gene prediction algorithms and the increasing availability of RNA-Seq data, gene predictions produced by computational algorithms can still benefit from manual review and refinement (Mudge and Harrow 2016; Tello-Ruiz *et al.* 2019). This article describes a protocol, developed by the Genomics Education Partnership (GEP; thegep.org), to engage undergraduate students in the comparative annotation of protein-coding genes involved in the insulin signaling pathway across 28 species of *Drosophila* (including *D. melanogaster* as the reference species).

As of December 2020, GEP students from 35 institutions have used this annotation protocol to construct 1,114 gene models across 27 *Drosophila* species. Despite differing instructional settings and instructors, this protocol ensures that GEP students use a uniform standard to construct gene models that are best supported by the available evidence. As an additional level of quality control, each gene is annotated by at least two students working independently, and the submitted gene models are then reconciled by an experienced student using the *Apollo* genome annotation editor (Dunn *et al.* 2019).

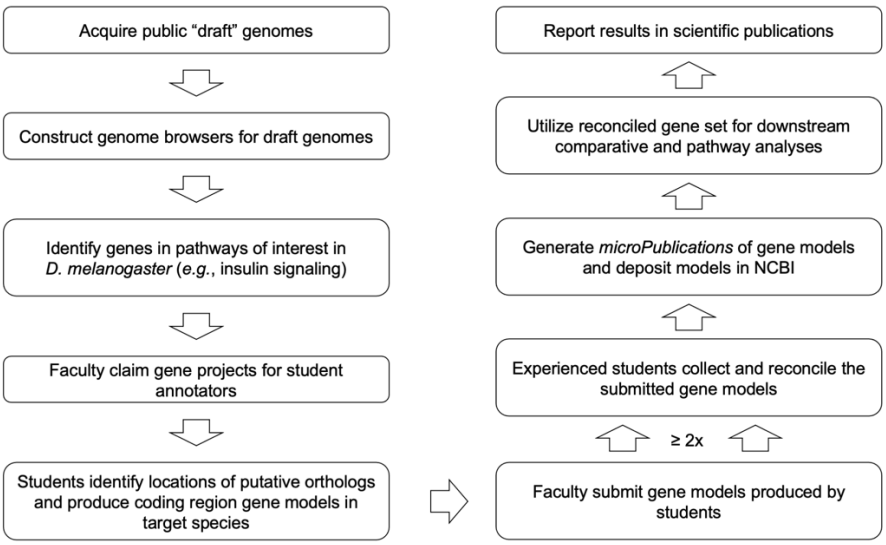


*Figure 2: Overall Model Creation Workflow*
Summarized workflow for annotation and reconciliation of a model to produce a finalized gene model

These reconciled gene models will be described in *microPublication* articles (Raciti *et al.* 2018) and submitted to the NCBI Third Party Annotation (TPA) database (Clark *et al.* 2016). Researchers can utilize the high-quality, manually curated gene models constructed by GEP students to investigate gene and genome evolution.

# Methods

## Overview of the Coding Region Annotation Protocol

*Drosophila* researchers and curators at FlyBase have produced high-quality gene annotations for *D. melanogaster* based on large amounts of genetic and sequencing data (Matthews *et al.* 2015). Our protocol utilizes the high-quality and comprehensive gene annotations from *D. melanogaster* (informant species) to facilitate the annotation of the coding regions of orthologous genes in other *Drosophila* species (target species). In the absence of compelling evidence (e.g., RNA-Seq data) indicating significant differences in the gene model, the proposed gene model in the target species minimizes the number of changes compared to the ortholog in the informant species (i.e., construct the most parsimonious gene model assuming evolutionary conservation).

In order to generate manual annotations for the protein-coding sequence (CDS) of a particular gene in the target species, we need to (1) identify the ortholog of that gene from our reference species in our target species using sequence similarity and synteny, (2) determine the structure and approximate coordinates of each isoform and its coding exons, and then (3) refine those coordinates for each isoform. The key analysis steps are summarized in the Annotation Workflow for the Pathways Project (Supplement 1). A walkthrough illustrating each step of the annotation protocol on an example gene is available on the GEP website (Supplement 2).

The annotation and reconciliation protocols described below utilize multiple bioinformatics tools. The "Data (and Software) Availability" section at the end of the manuscript provides a brief overview of each tool, including version information and references.

### CDS Annotation Procedure

#### *Project Claiming*

A GEP faculty member selects a *D. melanogaster* gene involved in the insulin signaling pathway (i.e., target gene in the informant genome) and one of the 27 *Drosophila* species (target species) for their students to annotate. A gene projects can be chosen by the faculty member to best suit their students and the pedagogical goals of the faculty member's course based on the estimated difficulty of the project, as well as any specific interests in the biological function of the gene.

#### *Identify the Ortholog*

The ortholog assignment in the target species is based on the analysis of protein sequence similarity and synteny (i.e., the relative gene order and orientation) (Jun *et al.* 2009; Jahangiri-Tazehkand *et al.* 2017). The key analysis steps are summarized in Supplement 3.

To identify the ortholog of our target gene, we examine the genomic neighborhood surrounding the gene in the GEP UCSC Genome Browser (further described in the Software section below;

https://gander.wustl.edu) in both the informant species and the target species. This includes identifying the orientation and nearest two upstream and downstream genes relative to our target gene. We also compare any nested genes to the locus surrounding our putative ortholog in the target species (Figure 3).

