Check for updates

METHOD ARTICLE

REVISED Manual annotation of *Drosophila* genes: a Genomics Education Partnership protocol [version 2; peer review: 1 approved, 1 approved with reservations]

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, Alabama, 35487, USA
[2]Department of Biology, Washington University in St. Louis, St. Louis, Missouri, 63130, USA

v2 **First published:** 23 Dec 2022, **11**:1579
https://doi.org/10.12688/f1000research.126839.1
Latest published: 31 Jul 2023, **11**:1579
https://doi.org/10.12688/f1000research.126839.2

Abstract

Annotating the genomes of multiple species allows us to analyze the evolution of their 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; https://thegep.org/) developed a structural annotation protocol for protein-coding genes that enables undergraduate student and faculty 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 27 *Drosophila* species, using *D. melanogaster* as the reference genome. Students construct gene models using multiple lines of computational and empirical evidence including expression data (e.g., RNA-Seq), sequence similarity (e.g., BLAST and multiple sequence alignment), and computational gene predictions. Quality control measures require each gene be 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. The protocol can be adapted to other scientific questions (e.g., expansion of the *Drosophila* Muller F element) and 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, CURE, structural gene annotation

Open Peer Review

Approval Status ? ✓

	1	2
version 2 (revision) 31 Jul 2023	? view	✓ view
version 1 23 Dec 2022	? view	? view

1. **Terence D. Murphy**, National Institutes of Health, Bethesda, USA

2. **Madeline Crosby**, Harvard University, Cambridge, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Bioinformatics gateway.

This article is included in the Genomics and Genetics gateway.

Corresponding author: Chinmay P. Rele (cprele@ua.edu)

Author roles: Rele CP: Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sandlin KM**: Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Leung W**: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Reed LK**: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The Genomics Education Partnership is funded by the National Science Foundation (#1915544) and National Institute of General Medical Sciences of the National Institutes of Health (#R25GM130517) and is hosted by the Department of Biological Sciences at The University of Alabama with continuing support from the Department of Biology at Washington University in St. Louis. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation nor the National Institutes of Health.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Rele CP *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Rele CP, Sandlin KM, Leung W and Reed LK. **Manual annotation of *Drosophila* genes: a Genomics Education Partnership protocol [version 2; peer review: 1 approved, 1 approved with reservations]** F1000Research 2023, **11**:1579 https://doi.org/10.12688/f1000research.126839.2

First published: 23 Dec 2022, **11**:1579 https://doi.org/10.12688/f1000research.126839.1

REVISED **Amendments from Version 1**

A more detailed description of the changes that were made between version one and version two of the article are provided in the detailed response to reviewers. We appreciate the constructive feedback from the reviewers. In summary, the changes that were incorporated are:

- A description of the goals and the mission statement for the GEP.
- Changing the Figure 1 caption to make clear that some prediction algorithms identify unique genes, while others identify unique transcripts.
- Adding the canonical names for the assemblies used by the pathways project as a new column in Table 1, and also making clear our plan to use newer/higher quality assemblies when they become available.
- An update to the data citing the number of models reconciled.
- Adding a new figure explaining the use of synteny to define orthology.
- Updating an incorrect isoform name in the caption of Figure 6.
- Making clear that manually-curated models tend to be of higher quality at the level of the individual gene. In the previous version, we did not mention the scale at which manually-curated models have a higher quality.
- An update to a reference that was previously citing the redacted version of an article.
- We also thought adding captions to Supplements 7 and 8 would be beneficial.

We did not add a more detailed analysis of the types of errors student annotators make. We omit this recommended change because the goal of this article is to provide the protocol the students use to generate the gene models. Also, we do not have sufficient sample size to make meaningful conclusions about the relationship between the gene's properties and student error profiles. We do have plans to study the error profiles within student models once we have enough data to make our conclusions statistically robust.

Any further responses from the reviewers can be found at the end of the article

Introduction

Genome annotation requires assessing and integrating multiple lines of computational and empirical evidence. Several computational pipelines have been developed (e.g., BRAKER and MAKER) for constructing an initial set of structural gene annotations for eukaryotic genomes.[1,2] Accuracy of the gene models produced by gene prediction algorithms depends on multiple biological (e.g., genome size, ploidy, repeat density, and complexity of the transcriptome) and technical factors (e.g., quality of the genome assembly, evolutionary distance from the reference species, and availability of transcriptome data).[3] These factors can contribute to differing numbers of gene predictions in closely related species (Figure 1).

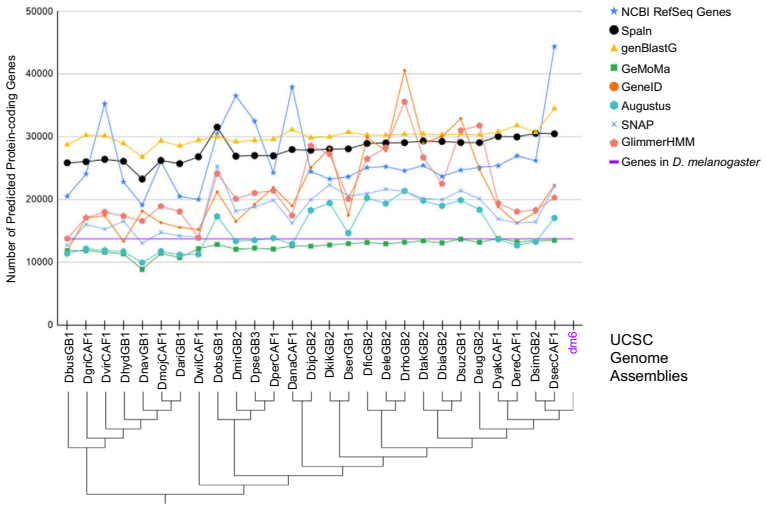

Figure 1. Number of protein-coding genes predicted by different gene predictors for the 27 *Drosophila* species analyzed for the Pathways Project. The number of predicted genes can show large variations across algorithms (algorithm information in extended data[4]) and species, particularly for gene predictors using sequence similarity to genes in a reference species as their primary source of evidence. Some algorithms consistently either predict more (e.g., genBlastG) or less (e.g., GeMoMa) genes than the number of *D. melanogaster* genes as curated by FlyBase (purple line). Prediction differences can partly be attributed to some algorithms predicting a single transcript in a genomic region (e.g., GeneID), while others can predict multiple transcripts per genomic region (e.g., genBlastG, Spaln). The genome assemblies indicated in the cladogram[5,6] correspond to those listed in the "GEP Assembly Identifier" column of Table 1.

Evidence-based gene prediction algorithms (e.g., Gnomon[7]) use extrinsic evidence, such as RNA sequencing (RNA-Seq) data derived from the target species, and sequence similarity to proteins from a reference species, to predict genes within the target assembly. The advent of high-throughput RNA sequencing technologies have led to substantial improvement in the quality of gene annotations,[8,9] particularly for species that lack high-quality gene annotations from a closely related reference species; however, the efficacy of assembling transcripts from RNA-Seq data depends on transcript expression levels in the specific developmental stages and tissue types that are sampled.[10] Long-read RNA sequencing technologies (e.g., Iso-Seq by Pacific Biosciences and Direct RNA sequencing by Oxford Nanopore) can produce reads that span the entire transcript, which facilitates identification of alternative splicing patterns and characterization of different gene isoforms. There are, however, several challenges for producing high quality long-read data including the robustness of RNA extraction methods, bias towards short transcripts, low sequencing throughput, and low read accuracy (reviewed in Ref. 11). Additionally, past studies have shown that transcriptomes constructed from long-read RNA-Seq data have high sensitivity but low precision.[12]

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.[13,14] This article describes a protocol, developed by the Genomics Education Partnership (GEP; https://thegep.org), to engage undergraduate students in the comparative annotation of protein-coding genes involved in the insulin signaling pathway (ISP) across 27 species of *Drosophila*, using *D. melanogaster* as the reference species (Table 1). The species/assemblies described in Table 1 will be updated periodically to reflect the scientific goals of the project, and to conform to the latest NCBI assemblies for those species.

Table 1. Genome assemblies and RNA-Seq data for the comparative analysis of ISP genes in 27 *Drosophila* species and the reference species *D. melanogaster*. For each Assembly, the table shows the corresponding Assembly name, the GEP internal identifier, the NCBI RefSeq Accession Numbers, species names, and BioProject Accession Numbers for the RNA-Seq data.

Assembly Name	GEP Assembly Identifier	NCBI GenBank Assembly Accession	Species Name	NCBI BioProject Accession Numbers for RNA-Seq data
ASM127793v1	DbusGB1	GCA_001277935.1	*D. busckii*	PRJNA274996
dgri_caf1	DgriCAF1	GCA_000005155.1	*D. grimshawi*	PRJNA317989
dvir_caf1	DvirCAF1	GCA_000005245.1	*D. virilis*	PRJNA200701
ASM278046v1	DhydGB1	GCA_002780465.1	*D. hydei*	PRJNA373926
ASM165401v1	DnavGB1	GCA_001654015.1	*D. navajoa*	No RNA-Seq data available
dmoj_caf1	DmojCAF1	GCA_000005175.1	*D. mojavensis*	PRJNA200701
ASM165402v1	DariGB1	GCA_001654025.1	*D. arizonae*	PRJNA395148
dwil_caf1	DwilCAF1	GCA_000005925.1	*D. willistoni*	PRJNA388952
Dobs_1.0	DobsGB1	GCA_002217835.1	*D. obscura*	PRJDB4576
DroMir_2.2	DmirGB2	GCA_000269505.2	*D. miranda*	PRJNA77213
Dpse_3.0	DpseGB3	GCA_000001765.2	*D. pseudoobscura*	PRJNA200701
dper_caf1	DperCAF1	GCA_000005195.1	*D. persimilis*	PRJNA388952
dana_caf1	DanaCAF1	GCA_000005115.1	*D. ananassae*	PRJNA200701, PRJNA72165, PRJNA257286, PRJNA388952
Dbip_2.0	DbipGB2	GCA_000236285.2	*D. bipectinata*	PRJNA63469
Dkik_2.0	DkikGB2	GCA_000224215.2	*D. kikkawai*	PRJNA63469
Dser1.0	DserGB1	GCA_002093755.1	*D. serrata*	PRJNA355616
Dfic_2.0	DficGB2	GCA_000220665.2	*D. ficusphila*	PRJNA63469
Dele_2.0	DeleGB2	GCA_000224195.2	*D. elegans*	PRJNA63469
Drho_2.0	DrhoGB2	GCA_000236305.2	*D. rhopaloa*	PRJNA63469
Dtak_2.0	DtakGB2	GCA_000224235.2	*D. takahashii*	PRJNA63469