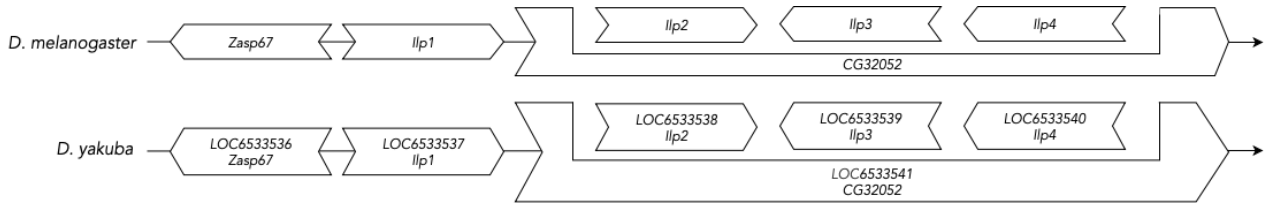


*Figure 3: Synteny of Ilp2 in D. yakuba*
Schematic of genomic neighborhood around the *Ilp2* gene in *D. melanogaster* and *D. yakuba*, showing that *Ilp2*, *Ilp3*, and *Ilp4* are nested within *CG32052.*

Locating the putative ortholog requires us to obtain the protein sequence for the target gene in our informant species from the *Gene Record Finder* and use it as the query to perform a *tblastn* search against the target species' genome assembly. We perform *tblastn* against a specific assembly of interest on NCBI (https://www.ncbi.nlm.nih.gov/). We use *tblastn* since we are comparing our protein sequence query against a nucleotide database. Amino acid sequences are more conserved than nucleotide sequences across evolutionary time, due the degeneracy of the genetic code sequence similarity searches at the amino acid level are more sensitive than searches at the nucleotide level (States *et al.* 1991; Gonzalez *et al.* 2019). There are three possible outcomes of our *tblastn* search: (1) getting zero matches, (2) getting one match, or (3) getting more than one match.

A lack of *tblastn* matches could be caused by gaps or misassemblies in the genome assembly of the target species, or due to lack of sequence conservation between the orthologs. If this is the case, we can use the ortholog assignments for the flanking genes in the informant species and synteny (orthology of neighboring genes) to help establish the location of the ortholog for the target gene. If the location and candidate for the ortholog still cannot be established, the ortholog may not be present in the target species. The primary usage of our *tblastn* search is comparing the metrics of the best and second-best matches to confidently assign orthology. The presence of a single or multiple matches implies potential orthology that is then confirmed by comparing the local synteny of our target gene between the target and informant species. If there is no match for neighboring genes in an evolutionarily diverged species but a singular match exists for the target gene, there still may be an orthologous relationship between the reference and target sequences if they are reciprocal best hits in a BLAST search.

Figure 3 shows an example of how we can use synteny to establish the ortholog of the target gene. *Ilp2* in *D. melanogaster* and *D. yakuba* have the same genes surrounding them. A few of these genes are also nested within another gene, which further supports that the genomic neighborhood is conserved, and thus, the ortholog has been identified.

## *Identify the Approximate Coordinates of Each Coding Exon*

Once the putative ortholog is identified, we begin constructing our gene model by separately mapping each coding exon to determine their approximate locations using the "Align two or more sequences" (*bl2seq*) feature provided by the NCBI BLAST web service.

For each isoform of the target gene in *D. melanogaster*, we performed *tblastn* searches of each coding exon of the isoform (query) in *D. melanogaster* against the scaffold which contains the putative ortholog of the target gene in the target species (subject). The amino acid sequence for each coding exon in the *D. melanogaster* isoform is obtained from the Gene Record Finder (further described in the Software section below). The scaffold in the target species which contains the putative ortholog of the target gene was determined from the "Identify the Ortholog" step above. For scaffolds larger than 10 Mbp, the "From" and "To" fields under "Subject subrange" in the NCBI *tblastn* search interface are used to limit the size of the search region to the approximate location of the target gene (inferred from results of the "Identify the Ortholog" step).

The default *tblastn* search parameters for NCBI Web BLAST are used in this search except for the following parameters:

1. Select the "Align two or more sequences" checkbox
2. Specify a subject subrange that corresponds to the coding span estimated at the "Identify the Ortholog" step
3. Select "No adjustment" under "Compositional adjustments"
4. Uncheck the "Low complexity regions" filter (under "Filters and Masking")

At the end of this process, we have identified a collinear set of coordinates for most of the coding exons of the isoform. The results of these *tblastn* searches provide further supporting evidence for the ortholog assignment and provide anchors from which to narrow the size of the search region for identifying small or weakly conserved coding exons.

Since *tblastn* only aligns to complete codons, the BLAST alignments will not include partial codons adjacent to the splice junctions. In addition, *tblastn* does not take the locations of potential splice sites into account when it generates the BLAST alignment. Consequently, other lines of evidence (e.g., computational gene predictions, RNA-Seq data) must be used to refine the start and end coordinates of each coding exon.