Table 1. *Continued*

Assembly Name	GEP Assembly Identifier	NCBI GenBank Assembly Accession	Species Name	NCBI BioProject Accession Numbers for RNA-Seq data
Dbia_2.0	DbiaGB2	GCA_000233415.2	*D. biarmipes*	PRJNA63469
Dsuzukii.v01	DsuzGB1	GCA_000472105.1	*D. suzukii*	PRJNA221549
Deug_2.0	DeugGB2	GCA_000236325.2	*D. eugracilis*	PRJNA63469
dyak_caf1	DyakCAF1	GCA_000005975.1	*D. yakuba*	PRJNA200701
dere_caf1	DereCAF1	GCA_000005135.1	*D. erecta*	PRJNA414017, PRJNA264407
ASM75419v3/ ASM75419v2	DsimGB2	GCA_000754195.3	*D. simulans*	PRJNA200701
dsec_caf1	DsecCAF1	GCA_000005215.1	*D. sechellia*	PRJNA205470, PRJNA414017
Release 6 plus ISO1 MT	dm6	GCA_000001215.4	*D. melanogaster*	PRJNA481740

The GEP is a collaborative effort involving 200+ institutions across the nation. It aims at incorporating active learning into the undergraduate curriculum by implementing Course-based Undergraduate Research Experiences (CUREs) focused on bioinformatics and genomics. The primary objectives of the GEP are: (1) to introduce bioinformatics and genomics into undergraduate education, (2) to offer undergraduates research opportunities, and (3) to foster an inclusive and open partnership. Since its establishment in 2006, the GEP has made significant contributions to the field, advancing our understanding of genome evolution and effective teaching practices in STEM. Through its research initiatives, over 1,100 students have actively participated as co-authors in published scientific research papers.[15–27]

As of January 2023, GEP students from 79 institutions have used this annotation protocol to construct 2,394 gene models across 31 *Drosophila* species (number of species/gene combinations that have been constructed can be found in Supplement 7). Despite differing in instructional settings and teaching modalities, 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.[28] Reconciled gene models will typically be described in *microPublication* articles[29] and submitted to the NCBI Third Party Annotation (TPA) database.[30] Researchers can utilize the high-quality, manually curated gene models constructed by GEP students to investigate the evolution of genes and genomes (Figure 2).

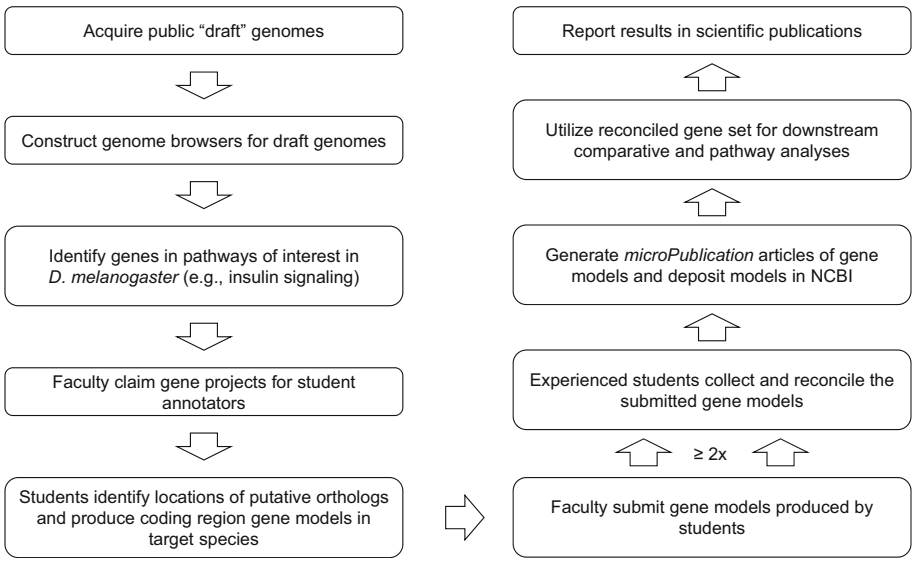

Figure 2. Gene Model Creation Workflow. Summarized workflow for annotation and reconciliation to produce a set of high-quality gene models suitable for comparative and pathway analyses.

Methods

Overview of the coding region annotation protocol

Drosophila researchers and curators at FlyBase have produced high-quality, comprehensive gene annotations for *D. melanogaster* based on a large amount of genetic and sequencing data.[31] Our protocol utilizes the gene annotations from *D. melanogaster* (reference species) to facilitate annotation of the protein-coding sequence (CDS) 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 reference species (i.e., construct the most parsimonious gene model assuming evolutionary conservation).

In order to generate manual annotations for the CDS of a gene in the target species, we need to (1) identify the ortholog of that gene from *D. melanogaster* in the target species using sequence similarity and local synteny, (2) determine the structure and approximate coordinates of each isoform and their coding exons, and then (3) refine those coordinates for each isoform. The key analysis steps are also summarized in the Results section and in the Annotation Workflow for the Pathways Project.[32] A walkthrough illustrating each step of the annotation protocol from the perspective of a naive student annotator on an example gene is both available on the Pathways Project page of the GEP website (https://thegep.org/pathways/) and in Ref. 33. Here we highlight the essential conceptual steps the student annotator will follow.

The annotation and reconciliation protocols described below utilize multiple bioinformatics tools that are briefly summarized in the "Data (and Software) Availability" section.

CDS annotation procedure

Project claiming

GEP faculty members select at least one *D. melanogaster* gene involved in the ISP (i.e., target gene in the reference genome) and one or more of the 27 *Drosophila* species (i.e., target species) for their students to annotate. Each gene project includes an estimated difficulty level based on the number of isoforms, number of coding exons, and evolutionary distance from the reference genome. Faculty members can take the estimated difficulty of the gene projects into consideration when selecting projects that best suit the pedagogical goals of their courses, the amount of class time devoted to the annotation project, the academic levels of their students, and specific interests in the biological function of a gene. For example, faculty members might select the same gene in multiple *Drosophila* species for their students to annotate (working individually or in groups) in order to teach students about conservation relative to divergence time.

Identify the ortholog

Ortholog assignment in the target species is based on the analysis of protein sequence similarity and local synteny (i.e., relative gene order and orientation within a syntenic chromosomal region) compared to the reference species.[34,35] The key analysis steps for identifying the ortholog are also summarized in the extended data.[36]

To identify the ortholog of the target gene, the student annotator examines the genomic neighborhood surrounding the gene in both the reference species and the target species using the GEP UCSC Genome Browser (further described in the Software section below; https://gander.wustl.edu). This analysis includes identifying the nearest two upstream and downstream genes and their orientations relative to the target gene. The local synteny analysis also includes any nested genes in the locus surrounding the putative ortholog in the target species.

Locating the putative ortholog requires the student to obtain the protein sequence for the target gene in the reference species from the Gene Record Finder, and use it as the query to perform a *tblastn* search against the genome assembly of the target species via NCBI Web BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). This protocol uses the *tblastn* program to compare the protein sequence query against a nucleotide database because amino acid sequences show higher sequence conservation than nucleotide sequences across evolutionary time. In addition, due to the degeneracy of the genetic code, sequence similarity searches at the amino acid level are more sensitive than searches at the nucleotide level.[37,38] There are three possible outcomes of their *tblastn* search: (1) zero matches, (2) one match, or (3) more than one match. If there is a single high-quality match, that match is a good candidate for being the ortholog.

The *tblastn* search may report zero significant matches even if the ortholog exists in the target species due to gaps or misassemblies in the genome assembly or the lack of sequence conservation between orthologs. In the latter case, the two genes upstream and downstream of the target gene in the reference species will be used to infer the location of the target

gene in the target species (i.e., local synteny). This is done for *Ilp3* in *D. grimshawi* as shown in Figure 3A. Here, part of the *Ilp3* CDS erroneously maps to the end of the *Ilp1* CDS, and does not map to the location assessed as correct by conforming to the conservation of local synteny. Figure 3B also shows the change in duplication of *Ilp2* in the local neighborhood of *D. grimshawi* (as indicated by XM_001983645 and XM_001983644 for *Ilp2-p* and *Ilp2-d* respectively). The student first identifies the orthologs to the genes flanking the target gene in the target species. If there is a locally syntenic region in the target species, then there will likely be an additional feature located between the flanking orthologs in the target species. This additional feature can then be considered as the likely ortholog to the target gene in the target species based on local synteny. Orthology should be confirmed using a *blastp* search of the predicted protein sequence of the feature in the target species against a database of annotated proteins for the reference species. If the location and candidate for the ortholog cannot be established using this strategy, then the ortholog may be absent from the genome assembly for the target species.

If the *tblastn* search reports two or more significant matches, then the ortholog assignment will be based on parsimony with the reference species. The match with the lowest E-value, highest percent identity, and highest alignment coverage to the target gene will be assigned as the putative ortholog in the target species. The student then confirms the putative ortholog assignment derived from the best *tblastn* match via local synteny by comparing the genes surrounding the target gene in the target and reference species. If there is no match for neighboring genes in an evolutionarily diverged species but a single match exists for the target gene, then the ortholog assignment will be determined by BLAST searches using the Reciprocal Best Hits (RBH) strategy.[39]

Figure 4 shows an example of how the student can use local synteny to establish the ortholog of the target gene *Ilp2*. The coding region of the same two genes are located upstream (i.e., *Zasp67* and *Ilp1*) and downstream (i.e., *Ilp3* and *Ilp4*) of