## *Refine Coordinates*

The coding exon coordinates determined by the *tblastn* searches are approximate as they only detect regions of similarity with complete codons and are only estimates of the portion of the coding exon which shows significant similarity to the coding exon in the informant genome. These coordinates do not account for intron splicing rules. We refine the collinear coordinates identified above utilizing the RNA-Seq track, other homology-based alignment algorithms, and gene predictions, as well as comparing the sequence to other species within the phylogeny (more details about the tracks used can be found in Supplement 4). If available, the RNA-Seq data can provide

empirical support for the proposed gene model in the target species. Canonical splice sites (GT donor and AG acceptor) are adhered to unless there is good evidence to the contrary, such as spliced RNA-Seq reads from the target species and conservation of non-canonical splice sites across the phylogeny. We also need to ensure the donor and acceptor have compatible phases. A more detailed version of the workflow for coordinate refinement can be found within Supplement 5. Figure 4 gives an example coordinate refinement to account for the complete exon boundaries considering RNA-Seq data and splice site compatibility, and Table 2 shows the coordinates for the entire example gene along with their refined counterparts.

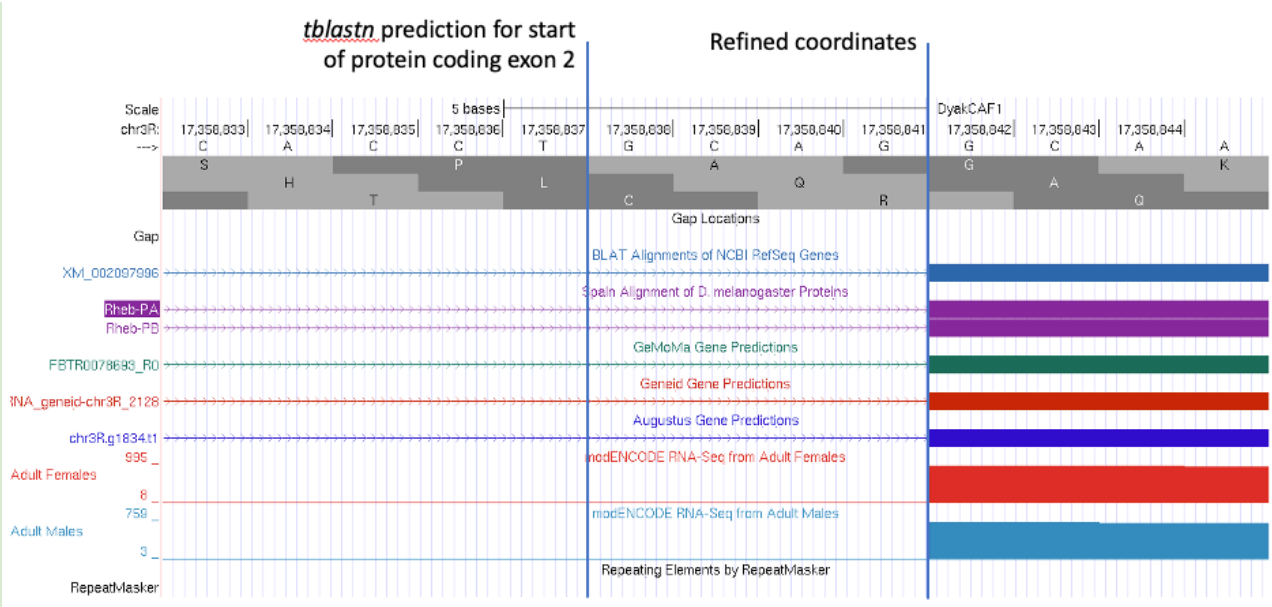


*Figure 4: Refined Coordinates*
Coordinate refinement for the second CDS of *Rheb* in *D. yakuba* since tblastn predicted the exon to be larger than it actually is based on other lines of evidence.

| | *tblastn* | | Refined | | Δ | |
|---|---|---|---|---|---|---|
| Exon Number | Start | Stop | Start | Stop | Start | Stop |
| 1 | 17,358,666 | 17,358,725 | 17,358,666 | 17,358,666 | 0 | -11 |
| 2 | 17,358,838 | 17,358,927 | 17,358,842 | 17,358,913 | 4 | -14 |
| 3 | 17,359,007 | 17,359,216 | 17,359,011 | 17,359,218 | 4 | 2 |
| 4 | 17,359,279 | 17,359,407 | 17,359,407 | 17,359,407 | -1 | 0 |
| 5 | 17,359,467 | 17,359,556 | 17,359,470 | 17,359,556 | 3 | 0 |

*Table 2:Refining rough tblastn coordinates*
Comparison between the *tblastn* and refined coordinates for the model of *Rheb* in *D. yakuba*

After refining the coding exon coordinates, we use the *Gene Model Checker* to verify the coordinates for the proposed gene model satisfies the biological constraints for protein-coding genes in most eukaryotes, and reflects the structure of the *D. melanogaster* ortholog. The dot plot and protein alignment between the model and the *D. melanogaster* ortholog help us verify that our proposed gene model is the most parsimonious compared to the *D. melanogaster* ortholog. For more information on the *Gene Model Checker*, refer to the "Data (and Software) Availability" section.

We repeat the "Identify the Approximate Coordinates of Each Coding Exon" and the "Refine Coordinates" steps to construct gene models for each protein coding isoform of our target gene.

### *Final Submission*

We then submit a GFF and FASTA (FAA and FNA) of each of the isoforms (see *Annotation Files Merger* tool below) as well as a complete Annotation Report (Supplement 6) to document the evidence supporting the proposed gene model.

## Exceptions to the Standard Annotation Workflow

There are many challenges that can require the use of different annotation strategies than the basic annotation protocol outlined above. Gene models with these exceptions require support from multiple lines of evidence (e.g., RNA-Seq data, congruence with multiple gene predictors, conservation with *D. melanogaster* and other species in the clade). Below we describe the process for two such more common challenges:

### *Non-canonical Splice Site*

GT and AG are the most commonly used splice donor and acceptor sites, respectively, and variant splice sites are termed non-canonical. For example, the GC splice donor site appears in ~1% (604 / 71,909) of the unique introns in *D. melanogaster* (FlyBase release 6.34). The use of a non-canonical splice site in a gene model will typically be supported by splice junction predictions derived from spliced RNA-Seq reads. The presence of a non-canonical splice site in the orthologous intron in the informant species (*D. melanogaster*) or in multiple *Drosophila* species more closely-related to the target species could also be used to support the annotation of a non-canonical splice site.

### *Assembly Errors*

Each sequencing platform (e.g., Sanger, Illumina, PacBio, Nanopore) has a distinct error profile (Huang *et al.* 2012; Yang *et al.* 2017) that could introduce errors (e.g., base substitutions, insertions, and deletions) into the consensus sequence of an assembly. Transposons and other

repetitive sequences (e.g., tandem repeats) in eukaryotic genomes can also lead to gaps and misassemblies (Mikheenko *et al.* 2018). These assembly issues can have substantial impacts on gene annotations — leading to frame shifts within coding exons, partial genes, missing genes or isoforms, and errors in ortholog/paralog assignments (e.g., due to "duplications" caused by misassemblies).

The publicly available genome assemblies our projects utilize are constructed using different sequencing technologies and assembly protocols. Since these genome assemblies have not been manually improved, they might contain assembly errors that could interfere with coding region annotations. Consequently, in cases where the proposed gene models for the target species include changes in gene structure compared to the *D. melanogaster* ortholog (e.g., novel/missing isoforms, exons, and/or introns), further investigations are needed to ascertain if the difference is caused by an assembly error, or reflects true divergence. As part of this assessment, we evaluate multiple lines of evidence such as: (1) conservation across other species in the *Drosophila* phylogeny, (2) consistency with genomic reads and RNA-Seq reads in the NCBI Sequence Read Archive (SRA) database, or (3) comparison with other genome assemblies for the same species (Miller *et al.* 2018).

### *Other Exceptions*

Other potential challenges include whether the orthology of the gene in the target species has been correctly assigned in instances such as changes in gene structure relative to the reference species, absence of an isoform in the target species, or presence of a novel paralog are observed. We might also have cases where we have a non-canonical start codon (e.g., *Akt1*), stop codon readthrough (e.g., *jim*), or trans-splicing (e.g., *mod(mdg4)*), none of which conform to a standardized genetic ruleset. Challenges caused by assembly errors could be partial genes due to either gaps in the assembly, or the gene being near the end of a scaffold, as well as potential frameshifts within CDSs caused by errors in the consensus sequence.

## Common Issues in Gene Models Prior to Submission

### *Multiple Fails in the Gene Model Checker Checklist*

Gene models sometimes produce multiple "fails" for specific biological rules that were violated in the *Gene Model Checker*. Typically, fails are attributed to errors in upstream regions of the gene model that propagate downstream.

One common cause of multiple failures is frameshifts caused by selecting incompatible donor and acceptor splice sites. When a coding exon ends in an incomplete codon, the 3' end of that coding exon and the 5' beginning of the nearest downstream coding exon must include nucleotides that form a complete codon once the intron has been spliced out. The number of nucleotides between the end of the last complete codon and the splice donor site is defined as the phase of the splice donor site. Similarly, the number of nucleotides between the splice acceptor site and the start of the

first complete codon is defined as the phase of the splice acceptor site. In order to maintain the open reading frame after the intron has been spliced out, the sum of the donor and acceptor phases for adjacent coding exons must either be zero (no extra codon) or three (one extra codon). Selecting incompatible donor and acceptor splice sites causes a frameshift that changes the reading frame of any coding exons downstream of that splice junction, thereby triggering multiple fails in the *Gene Model Checker*.

To resolve multiple fails, start troubleshooting at the beginning of the gene. In many cases, correcting errors in the upstream portion of the gene model resolves the fails reported downstream.

# Reconciliation

While most gene models produced by students using the annotation protocol described above are congruent with each other, incongruent models require further examination by a student reconciler. Reconciliation is carried out by experienced students who have received additional training on reconciliation under the guidance of a faculty and/or senior-scientist mentor.

The strategy of collecting at least two student models for each gene of interest in a target species and then comparing them with each other to reconcile a final model is essential for quality control. Student reconcilers look for differences in the submitted gene models, paying special attention to the three most common errors that might invalidate a model (described below), and investigate any large-scale anomalies (e.g., proposed novel isoform, missing specific exons or isoforms). It is relatively common for one student to make a single error but less common for another student, working independently of the first, to make the same error.

## Reconciliation Process

Reconciliation is performed using Apollo (Dunn et al. 2019), a web-based collaborative gene annotation editor that allows reconcilers to view student-generated models alongside the evidence tracks (Figure 5) used for annotating those models.