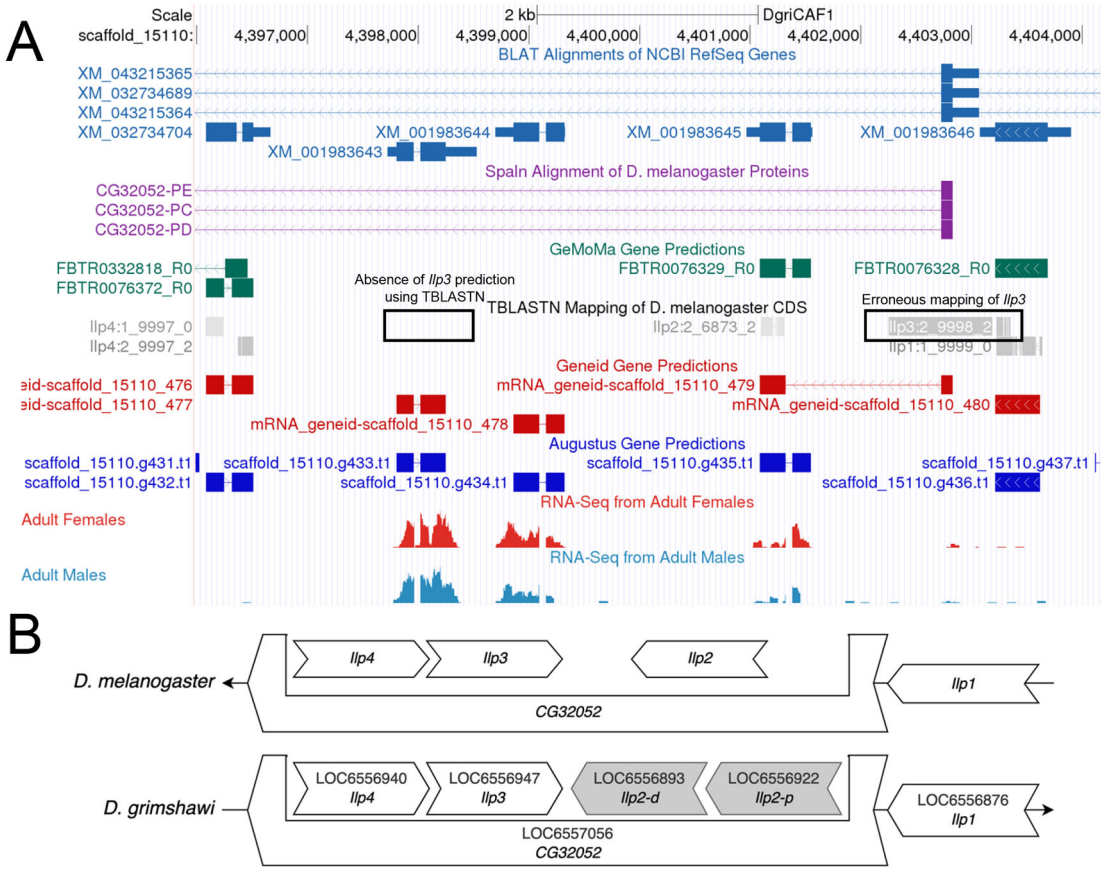

Figure 3. Identifying the Orthologoy of *Ilp3* in *D. grimshawi* based on local synteny. (A) The genome browser for the DgriCAF1 assembly of *D. grimshawi* shows that the *tblastn* mapping of the CDS of *Ilp3* does not match the correct location of *Ilp3* (based on local synteny). Instead, it aligns to the end of the CDS for *Ilp1*. (B) The comparison of the local neighborhood of *Ilp3* in *D. melanogaster* and *D. grimshawi* shows that *Ilp3* should be nested within *CG32052*, and should be between of *Ilp4* and both copies of *Ilp2* in *D. grimshawi*.

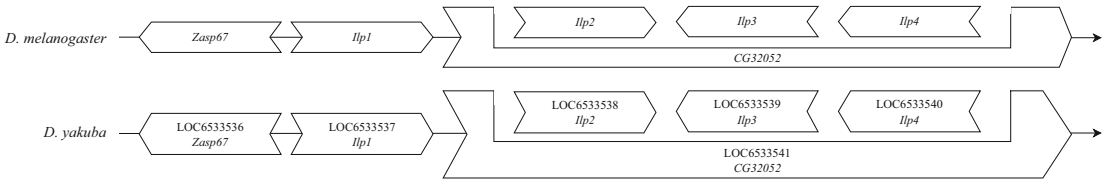

Figure 4. Local synteny analysis of the *Ilp2* gene in *D. yakuba*. Schematic of genomic neighborhood surrounding the *Ilp2* gene in *D. melanogaster* and *D. yakuba*, showing that *Ilp2*, *Ilp3*, and *Ilp4* are nested within *CG32052* in both species.

Ilp2 in *D. melanogaster* and *D. yakuba.* In addition, *Ilp2, Ilp3*, and *Ilp4* are all nested within the coding span of *CG32052* in both species, which further supports the hypothesis that the genomic neighborhood is conserved, and thus, the ortholog of *Ilp2* has likely been identified.

Identify the approximate coordinates of each coding exon

Once the student identifies the putative ortholog, they begin constructing the gene model by separately mapping each coding exon to determine their approximate locations and their reading frames in the target genome using the "Align two or more sequences" (*bl2seq*) feature provided by NCBI Web BLAST.

For each isoform of the target gene in *D. melanogaster* (reference genome), they perform *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, is identified by the "Identify the ortholog" step above. For scaffolds larger than 10 Megabases (Mb), 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 approximate location of the target gene 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, the student usually identifies a collinear set of coordinates for most of the coding exons of the isoform. Results of the *tblastn* searches provide further supporting evidence for the ortholog assignment and provide anchors from which to define the search regions for small or weakly conserved coding exons.

Refine coding exon coordinates

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 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. The student manually refines the collinear coding exon coordinates identified above through visual inspection of the region in the Genome Browser utilizing the RNA-Seq track, other homology-based alignment algorithms, and gene predictions, as well as comparing the sequence to other *Drosophila* species based on whole genome multiple sequence alignments (more details about the tracks used can be found in the extended data).[4]

If available, the RNA-Seq data can provide empirical support for the proposed gene model in the target species. RNA-Seq data for the target species were obtained from the NCBI Sequence Read Archive (Table 1[40]). The RNA-Seq reads were

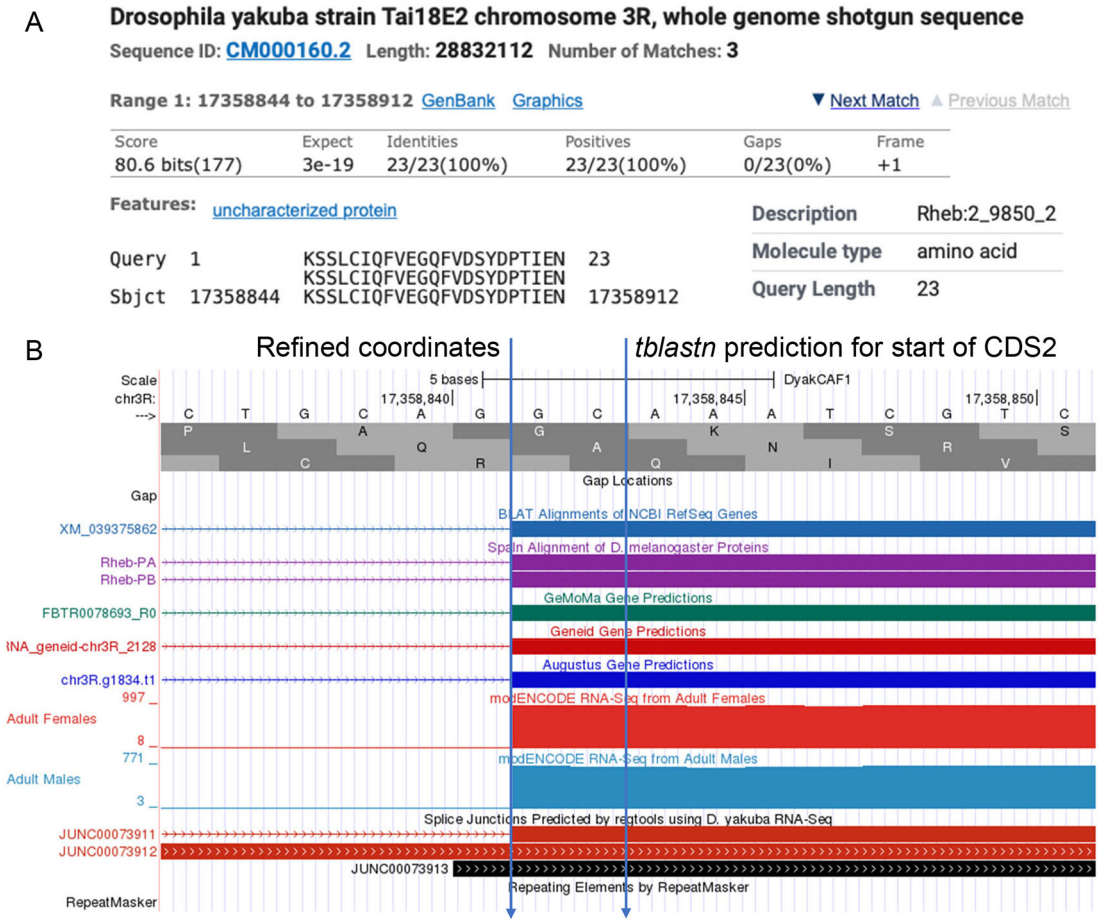

Figure 5. Refined Coordinates for CDS 2 of *Rheb* in *D. yakuba*. (A) The *tblastn* alignment for the *D. melanogaster* CDS 2_9850_2 (query) against the *D. yakuba* scaffold CM000160.2 (chr3R) placed the start of the CDS at 17,358,844 in frame +1. (B) Examination of the other lines of evidence (e.g., RNA-Seq read coverage, splice junction predictions, gene predictions, protein alignments) using the Genome Browser placed the start of the coding exon at 17,358,842. Since there are two nucleotides (GC; blue lines) prior to the first complete codon (AAA) that codes for first amino acid in the *tblastn* alignment (K), CDS 2_9850_2 in *D. yakuba* has a phase 2 splice acceptor site relative to reading frame +1.

mapped against the target genome using HISAT2,[41] and putative introns were inferred from the alignments of spliced RNA-Seq reads using the *junctions extract* command provided by RegTools.[42]

When defining the intron coordinates, canonical splice sites (i.e., 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 clade. The student also needs to ensure the donor and acceptor sites have compatible phases (i.e., the sum of the phases of the donor and acceptor sites of an intron is either zero or three) in order to maintain the open reading frame after splicing. A more detailed version of the workflow for refining the coding exon coordinates, from the perspective of a naive student, can be found within extended data.[43] Figure 5 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 comparison of the approximate coding exon coordinates determined by the *tblastn* searches with their refined counterparts for the *Rheb* gene in *D. yakuba.*

Verify gene model

The student uses the Gene Model Checker to verify that the refined coordinates for the gene model in the target species satisfy the biological constraints for protein-coding genes in most eukaryotes and reflect the gene structure of the *D. melanogaster* ortholog. The dot plot and protein alignment identify differences between the proposed gene model and the *D. melanogaster* ortholog, and they help to verify that their proposed gene model is the most parsimonious compared

Table 2. Refining approximate *tblastn* coordinates for coding exons of *Rheb* in *D. yakuba* (DyakCAF1). Comparison between the *tblastn* and refined coordinates for the model of *Rheb* in *D. yakuba.* Note that in most cases, the coordinates identified by the *tblastn* search are adjusted by a few nucleotides (difference columns) via visual curation to account for incomplete codons and over-extensions of the *tblastn* alignment.

Exon Number	Approximate coordinates from *tblastn* search		Refined exon coordinates		Difference (refined - *tblastn*)	
	Start	Stop	Start	Stop	Start	Stop
1	17,358,666	17,358,713	17,358,666	17,358,714	0	1
2	17,358,844	17,358,912	17,358,842	17,358,913	-2	1
3	17,359,013	17,359,216	17,359,011	17,359,218	-2	2
4	17,359,279	17,359,407	17,359,278	17,359,407	-1	0
5	17,359,470	17,359,559	17,359,470	17,359,559	0	0

to the *D. melanogaster* ortholog. For more information on the Gene Model Checker, refer to the "Data (and Software) Availability" section.

The student repeats the "Identify the approximate coordinates of each coding exon" and "Refine coding exon coordinates" steps to construct gene models for each unique protein-coding isoform of their target gene and then verifies them with the Gene Model Checker.

Final submission

The student submits a file containing the coordinates of the coding exons for all isoforms in the Generic Feature Format (GFF), a file containing the transcript sequence for the coding region of all isoforms in FASTA format (FNA), and a file containing the peptide sequence for all isoforms in FASTA format (FAA). These files are generated for each isoform by the Gene Model Checker and then concatenated by the Annotation Files Merger tool described below. They also submit an Annotation Report[44] to document the evidence supporting their proposed gene models.