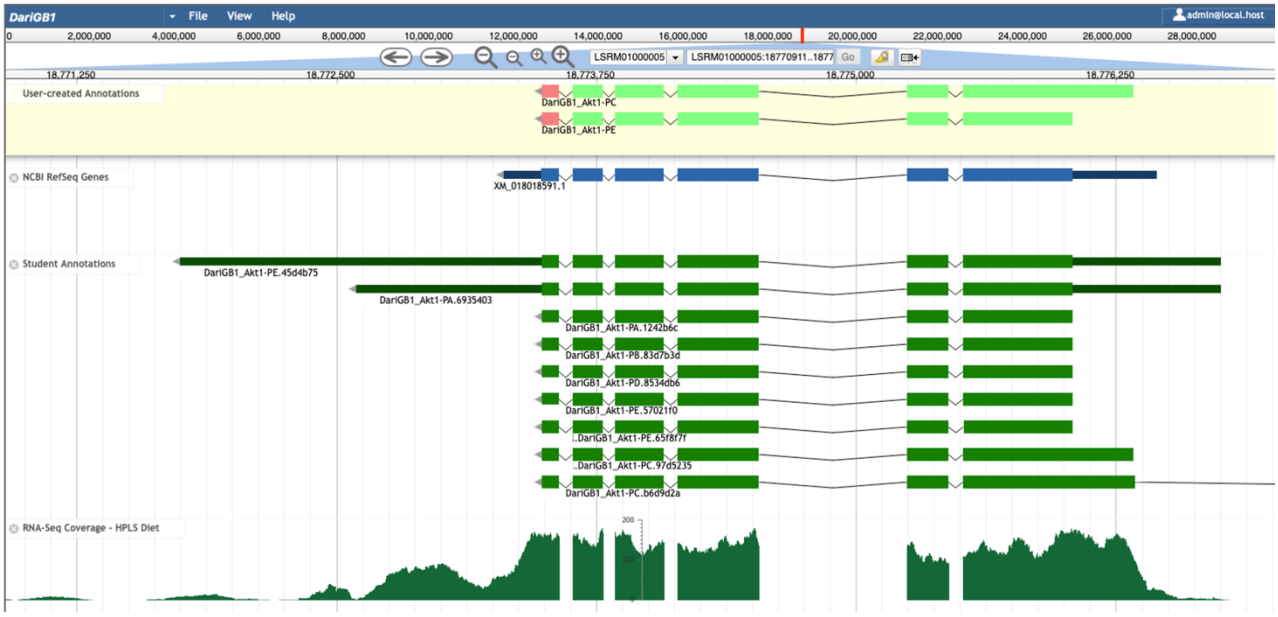


*Figure 5: Apollo Screenshot for the Akt1 gene model in D. arizonae*
Final gene model for *Akt1* in *D. arizonae*, along with the submitted student models, and RNA-Seq data aligning to the region. The final model shows that despite there being only a single isoform prediction for a protein coding gene by RefSeq, there are likely two protein coding isoforms for this gene, which were annotated using multiple lines of evidence. The second isoform has a larger coding region in the reconciled gene model that is missed by the RefSeq genome predictor.

Reconcilers evaluate the available student annotations for each isoform in conjunction with the other evidence tracks (e.g., sequence similarity, RNA-Seq data, gene predictions) to construct the final gene model that is best supported by the available evidence. After creating a reconciled model, reconcilers draft a *microPublication* describing the supporting evidence for the final gene model (Raciti *et al.* 2018). The student annotators and their faculty mentors then review and approve the article draft. Reconciled models are also used in downstream meta-analysis and deposited into the NCBI Third Party Annotation (TPA) database.

## Most Common Annotation Errors

Reconciliation consists of reviewing two or more models created by student annotators. The main advantage of manually curated gene models relative to computational predictions is the ability for the curator (in our case, the student annotator) to evaluate and integrate across non-conforming pieces of evidence. The reconciler also pays close attention to each of the putative idiosyncrasies in the proposed model. Some of the most common idiosyncrasies/errors are listed below.

### *Selection of Incorrect Splice Sites*

The use of a non-canonical splice site in the orthologous intron in other *Drosophila* species provides supporting evidence for the use of the non-canonical splice site in the target species (Sheth *et al.* 2006; Parada *et al.* 2014; Sibley *et al.* 2016). There are cases where annotators place too high of a priority on the use of canonical GT and AG donor and acceptor splice sites, when

there is strong evidence from spliced RNA-Seq reads and conservation in other *Drosophila* species that support the use of non-canonical splice sites.

### *Missing/Extra Exons*

There may be cases where the student model is missing exons or has predicted an additional exon relative to the *D. melanogaster* ortholog. Sometimes these predictions are well supported by the data, but often the mismatch in exon number is due to the student annotator failing to fully account for all lines of evidence.

### *Incorrect Ortholog Assignment*

Correctly identifying an ortholog depends on an appropriate BLAST search, which provides matches to the gene of interest, and properly evaluating the conservation of synteny. If a student annotates a gene that is not the ortholog, a reconciler must evaluate based on BLAST and synteny what the correct position of the ortholog is in the target species, and if the particular model does not have two congruent annotations, it must be returned to the claim list for annotation by other students.

# Discussion

This annotation protocol has been used by GEP faculty to engage undergraduate students in the comparative annotation of insulin signaling pathway genes in 27 *Drosophila* species. Similar protocols have been used by the GEP on other scientific projects in the past (Slawson *et al.* 2006; Leung *et al.* 2010, 2015, 2017). Collectively, in the last two years, GEP faculty and students across the country have helped us generate 1114 gene models (783 unique models) for the Insulin Pathways project as of December 2020.

To ensure that the gene models produced by GEP students are of high quality, each gene is annotated by at least two students working independently and then reconciled by a more experienced student. This manual curation process ensures the availability of an accurate set of gene models from multiple *Drosophila* species for the comparative study of network architecture and the evolution of genes involved in signaling pathways (e.g., the insulin signaling pathway).