Exceptions to the standard annotation workflow

While the standard annotation workflow provides a good starting point for the annotation of target genes in the Pathways Project, additional tools and strategies are needed to address challenges with the annotations of a subset of genes. Resolving these challenges typically require the integration of multiple lines of empirical and computational evidence. Below we describe the process for resolving two common challenges—non-canonical splice sites and assembly errors—and provide a list of other potential challenges annotators may encounter.

Non-canonical splice sites

The most common (canonical) sequence for the splice donor site is GT (GU in the pre-mRNA), and the most common sequence for the splice acceptor site is AG. Variant splice sites are termed non-canonical. For example, the GC splice donor site appears in ~0.8% (603/71,922) of the unique introns in *D. melanogaster* (FlyBase release 6.43). 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 reference species (*D. melanogaster*), or in multiple *Drosophila* species closely related to the target species, can also be used as supporting evidence for the annotation of a non-canonical splice site.

Assembly errors

Each sequencing platform (e.g., Sanger, Illumina, PacBio, and Nanopore) has a distinct error profile[45,46] 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.[47] Gene annotation challenges caused by assembly errors include partial or missing genes due to gaps in the assembly, apparent frameshifts within CDSs due to extra or missing nucleotides in the consensus sequence, and errors in ortholog/paralog assignments (e.g., due to "duplications" caused by misassemblies).

The publicly available genome assemblies utilized by the Pathways Project were 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 proposed gene models for the target species includes 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, the student evaluates multiple lines of evidence including: (1) sequence conservation with other *Drosophila* species besides *D. melanogaster*, (2) consistency with Illumina genomic reads and RNA-Seq reads in the NCBI Sequence Read Archive (SRA), and (3) consistency with other genome assemblies for the same species.[48]

Other exceptions

Past studies have shown that the number of genes involved in the ISP varied in the different *Drosophila* species due to gene duplications and pseudogenization,[49] which leads to challenges in ortholog assignments. Other annotation challenges are caused by changes in gene structure (e.g., gain or loss of coding exons and isoforms) compared to the target gene in the reference species.

Another set of challenges pertain to gene models that do not conform to the typical characteristics of protein-coding genes, including genes with a non-canonical start codon (e.g., *Akt*), stop codon readthrough (e.g., *jim*), or trans-splicing (e.g., *mod (mdg4)*).

Common issues in gene models prior to submission

The Gene Model Checker sometimes reports multiple "fails" for a proposed gene model because it deviates from the expected biological characteristics of most protein-coding genes (e.g., gene model with multiple in-frame stop codons). Typically, the multiple failures can be attributed to an error in the upstream regions of the proposed gene model that propagates downstream.

For example, 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 next 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 (ORF) after the intron has been spliced out, the sum of the donor and acceptor phases for adjacent coding exons must either be zero (i.e., no extra codon) or three (i.e., one extra codon). Selecting incompatible donor and acceptor splice sites causes a frameshift that changes the reading frames of the coding exons downstream of that splice junction. Using the incorrect reading frame to translate the downstream coding exons will likely introduce stop codons in the translation, thereby triggering multiple failures in the Gene Model Checker.

To resolve gene models with multiple fails, the student starts 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 of the gene models produced by GEP 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 under the guidance of a GEP faculty, and/or senior-scientist, mentor.

Each target gene in a target species is annotated by at least two students working independently. This quality control step is predicated on the assumption that it is relatively common for one student to make a single error but relatively rare for multiple students working independently to make the same error.

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 and missing specific exons or isoforms).

Reconciliation process

Reconciliation is performed using Apollo,[28] a web-based collaborative genome annotation editor that allows reconcilers to view student-generated models alongside the evidence tracks (Figure 6) used for annotating those models.

Reconcilers evaluate the available student annotations for each isoform in conjunction with the other evidence tracks (e.g., sequence similarity, RNA-Seq data, and gene predictions) to construct the final gene model for the protein-coding isoform(s) that is best supported by the available evidence. Reconcilers then draft a *microPublication* describing the supporting evidence for the final gene model.[29] The student annotators and their faculty mentors review and approve the article draft. Reconciled models are also used in downstream meta-analyses and deposited into the NCBI Third Party Annotation (TPA) database. TrackHub links are currently hosted on GEP servers and all the models that have been reconciled to date have TrackHub links listed in Supplement 8.

Results

CDS annotation procedure

Identify the ortholog

Identification of the ortholog is done using BLAST and local synteny analysis of the genomic neighborhood of the target gene. For example, to locate the *Ilp2* gene (target gene) in *D. yakuba* (target species), a *tblastn* search was used to compare the protein sequence for *D. melanogaster* Ilp2-PA (query) against the *D. yakuba* DyakCAF1 genome assembly (subject). The two collinear alignment blocks that correspond to the best match to the *D. melanogaster* Ilp2-PA protein (137aa) are located in the 9,766,395-9,766,887 region of the *D. yakuba* scaffold CM000159.2 (chr3L) (Figure 7). The alignment block for the first 54aa of Ilp2-PA is located at 9,766,395-9,766,556 in frame +3, with a normalized score of 171 bits and 94% identity. The alignment block for residues 56-137 of Ilp2-PA is located at 9,766,642-9,766,887 in frame +1, with a normalized score of 261 bits and 94% identity. The joint E-value for the two collinear alignment blocks is 3e-114. The two alignment blocks account for 136aa out of 137aa of *D. melanogaster* Ilp2-PA (residue 55 of the protein is not covered by the two alignment blocks). Local synteny analysis shows that the genomic neighborhood surrounding *Ilp2* is conserved between *D. melanogaster* and *D. yakuba* (Figure 4).

Collectively, the available evidence supports the hypothesis that the 9,766,395-9,766,887 region of the *D. yakuba* scaffold CM000159.2 contains the putative ortholog of *Ilp2*.

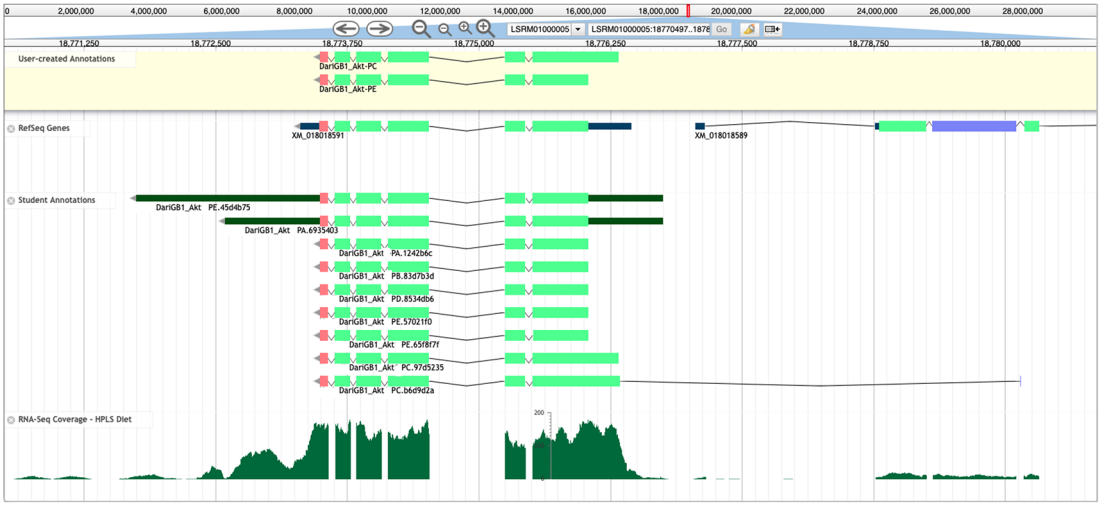

Figure 6. Apollo Screenshot for the *Akt* gene model in *D. arizonae*. Final gene models for *Akt* in *D. arizonae* (User-created Annotations track, yellow background), along with the NCBI RefSeq gene model (RefSeq Genes track), submitted student models (Student Annotations track), and RNA-Seq data aligning to the region (RNA-Seq Coverage – HPLS Diet track, histograms). Despite RefSeq predicting only one isoform (XM_018018591) for this gene, the final model contains two unique protein-coding isoforms (Akt-PC and Akt-PE), which were annotated using multiple lines of evidence. The Akt-PC isoform has a larger coding region in the reconciled gene model that is missed by the RefSeq gene predictions. In the Student Annotations track, the top two models included annotations of the 5' and 3' untranslated regions (UTRs, thinner dark green rectangles), and the bottom student model (DariGB1_Akt-PC. b6d9d2a) incorporated a coding exon that was likely not part of *Akt*.

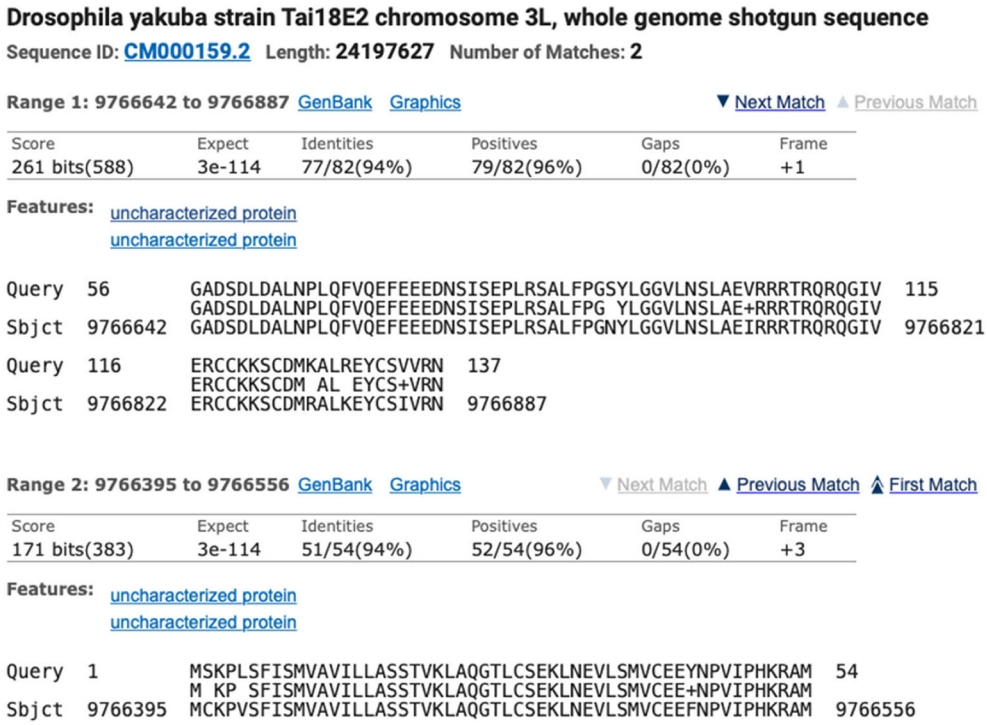

Figure 7. *tblastn* alignment for *Ilp2* from the *D. melanogaster* dm6 assembly (query) against the *D. yakuba* DyakCAF1 assembly (GCA_000005975.1; subject).