For example, GEP students were able to annotate a new isoform for *Akt1* in *D. arizonae*. This isoform was missed by the NCBI RefSeq gene predictor due to a non-canonical start codon that is conserved in *Akt1* in *D. melanogaster*.

The annotation and reconciliation protocols described in this article have been adapted to two additional scientific projects within the GEP: investigation of the expansion of the Muller F element in four *Drosophila* species and investigation of the evolution of venom proteins in four parasitoid wasp species. The annotation protocol described here is focused only on the annotation coding regions. Once additional experimental data (e.g., CAGE, RAMPAGE, long-read RNA-Seq) becomes available in more species, future analyses will focus on the annotation of untranslated regions and transcription start sites. Faculty can also use this protocol to create new Course-based Undergraduate Research Experiences (CUREs) that engage students in the comparative analysis of genes involved in other metabolic and signaling pathways.

# Data (and Software) Availability

## Software

All custom software and tools generated by the GEP can be accessed from the GEP website (thegep.org). The tools are all web-based applications, thus requiring no special installation on the part of the user.

### GEP UCSC Genome Browser

Since GEP materials are catered towards undergraduates that may have limited or no experience with programming and command-line interfaces, we endeavor to make all data easily accessible, thereby reducing the barrier to engagement in genomics research. The GEP maintains a mirror of the UCSC Genome Browser with *Drosophila* genomes (https://gander.wustl.edu). We created a these genome browsers (Shaffer *et al.* 2014) with multiple evidence tracks using data generated by various algorithms (Supplement 2).

### Gene Record Finder

The *Gene Record Finder* summarizes the FlyBase annotations for protein-coding genes in *D. melanogaster*. It provides information about the structure of each *D. melanogaster* gene, such as the number of isoforms (and number of isoforms with unique coding regions), as well as the amino acid sequence for each coding exon and the nucleotide sequence for each transcribed exon. The *Gene Record Finder* also provides exon usage maps that demarcate the exons used by each isoform. This information is used in step two of the annotation protocol ("Identify the Approximate Coordinates of Each Coding Exon"). While this information can be obtained from FlyBase directly, it is easier for annotators to retrieve the gene structure information from a single page instead of through multiple FlyBase Transcript and Polypeptide Reports.

### Gene Model Checker

The *Gene Model Checker* provides a way for annotators to check their own work when constructing gene models. It verifies that the proposed gene model satisfies basic biological constraints (e.g., maintains an open reading frame, uses canonical start and stop codons, and canonical splice donor and acceptor sites). It also indicates whether the predicted exon number matches that in the orthologous isoform in *D. melanogaster.* It does not inform us whether our proposed model is otherwise accurate. If there is empirical evidence indicating the gene has unusual characteristics (e.g., the use of a non-canonical start codon, stop codon read-through, polycistronic transcripts (Crosby *et al.* 2015)), we provide the supporting evidence for our claims in the annotation report form. The "Dot Plot" section of the *Gene Model Checker* output compares the proposed gene model against the putative ortholog in *D. melanogaster* using protein alignment algorithms. Large gaps in the dot plot or protein alignment might indicate the selection of an

incorrect splice site, missing exons, or extra exons in the proposed gene model. Note that this tool does not compare the proposed gene model against other lines of evidence, such as RNA-Seq data or computational gene predictions. The *Gene Model Checker* also produces the three annotation files for the proposed gene model that are required for project submission (i.e., the GFF and FASTA files for the transcript and protein sequences (FAA and FNA)).

The *Gene Model Checker* requires input of the species, assembly, and the scaffold where the putative ortholog is located. The gene model specific information requested by the *Gene Model Checker* includes the name of the ortholog in *D. melanogaster,* the set of coding exon coordinates for the gene in the target species, and the orientation of the gene. Further model specific information requested includes noting whether or not the untranslated regions (UTRs) are included, the gene model is complete (i.e., encompasses all CDSs/UTRs), and whether the genomic region containing the gene in the target species has consensus errors (Figure 6).

Please note that, at present, this tool only supports the 28 *Drosophila* species that are currently in the GEP UCSC Genome Browser.

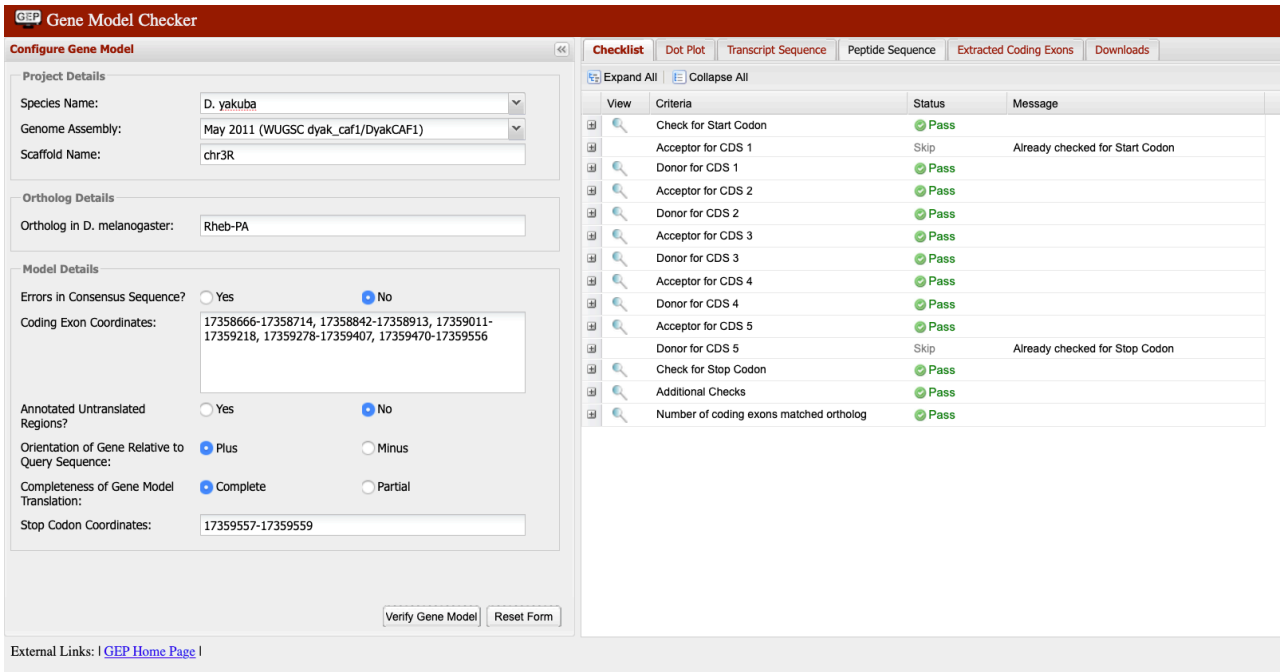


*Figure 6: Gene Model Checker*
Gene Model Checker with information for *Rheb*-PA in *D. yakuba*: visualizing how we establish biological sanity for one of the two proposed isoform models.

## Small Exons Finder

The typical use case for the *Small Exons Finder* tool is for identifying CDSs that are too small or too weakly conserved to be detected by BLAST. The tool is designed to look for open reading frames (ORFs) that satisfy a set of biological constraints. These constraints include the type of CDS (i.e., initial, internal, or terminal CDS), the phase of the donor or acceptor site, and the expected CDS size according to the *D. melanogaster* model. The *Small Exons Finder* then looks

for ORFs in the provided sequence that conform to the aforementioned constraints. Compared to the *ORFfinder* tool developed by NCBI (Rombel *et al.* 2002), the *Small Exons Finder* allows users to search for open reading frames that are less than 30bp in size, and it can search for initial, internal, and terminal coding exons with constraints on the phases of the donor and acceptor sites.

## Sequence Updater

The *Sequence Updater* tool is primarily used to create a VCF (Danecek *et al.* 2011) file to correct assembly errors in the consensus sequence. It provides an updated assembly sequence with the base(s) hypothesized to be more accurate within a specific region in the assembly. This tool is used to correct errors (i.e., base substitutions, insertions, deletions) in the genomic sequence of the project region. The VCF file produced by the *Sequence Updater* can be used with the *Gene Model Checker* to validate a gene model with consensus errors.

## Annotation Files Merger

The *Gene Model Checker* produces GFF, FNA, and FAA files for each isoform. The submission pipeline requires a single GFF, FNA, and FAA file that includes all of the isoforms in a project. For each type of annotation file, the *Annotation Files Merger* is used to combine the annotations for all of the isoforms in a project into a single file. The *Annotation Files Merger* also enables the user to view the combined GFF file as a custom track on the GEP UCSC Genome Browser.

# Grant Information

GEP is supported by NSF grant #1915544, NIH grant #R25GM130517, and hosted by The University of Alabama and Washington University in St. Louis.

# Acknowledgements

The work done by GEP Faculty and their students has generated student models that have benefited us in refining our protocol. We would also like to thank the reconcilers who have generated models that can be used for downstream analysis.

# References


Brůna T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2020 BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. bioRxiv 2020.08.10.245134. https://doi.org/10.1101/2020.08.10.245134

Byrne A., C. Cole, R. Volden, and C. Vollmers, 2019 Realizing the potential of full-length transcriptome sequencing. Philos. Trans. R. Soc. London. Ser. B, Biol. Sci. 374: 20190097. https://doi.org/10.1098/rstb.2019.0097

Clark K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, 2016 GenBank. Nucleic Acids Res. 44: D67-72. https://doi.org/10.1093/nar/gkv1276

Crosby M. A., L. S. Gramates, G. Dos Santos, B. B. Matthews, S. E. St Pierre, *et al.*, 2015 Gene Model Annotations for Drosophila melanogaster: The Rule-Benders. G3 (Bethesda). 5: 1737–1749. https://doi.org/10.1534/g3.115.018937

Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dunn N. A., D. R. Unni, C. Diesh, M. Munoz-Torres, N. L. Harris, *et al.*, 2019 Apollo: Democratizing genome annotation. PLoS Comput. Biol. 15: e1006790. https://doi.org/10.1371/journal.pcbi.1006790

Gonzalez D. L., S. Giannerini, and R. Rosa, 2019 On the origin of degeneracy in the genetic code. Interface Focus 9: 20190038. https://doi.org/10.1098/rsfs.2019.0038

Hoff K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2016 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff K. J., A. Lomsadze, M. Borodovsky, and M. Stanke, 2019 Whole-Genome Annotation with BRAKER. Methods Mol. Biol. 1962: 65–95. https://doi.org/10.1007/978-1-4939-9173-0_5

Holt C., and M. Yandell, 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491. https://doi.org/10.1186/1471-2105-12-491