Identify the coordinates of each coding exon

The approximate coding exon coordinates for the target gene in the target species are defined by *tblastn* searches of the coding exons of the target gene in *D. melanogaster* against the genomic scaffold in the target species determined by the "Identify the ortholog" step. The approximate coding exon coordinates are then refined by examining the evidence tracks in the GEP UCSC Genome Browser. For example, Figure 5A shows the approximate placement of the second coding exon of the *Rheb* gene in the *D. yakuba* DyakCAF1 assembly based on the results of the *tblastn* search (i.e., scaffold CM000160.2 (chr3R) at 17,358,844-17,358,912 in frame +1). In Figure 5B, the refined start coordinate for the second coding exon of *Rheb* (i.e., at 17,358,842) was determined by RNA-Seq read coverage, splice junction predictions, gene predictions, and protein alignments evidence tracks on the GEP UCSC Genome Browser. Table 2 shows the refinement of all CDS exons of *Rheb* in *D. yakuba.* The "Difference" column indicates the necessity of manually refining BLAST derived coordinates.

Verify gene model

The final step prior to submission to the GEP is for the student to confirm the proposed gene model using GEP's Gene Model Checker. This tool can help the annotator verify that the proposed gene model satisfies the biological constraints of most protein-coding genes and identify differences between the proposed gene model and the *D. melanogaster* ortholog. In Figure 8, we can see the checklist for the model of Rheb-PA in *D. yakuba* passed the Gene Model Checker (i.e., the proposed gene model begins with a start codon, ends with a stop codon, and the five coding exons use the canonical splice donor and acceptor sites). It is, however, possible to "pass"' all the checks in the tool and still have an entirely incorrect model since the tool does not specifically test for protein sequence conservation or orthology.

Most common annotation errors

Reconciliation consists of reviewing two or more gene models created by student annotators working independently. The main advantage of manually curated gene models relative to computational predictions is the ability for the curator (in this case the student annotator) to evaluate and integrate across non-conforming pieces of evidence. Reconcilers provide further quality control measures as they closely scrutinize each idiosyncrasy in the gene models proposed by student annotators. The most common idiosyncrasies/errors are listed below.

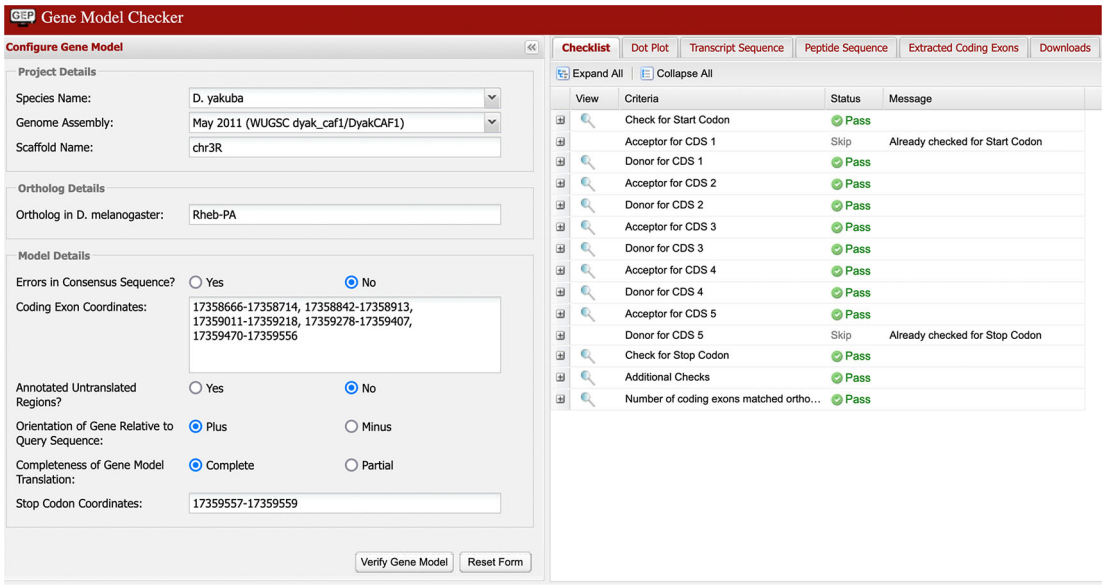

Figure 8. Confirmation that the coding exon coordinates of the putative ortholog of Rheb-PA in the *D. yakuba* DyakCAF1 assembly (GCA_000005975.1) follow molecular biology rules for protein-coding genes and match the expected number of exons in *D. melanogaster* using the GEP's Gene Model Checker.

Selection of incorrect splice sites

One of the most common annotation errors is the selection of splice donor and acceptor sites that are not the most parsimonious candidates compared to the target gene in the reference genome. Another type of splice site error is caused by the annotators placing too high of a priority on the use of canonical sequences for splice donor sites (GT) and splice acceptor sites (AG). While non-canonical splice sites are rare in *Drosophila* (i.e., found in < 1% of introns), they should be used if the non-canonical splice sites are supported by RNA-Seq data and conservation in other *Drosophila* species.[50–52]

Multiple failures in the Gene Model Checker checklist (as shown in Figure 9) are typically caused by the selection of incompatible splice sites during the "Refine coding exon coordinates" stage of the analysis. Most of these errors can be resolved by scrutiny of the coordinates for the checklist item where the first error is reported by the Gene Model Checker. Figure 10 shows the incorrect assignment of a splice site in *D. eugracilis* indicated by the first "fail" in the Gene Model Checker in Figure 9. The proposed splice donor site is one nucleotide away from the correct splice donor site with the canonical sequence of GT. Reconcilers are well equipped to identify the proper splice site when the student model has an error.

Missing/extra exons

Another common error is gene models with missing or extra coding exons compared to the *D. melanogaster* ortholog. Sometimes these proposed changes in gene structure are well supported by the data (e.g., a novel intron supported by splice junction predictions), but often the mismatch in the number of coding exons is due to the student annotator failing to fully account for all lines of evidence.

Incorrect ortholog assignment

Correct assignment of an ortholog depends on the proper configuration of the BLAST search parameters, proper interpretation of the BLAST search result, and proper consideration of local synteny compared to the reference species. If a student annotates a gene that is not the ortholog, then their gene model cannot be used to establish the final ortholog model in the target species. In instances where the student annotator has failed to annotate the correct ortholog for a given gene, the reconciler asks another member of the reconciliation team to generate a new annotation for the offending model, thus the team member creates another independent annotation to give us at least two reasonable models. This additional model is then incorporated into the standard reconciliation pipeline to identify congruence with another already submitted model.

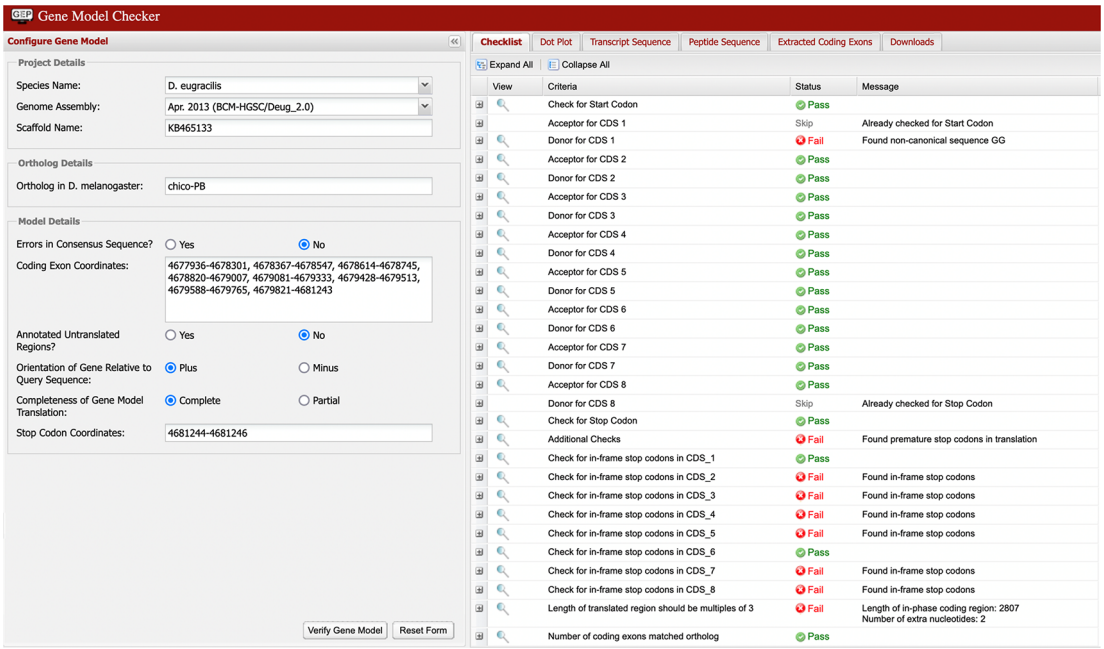

Figure 9. The Gene Model Checker checklist indicated the presence of in-frame stop codons within coding exons 2, 3, 4, 5, 7, and 8 of the proposed gene model for chico-PB in *D. eugracilis*. The checklist also reports the use of a non-canonical GG splice donor site for CDS 1, and that the total length of the coding region (2,807 nt) is not divisible by three. To address these errors, the annotator should verify the annotation for the item associated with the first error in the checklist (i.e., the end coordinate for the first coding exon and its corresponding splice donor site). In this example, the end of the first coding exon should be changed from 4,678,301 to 4,678,302 based on the available evidence on the GEP UCSC Genome Browser. This change will resolve the remaining failures in the checklist.

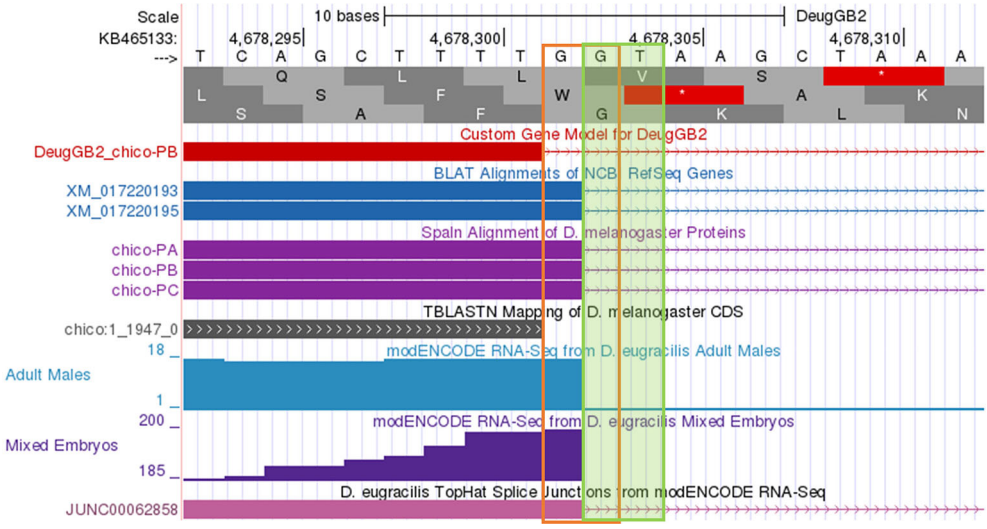

Figure 10. Use of incorrect splice site (GG) in the annotation of the *chico* ortholog in *D. eugracilis*. Extending the end of the proposed coding exon by one nucleotide (to 4,678,302) will utilize the canonical splice donor site (GT) and is more consistent with the available gene predictions and RNA-Seq data. The revised coding exon coordinate is supported by the splice junction prediction JUNC00062858 (score = 216).