Huang W., L. Li, J. R. Myers, and G. T. Marth, 2012 ART: A next-generation sequencing read simulator. Bioinformatics 28: 593–594. https://doi.org/10.1093/bioinformatics/btr708

Jahangiri-Tazehkand S., L. Wong, and C. Eslahchi, 2017 OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation. Genomics. Proteomics Bioinformatics 15: 361–370. https://doi.org/10.1016/j.gpb.2017.07.002

Jun J., I. I. Mandoiu, and C. E. Nelson, 2009 Identification of mammalian orthologs using local synteny. BMC Genomics 10: 630. https://doi.org/10.1186/1471-2164-10-630

Keilwagen J., F. Hartung, M. Paulini, S. O. Twardziok, and J. Grau, 2018 Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics 19: 189. https://doi.org/10.1186/s12859-018-2203-5

Kovaka S., A. V Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, *et al.*, 2019 Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20: 278. https://doi.org/10.1186/s13059-019-1910-1

Leung W., C. D. Shaffer, T. Cordonnier, J. Wong, M. S. Itano, *et al.*, 2010 Evolution of a distinct genomic domain in Drosophila: comparative analysis of the dot chromosome in Drosophila melanogaster and Drosophila virilis. Genetics 185: 1519–1534. https://doi.org/10.1534/genetics.110.116129

Leung W., C. D. Shaffer, L. K. Reed, S. T. Smith, W. Barshop, *et al.*, 2015 Drosophila muller f elements maintain a distinct set of genomic properties over 40 million years of evolution. G3 (Bethesda). 5: 719–740. https://doi.org/10.1534/g3.114.015966

Leung W., C. D. Shaffer, E. J. Chen, T. J. Quisenberry, K. Ko, *et al.*, 2017 Retrotransposons are the major contributors to the expansion of the Drosophila ananassae muller F element. G3 Genes, Genomes, Genet. https://doi.org/10.1534/g3.117.040907

Matthews B. B., G. Dos Santos, M. A. Crosby, D. B. Emmert, S. E. St Pierre, *et al.*, 2015 Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data. G3 (Bethesda). 5: 1721–1736. https://doi.org/10.1534/g3.115.018929

Mikheenko A., A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich, 2018 Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150. https://doi.org/10.1093/bioinformatics/bty266

Miller D. E., C. Staber, J. Zeitlinger, and R. S. Hawley, 2018 Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 (Bethesda). 8: 3131–3141. https://doi.org/10.1534/g3.118.200160

Mudge J. M., and J. Harrow, 2016 The state of play in higher eukaryote gene annotation. Nat. Rev. Genet. 17: 758–772. https://doi.org/10.1038/nrg.2016.119

Parada G. E., R. Munita, C. A. Cerda, and K. Gysling, 2014 A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Res. 42: 10564–10578. https://doi.org/10.1093/nar/gku744

Raciti D., K. Yook, T. W. Harris, T. Schedl, and P. W. Sternberg, 2018 Micropublication: incentivizing community curation and placing unpublished data into the public domain. Database (Oxford). 2018. https://doi.org/10.1093/database/bay013

Rombel I. T., K. F. Sykes, S. Rayner, and S. A. Johnston, 2002 ORF-FINDER: a vector for high-throughput gene identification. Gene 282: 33–41. https://doi.org/10.1016/s0378-1119(01)00819-8

Shaffer C. D., C. J. Alvarez, A. E. Bednarski, D. Dunbar, A. L. Goodman, *et al.*, 2014 A course-based research experience: how benefits change with increased investment in instructional time. CBE Life Sci. Educ. 13: 111–130. https://doi.org/10.1187/cbe-13-08-0152

Shao M., and C. Kingsford, 2017 Accurate assembly of transcripts through phase-preserving graph decomposition. Nat. Biotechnol. 35: 1167–1169. https://doi.org/10.1038/nbt.4020

Sheth N., X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer, *et al.*, 2006 Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. 34: 3955–3967. https://doi.org/10.1093/nar/gkl556

Sibley C. R., L. Blazquez, and J. Ule, 2016 Lessons from non-canonical splicing. Nat. Rev. Genet. 17: 407–421. https://doi.org/10.1038/nrg.2016.46

Slawson E. E., C. D. Shaffer, C. D. Malone, W. Leung, E. Kellmann, *et al.*, 2006 Comparison of

dot chromosome sequences from D. melanogaster and D. virilis reveals an enrichment of DNA transposon sequences in heterochromatic domains. Genome Biol. 7: R15. https://doi.org/10.1186/gb-2006-7-2-r15

Souvorov a, Y. Kapustin, B. Kiryutin, V. Chetvernin, T. Tatusova, *et al.*, 2010 Gnomon–NCBI eukaryotic gene prediction tool. Natl. Cent. Biotechnol. Inf. 1–24.

States D. J., W. Gish, and S. F. Altschul, 1991 Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods 3: 66–70. https://doi.org/https://doi.org/10.1016/S1046-2023(05)80165-3

Tello-Ruiz M. K., C. F. Marco, F. M. Hsu, R. S. Khangura, P. Qiao, *et al.*, 2019 Double triage to identify poorly annotated genes in maize: The missing link in community curation. PLoS One 14: 1–13. https://doi.org/10.1371/journal.pone.0224086

Yang C., J. Chu, R. L. Warren, and I. Birol, 2017 NanoSim: Nanopore sequence read simulator based on statistical characterization. Gigascience 6: 1–6. https://doi.org/10.1093/gigascience/gix010