Missing/Mislabeled Isoforms

The annotation protocol assumes that all isoforms for a given gene from *D. melanogaster* will also be present in the target species unless there is evidence to the contrary. Orthologous isoform names are assigned based on sharing strong protein sequence conservation and a similar coding exon structure to the *D. melanogaster* isoform. If a student fails to annotate an

isoform model that is viable in the target species genome, the student's model is categorized as missing an isoform. Evidence that an orthologous isoform is not present in the target species might include that there are no viable splice sites to generate the orthologous isoform and/or an exon unique to that isoform is not present. If relevant, such evidence should be documented in the student's report form and the subsequent *microPublication* of the reconciled model.

Exceptions to the standard annotation workflow

Non-canonical splice sites

While non-canonical splice sites are rare in *Drosophila* (i.e., found in < 1% of introns), they should be included if they are supported by RNA-Seq read coverage, splice junction predictions, and conservation in other species.[50–52] An example of a gene with a non-canonical spice site is sgg-RN in *D. melanogaster*, which has a GC splice donor site rather than the GT donor site between exons sgg:15 and sgg:18 (Figure 11). The presence of a non-canonical splice site in a proposed model will be indicated by a "Warning" label in the Gene Model Checker. If a non-canonical splice site is substantiated by multiple lines of evidence (e.g., RNA-Seq data, conservation across multiple species), the student may conclude that the non-canonical splice site is also present in the target species.

Assembly errors

Assembly errors (e.g., gaps) can affect the coding region gene annotations. The Chained Alignments between the *D. melanogaster* dm6 assembly and the *D. pseudoobscura* DpseGB3 assembly (GCA_000001765.2) show that the *chico* gene is split across two different scaffolds in the *D. pseudoobscura* DpseGB3 assembly—the first coding exon of *chico* is located in the *D. pseudoobscura* scaffold CH673091 and the rest of the gene is located in scaffold CH475478 (Figure 12). These assembly errors will be provided to the TPA database as a BankIt submission that contains a VCF (Variant Call

Introns with Non-canonical Splice Sites

Transcript Name	FlyBase ID	Splice Donor	Splice Acceptor
sgg-RN	intron_sgg:15_sgg:18	GC	AG

Figure 11. The "Introns with Non-canonical Splice Sites" section of the Gene Record Finder for the N isoform of *sgg* shows the presence of a non-canonical GC splice donor site in *D. melanogaster*.

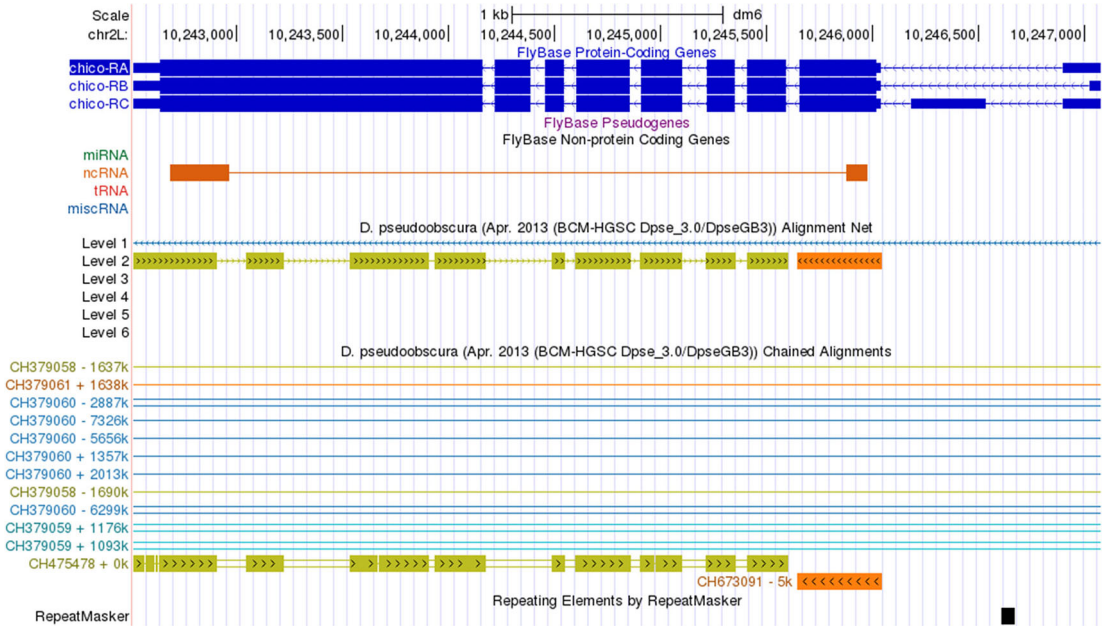

Figure 12. The Chained Alignments track in the *D. melanogaster* dm6 assembly shows that *chico* in the *D. pseudoobscura* DpseGB3 assembly (GCA_000001765.2) has been split across two scaffolds (CH475478 and CH673091).

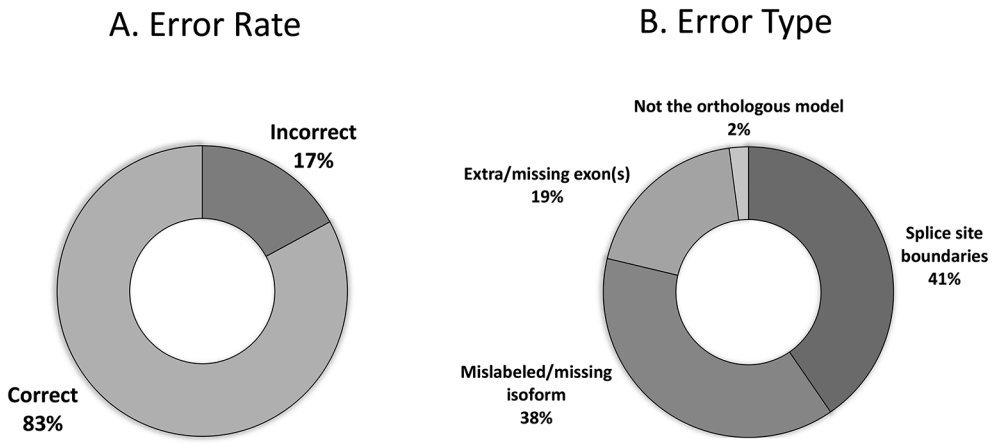

Figure 13. Error rate of student models that reach reconciliation (n=310). (A) The raw error rate of models, and the (B) breakdown of the error rates into the classes of errors.

Format) file that contains the suggested changes. The reason behind these changes will be detailed in the corresponding microPublication.

Reconciliation

The primary goal of reconciliation is to elucidate any inconsistencies or idiosyncrasies that might occur with two independent students generating the models. The majority of student models received for the Pathways Project (n=310) agree with the final reconciled model (83%; Figure 13A). Those with errors include mislabeled or missing isoforms (38%), incorrect splice sites (41%), extra or missing exon(s) (19%), and failing to identify the proper ortholog (2%) as seen in Figure 13B.

microPublication

After student models are reconciled and manuscripts are drafted, they are sent for review and publication via the *microPublication* journal (https://www.micropublication.org). The goal of this journal is to publish brief and novel findings whose results may lack a broader scientific narrative. The gene models generated and reconciled by students fit neatly within this description (e.g., Ref. 53).

Discussion

This annotation protocol has been used by GEP faculty to engage undergraduate students in the comparative annotation of ISP genes in 27 *Drosophila* species. Similar protocols have previously been used by the GEP on other scientific projects.[18,21,25,54] Those protocols were used as the basis to develop the protocol described here.

To ensure 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 ISP).

High-throughput computational gene annotators are extremely valuable at the scale of whole genomes. However, the annotation effort of the student annotators is of value because the students can outperform computational gene predictors at the level of an individual gene, or a small set of genes. For example, GEP students were able to annotate a conserved isoform not predicted by the NCBI RefSeq pipeline for *Akt* in *D. arizonae* (Figure 5). This isoform was not predicted computationally due to a non-canonical ACG start codon that is conserved in *Akt* in *D. melanogaster*.

Annotation and reconciliation protocols similar to the one described in this article are also currently used to investigate the expansion of the Muller F element in four *Drosophila* species and the evolution of venom proteins in four parasitoid wasp species. 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.

The protocol described here is focused only on the annotation of coding regions. Once additional experimental data (e.g., CAGE, RAMPAGE, and long-read RNA-Seq data) becomes available in more species, future analyses will focus on the annotation of untranslated regions and transcription start sites.

Genomics is a rapidly growing field and the research opportunities and careers in genomics are likely to increase. Therefore, the opportunity to provide students high-quality educations in genomics is required. Hands-on approaches to teaching about genomics both enhances student learning and provides us with additional high-quality datasets.[16,17] The GEP provides the central support structure to aid in genomics education that enhances the ability of faculty bring genomics research into their curriculum.[20] Lopatto et al. show that students benefit from learning through formative frustration and iteration within the GEP -CURE framework.[26,27]

Data and software availability

Repository-hosted data

Repository: NCBI

The GenBank accession numbers for the assemblies and BioProject accession numbers for the RNA-Seq datasets are available on NCBI under the accession numbers listed in Table 1. Data is available in the public domain through NCBI (https://www.ncbi.nlm.nih.gov).

Repository: FlyBase

FlyBase data used for tool creation mentioned below is publicly available through the FlyBase FTP site (https://ftp.flybase.net/releases/).

Data that cannot be shared

Data under license by a third party

The UCSC Genome Browser is developed by the Genomics Institute at the University of California Santa Cruz. The source code for the UCSC Genome Browser is covered by five different licenses. Most of the source code is available under the MIT License, and all the source code is freely available to non-commercial entities. The source code is available on GitHub (https://github.com/ucscGenomeBrowser/kent), and it can also be obtained through the UCSC Genome Browser Store (https://genome-store.ucsc.edu/).

Large data

The datasets displayed in the GEP UCSC Genome Browser are too large to be feasibly hosted by an F1000Research-approved repository, such as RNA-Seq read alignments and Whole-Genome Multiple Sequence Alignments. The data is available through the GEP UCSC Genome Browser and the UCSC Table Browser (https://gander.wustl.edu). Details on the datasets and tools used to construct each evidence track are available through the settings page for each track in the GEP UCSC Genome Browser. As stated under the "Repository-hosted Data" section, the genome assemblies and the RNA-Seq datasets used to construct the evidence tracks on the GEP UCSC Genome Browser are in the public domain, and they can be obtained through NCBI (https://www.ncbi.nlm.nih.gov).

Software

Repository: GitHub

- Annotation Files Merger (https://github.com/wilsonleung-gep/annotation-files-merger)
- Dot Plot Viewer (https://github.com/wilsonleung-gep/dot-plot-viewer)
- Gene Record Finder (https://github.com/wilsonleung-gep/gene-record-finder)
- Sequence Updater (https://github.com/wilsonleung-gep/sequence-updater)
- Small Exons Finder (https://github.com/wilsonleung-gep/small-exons-finder)

All the source code in the above repositories is available under the MIT License.

- Gene Model Checker (https://github.com/wilsonleung-gep/gene-model-checker)

  The Gene Model Checker is available under the GNU General Public License v3.0.

Table 3. Software Version Information and release dates for the web-based applications used by the manual gene annotation protocol for the Pathways Project.

Software	Version	Release Date
GEP UCSC Genome Browser	v400; FlyBase 6.43	Dec. 31, 2021
Gene Model Checker	v2.0; FlyBase 6.43	Dec. 31, 2021
Sequence Updater	v2.0	Jan. 15, 2021
BEDTools	v2.30.0	Jan. 23, 2021
Gene Record Finder	v1.3; FlyBase 6.43	Dec. 31, 2021
Small Exons Finder	Prototype; v1.0	Dec. 31, 2020
Annotation Files Merger	v2.0; FlyBase 6.43	Dec. 31, 2021
BLAST+	v2.12.0	June 28, 2021

All custom software and tools generated by the GEP can be accessed from the GEP website (https://thegep.org). The annotation tools are publicly available web-based applications, thus requiring no installation on the part of the user. Versions of the tools can be found in Table 3. All the GEP annotation tools are synchronized to the same FlyBase release. The FlyBase *D. melanogaster* gene annotations used by the GEP annotation tools are updated twice a year (i.e., just before the start of the Fall and the Spring semesters) in order to mitigate the impact of FlyBase updates to the *D. melanogaster* reference gene annotations during the semester.

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 these Genome Browsers[19] with multiple evidence tracks using data generated by various algorithms[4] and experimental data obtained from the NCBI SRA.

Gene Record Finder

The Gene Record Finder (https://gander.wustl.edu/%7ewilson/dmelgenerecord/index.html) 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 exon. The Gene Record Finder also provides exon usage maps that demarcate the exons used by each isoform. This information is used in the "Identify the approximate coordinates of each coding exon" step of the annotation protocol. While this information can be directly obtained from FlyBase, annotators can more easily 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 (https://gander.wustl.edu/%7ewilson/genechecker/index.html) 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 ORF, uses canonical start and stop codons, and canonical splice donor and acceptor sites). It also indicates whether the number of coding exons for the proposed gene model in the target species matches that in the orthologous isoform in *D. melanogaster*. 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, or polycistronic transcripts[31]), students provide the supporting evidence for their 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 (a GTF file to define genomic feature coordinates, a transcript sequence file in FASTA format (FNA), and a peptide sequence file in FASTA format (FAA)).

The Gene Model Checker requires input of the species, assembly, and the scaffold of the putative ortholog. In addition, the Gene Model Checker requires the name of the ortholog in *D. melanogaster,* the set of coding exon coordinates for the gene in the target species, orientation of the gene, noting whether or not the untranslated regions are included, whether 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 (e.g., Figure 8). Presently, this tool only supports the 28 *Drosophila* species that are currently in the GEP UCSC Genome Browser.

Small Exons Finder

The typical use case for the Small Exons Finder (https://gander.wustl.edu/%7ewilson/smallexonfinder/index.html) tool is identifying CDSs that are too small or too weakly conserved to be detected by BLAST. The tool is designed to look for ORFs that satisfy a set of biological constraints including the type of CDS (i.e., initial, internal, or terminal CDS), the phase of the donor or acceptor splice 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 ORF-FINDER tool developed by NCBI,[55] the Small Exons Finder allows users to search for ORFs that are less than 30bp in size and 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 (https://gander.wustl.edu/%7ewilson/sequence_updater/index.html) tool is primarily used to create a Variant Call Format (VCF) file[56] to correct errors (i.e., base substitutions, insertions, and deletions) in the consensus sequence of a genome assembly. 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 the unique isoforms in a project. For each type of annotation file, the Annotation Files Merger (https://gander.wustl.edu/%7ewilson/submissionhelper/index.php) is used to combine the annotations for all the unique 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.

Third party tools

BEDTools

BEDTools[57] was used to perform genomic arithmetic and compare locations of genomic features in multiple tracks of the GEP UCSC Genome Browser.

NCBI BLAST+

The local sequence alignments produced by the tools in NCBI BLAST+ suite were used for multiple aspects of the protocol including, but not limited to, sequence annotation and local synteny assignment.

Extended data

Figshare. Supplement 1.pdf, DOI: https://doi.org/10.6084/m9.figshare.21235341.[32]

Figshare. Supplement 2.pdf, DOI: https://doi.org/10.6084/m9.figshare.21235345.[33]

Figshare. Supplement 3.docx, DOI: https://doi.org/10.6084/m9.figshare.21235376.[36]

Figshare. Supplement 4.docx, DOI: https://doi.org/10.6084/m9.figshare.21235367.[4]

Figshare. Supplement 5.pdf, DOI: https://doi.org/10.6084/m9.figshare.21235343.[43]

Figshare. Supplement 6.docx, DOI: https://doi.org/10.6084/m9.figshare.21235380.[44]

Figshare. Supplement 7, DOI: https://doi.org/10.6084/m9.figshare.23600556.v1.[58]

Figshare. Supplement 8, DOI: https://doi.org/10.6084/m9.figshare.23600694.[59]

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Acknowledgements

Work done by GEP faculty and their students generated gene models that allowed us to refine our protocol. We would specifically like to acknowledge Dr. Alexa Sawa and her students at College of the Desert for providing input on initial drafts of the protocol and the reconcilers who generated models that can be used for downstream analyses.

References

1. Carson H, Mark Y: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics.* 2011; **12**: 491–491.
PubMed Abstract | Publisher Full Text

2. Hoff KJ, Lomsadze A, Borodovsky M, *et al.*: **Whole-Genome Annotation with BRAKER.** *Methods Mol. Biol.* 2019; **1962**: 65–95.
PubMed Abstract | Publisher Full Text

3. Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *Nar. Genom. Bioinform.* 2021; **3**: lqaa108.
Publisher Full Text

4. Leung W, Rele CP: **Supplement 4.docx.** *figshare.* 2022.
Publisher Full Text

5. Consortium D 12 G, *et al.*: **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature.* 2007; **450**: 203–218.
Publisher Full Text

6. Chen ZX, *et al.*: **Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation.** *Genome Res.* 2014; **24**: 1209–1223.
PubMed Abstract | Publisher Full Text

7. Souvorov A, Kapustin Y, Kiryutin B, *et al.*: **Gnomon – NCBI eukaryotic gene prediction tool.** *NCBI.* 2010.

8. Hoff KJ, Lange S, Lomsadze A, *et al.*: **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.** *Bioinformatics.* 2016; **32**: 767–769.
PubMed Abstract | Publisher Full Text

9. Keilwagen J, Hartung F, Paulini M, *et al.*: **Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi.** *BMC Bioinformatics.* 2018; **19**: 189–189.
PubMed Abstract | Publisher Full Text

10. Shao M, Kingsford C: **Accurate assembly of transcripts through phase-preserving graph decomposition.** *Nat. Biotechnol.* 2017; **35**: 1167–1169.
PubMed Abstract | Publisher Full Text

11. Byrne A, Cole C, Volden R, *et al.*: **Realizing the potential of full-length transcriptome sequencing.** *Philos. Trans. R. Soc. B.* 2019; **374**: 20190097–20190097.
PubMed Abstract | Publisher Full Text

12. Kovaka S, Zimin AV, Pertea GM, *et al.*: **Transcriptome assembly from long-read RNA-seq alignments with StringTie2.** *Genome Biol.* 2019; **20**: 278–278.
PubMed Abstract | Publisher Full Text

13. Mudge JM, Harrow J: **The state of play in higher eukaryote gene annotation.** *Nat. Rev. Genet.* 2016; **17**: 758–772.
PubMed Abstract | Publisher Full Text

14. Tello-Ruiz MK, *et al.*: **Double triage to identify poorly annotated genes in maize: The missing link in community curation.** *PLoS One.* 2019; **14**: e0224086–e0224013.
PubMed Abstract | Publisher Full Text

15. Slawson EE, Shaffer CD, Malone CD, *et al.*: **Comparison of dot chromosome sequences from D. melanogaster and D. virilis reveals an enrichment of DNA transposon sequences in heterochromatic domains.** *Genome Biol.* 2006; **7**(2): R15.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Lopatto D, Alvarez C, Barnard D, *et al.*: **Undergraduate research. Genomics Education Partnership.** *Science.* 2008 Oct 31; **322**(5902): 684–685.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Shaffer CD, Alvarez C, Bailey C, *et al.*: **The genomics education partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions.** *CBE Life Sci. Educ.* 2010 Spring; **9**(1): 55–69.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Leung W, *et al.*: **Evolution of a distinct genomic domain in *Drosophila*: comparative analysis of the dot chromosome in *Drosophila melanogaster* and *Drosophila virilis*.** *Genetics.* 2010; **185**: 1519–1534.
PubMed Abstract | Publisher Full Text

19. Shaffer CD, *et al.*: **A course-based research experience: how benefits change with increased investment in instructional time.** *CBE Life Sci. Educ.* 2014; **13**: 111–130.
PubMed Abstract | Publisher Full Text

20. Lopatto D, Hauser C, Jones CJ, *et al.*: **A central support system can facilitate implementation and sustainability of a Classroom-based Undergraduate Research Experience (CURE) in Genomics.** *CBE Life Sci. Educ.* 2014 Winter; **13**(4): 711–723.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Leung W, Shaffer CD, Reed LK, *et al.*: **Drosophila muller f elements maintain a distinct set of genomic properties over 40 million years of evolution.** *G3 (Bethesda).* 2015 Mar 4; **5**(5): 719–740.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Weisstein AE, Gracheva E, Goodwin Z, *et al.*: **A Hands-on Introduction to Hidden Markov Models.** *CourseSource.* 2016.
Publisher Full Text

23. Elgin SCR, Hauser C, Holzen TM, *et al.*: **Genomics Education Partnership. The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates.** *Trends Genet.* 2017 Feb; **33**(2): 81–85.
PubMed Abstract | Publisher Full Text

24. Laakso MM, Paliulis LV, Croonquist P, *et al.*: **An undergraduate bioinformatics curriculum that teaches eukaryotic gene structure.** *CourseSource.* 2017.
Publisher Full Text

25. Leung W, Shaffer CD, Chen EJ, *et al.*: **Retrotransposons Are the Major Contributors to the Expansion of the Drosophila ananassae Muller F Element.** *G3 (Bethesda).* 2017 Aug 7; **7**(8): 2439–2460.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Lopatto D, Rosenwald AG, DiAngelo JR, *et al.*: **Facilitating Growth through Frustration: Using Genomics Research in a Course-Based Undergraduate Research Experience.** *J. Microbiol. Biol. Educ.* 2020 Feb 28; **21**(1):21.1.6.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Lopatto D, Rosenwald AG, Burgess RC, *et al.*: **Student Attitudes Contribute to the Effectiveness of a Genomics CURE.** *J. Microbiol. Biol. Educ.* 2022 May 16; **23**(2): e00208–e00221.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Dunn NA, *et al.*: **Apollo: Democratizing genome annotation.** *PLoS Comput. Biol.* 2019; **15**: e1006790–e1006790.
PubMed Abstract | Publisher Full Text

29. Raciti D, Yook K, Harris TW, *et al.*: **Micropublication: incentivizing community curation and placing unpublished data into the**

public domain. *Database.* 2018; **2018**: bay013.
Publisher Full Text

30. Clark K, Karsch-Mizrachi I, Lipman DJ, *et al.*: **GenBank.** *Nucleic Acids Res.* 2016; **44**: D67–D72.
PubMed Abstract | Publisher Full Text

31. Matthews BB, *et al.*: **Gene Model Annotations for *Drosophila melanogaster*: Impact of High-Throughput Data.** *G3 Genes Genomes Genetics.* 2015; **5**: 1721–1736.
PubMed Abstract | Publisher Full Text

32. Rele CP, Sandlin KM: **Supplement 1.pdf.** *figshare.* 2022.
Publisher Full Text

33. Sandlin KM, Leung W, Reed LK: **Supplement 2.pdf.** *figshare.* 2022.
Publisher Full Text

34. Jun J, Mandoiu II, Nelson CE: **Identification of mammalian orthologs using local synteny.** *BMC Genomics.* 2009; **10**: 630–630.
Publisher Full Text

35. Jahangiri-Tazehkand S, Wong L, Eslahchi C: **OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation.** *Genom Proteom Bioinform.* 2017; **15**: 361–370.
PubMed Abstract | Publisher Full Text

36. Rele CP, Leung W, Sandlin KM, *et al.*: **Supplement 3.docx.** *figshare.* 2022.
Publisher Full Text

37. Gonzalez DL, Giannerini S, Rosa R: **On the origin of degeneracy in the genetic code.** *Interface Focus.* 2019; **9**: 20190038–20190038.
PubMed Abstract | Publisher Full Text

38. States DJ, Gish W, Altschul SF: **Improved sensitivity of nucleic acid database searches using application-specific scoring matrices.** *Methods.* 1991; **3**: 66–70.
Publisher Full Text

39. Tatusov R, Koonin E, Lipman D: **A genomic perspective on protein families.** *Science.* 1997; **278**: 631–637.
Publisher Full Text

40. Leinonen R, Sugawara H, Shumway M: **The sequence read archive.** *Nucleic Acids Res.* 2011; **39**: D19–D21.
PubMed Abstract | Publisher Full Text

41. Daehwan K, Ben L, Steven LS: **HISAT: a fast spliced aligner with low memory requirements.** *Nat. Methods.* 2015; **12**: 357–360.
PubMed Abstract | Publisher Full Text

42. Feng Y-Y, *et al.*: **RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer.** *Biorxiv.* 2018; **436634**.
Publisher Full Text

43. Wong J, Rele CP, Sandlin KM: **Supplement 5.pdf.** *figshare.* 2022.
Publisher Full Text

44. Rele CP, Reed LK: **Supplement 6.docx.** *figshare.* 2022.
Publisher Full Text

45. Huang W, Li L, Myers JR, *et al.*: **ART: A next-generation sequencing read simulator.** *Bioinformatics.* 2012; **28**: 593–594.
PubMed Abstract | Publisher Full Text

46. Yang C, Chu J, Warren RL, *et al.*: **NanoSim: Nanopore sequence read simulator based on statistical characterization.** *Gigascience.* 2017; **6**: 1–6.
Publisher Full Text

47. Mikheenko A, Prjibelski A, Saveliev V, *et al.*: **Versatile genome assembly evaluation with QUAST-LG.** *Bioinformatics.* 2018; **34**: i142–i150.
PubMed Abstract | Publisher Full Text

48. Miller DE, Staber C, Zeitlinger J, *et al.*: **Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing.** *G3 Genes Genomes Genetics.* 2018; **8**: 3131–3141.
PubMed Abstract | Publisher Full Text

49. Alvarez-Ponce D, Aguade M, Rozas J: **Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes.** *Genome Res.* 2009; **19**: 234–242.
PubMed Abstract | Publisher Full Text

50. Sheth N, Roca X, Hastings ML, *et al.*: **Comprehensive splice-site analysis using comparative genomics.** *Nucleic Acids Res.* 2006; **34**: 3955–3967.
PubMed Abstract | Publisher Full Text

51. Parada GE, Munita R, Cerda CA, *et al.*: **A comprehensive survey of non-canonical splice sites in the human transcriptome.** *Nucleic Acids Res.* 2014; **42**: 10564–10578.
PubMed Abstract | Publisher Full Text

52. Sibley CR, Blazquez L, Ule J: **Lessons from non-canonical splicing.** *Nat. Rev. Genet.* 2016; **17**: 407–421.
PubMed Abstract | Publisher Full Text

53. Lose B, Myers A, Fondse S, *et al.*: **Drosophila yakuba – Tsc1.** *MicroPubl. Biol.* 2021; **2021**.
PubMed Abstract | Publisher Full Text

54. Slawson EE, *et al.*: **Comparison of dot chromosome sequences from *D. melanogaster* and *D. virilis* reveals an enrichment of DNA transposon sequences in heterochromatic domains.** *Genome Biol.* 2006; **7**: R15–R15.
PubMed Abstract | Publisher Full Text

55. Rombel IT, Sykes KF, Rayner S, *et al.*: **ORF-FINDER: a vector for high-throughput gene identification.** *Gene.* 2002; **282**: 33–41.
PubMed Abstract | Publisher Full Text

56. Danecek P, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**: 2156–2158.
PubMed Abstract | Publisher Full Text

57. Quinlan AR, Hall IM: **BEDTools: A flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**: 841–842.
PubMed Abstract | Publisher Full Text

58. Rele CP, Sandlin KM: Supplement 7. [Dataset]. *figshare.* 2023.
Publisher Full Text

59. Rele CP, Sandlin KM: **Supplement 8.** *figshare.* Online resource. 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓

Version 2

Reviewer Report 07 September 2023

https://doi.org/10.5256/f1000research.147335.r192290

© **2023 Murphy T.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

?

Terence D. Murphy
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Thank you for the revisions to the manuscript. I particularly appreciate the data on curation error rates (figure 13), and look forward to the TrackHub being available in the TrackHub Registry.

I have three revisions, the first of which is significant:

1) Figure 1 is still inaccurate and does need to be fixed. The authors make the assertion that "The number of predicted **genes** can show large variations across algorithms", stated in the figure title, legend, Y axis, and text. The legend goes on to explain that some algorithms "can predict multiple transcripts per genomic region (e.g., genBlastG, Spaln)". Having multiple transcripts/isoforms per gene should only be interpreted as inflating the protein-coding **gene** count if the algorithm does not provide information to group them into genes. And in fact the authors are applying such grouping for D. melanogaster by showing a baseline of ~14k genes, as opposed to the ~31k isoforms annotated by FlyBase. NCBI RefSeq provides information to group isoforms into genes, with 14181+/-1083 protein-coding genes reported in current Drosophila annotations. See: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=7215&annotated_only=true&refseq_annotation=true And use "Select columns" to add the protein-coding gene count.

While this isn't a critical point of the paper, the figure implies many of these algorithms are predicting twice as many protein-coding **genes** as found in D. melanogaster, which isn't true in most cases.

a) The best fix is to correct the figure to show true counts of protein-coding genes, aggregating multiple isoforms where that information is provided by the algorithm. Then it's correct to compare to the 14k protein-coding genes annotated in D. melanogaster

b) Alternatively, the authors can revise the main text, figure legend title and text, Y axis, and legend in the figure itself to state everything is plotted for isoforms, and revise the D. melanogaster baseline to be 30799 (for Release 6.46). That alternative isn't ideal, but would be less misleading.

2) if the location will be reasonably permanent, please provide a URL for the TrackHub in the paper so that they can be accessed while the TrackHub Registry is pending.

3) the "Missing/Mislabeled Isoforms" heading is formatted in a different style

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Expertise in manual and automated gene annotation techniques and supporting datasets

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 24 August 2023

https://doi.org/10.5256/f1000research.147335.r192289

© 2023 Crosby M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Madeline Crosby
Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

Accepted as revised.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Relevant area of expertise: gene model annotation in Drosophila melanogaster.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 16 February 2023

https://doi.org/10.5256/f1000research.139289.r162520

© 2023 Crosby M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Madeline Crosby
Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

This paper is a detailed description of a protocol used for undergraduate-based annotation of gene models in various Drosophila species. Undergraduate students benefit from a hands-on experience using computational tools and real genomic data. As primarily a methods paper, it appears to be admirably thorough.

Two minor errors that need to be corrected:

1. Figure 5 legend confuses Akt-PC and Akt-PE.
2. Link for reference 40 goes to a retracted paper.

Additional comments, questions, and suggestions:

The initial project-claiming step includes an estimated difficulty level; this is potentially very useful, but the current assessment appears to be based on the overly simplistic criteria of the number of isoforms, number of coding exons, and evolutionary distance from the reference genome. If this difficulty assessment could be expanded to include atypical phenomena such as non-canonical splicing and non-canonical start codons, it would be much more useful.

Time permitting, it would be beneficial for a student to receive two genes to annotate: one simple case and a second more complicated case.

Were students asked to complete the gene model for the described case of an incomplete genome assembly (split scaffold, Figure 11)? I should think these would fall in the very difficult category and should be screened out.

No ortholog identified by tBLASTn: Can you give an example of successful annotation of a model that was located based on synteny and pieced together from 'additional features' in the syntenic region?

"Correct" naming of protein isoforms is mentioned. The policy appears to be that an isoform should be named with the same suffix as the analogous isoform for the orthologous D. melanogaster gene. This is perhaps helpful, but it is not strictly necessary.

Results using a group of 310 students who submitted gene models are very briefly described (Figure 12); it would be informative if this assessment were expanded. For example:

1. How effective was the strategy to have two independent students annotate a given gene? (Was an incorrect submission usually accompanied by a correct submission?)
2. What percentage of incorrect submissions were in the more difficult category?
3. What percentage of incorrect submissions involved atypical phenomena?

What percentage of students were unable to create a gene model on their own (initially)? What type of assistance was provided to these students?

Has feedback from students precipitated any changes?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Relevant area of expertise: gene model annotation in Drosophila melanogaster.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jun 2023

Chinmay Rele

Thank you for your time and detailed descriptions of the ways you think our methods article could be better improved. We have made most of the changes that you have suggested, and have detailed them below.

- *Two minor errors that need to be corrected:*
    1. *Figure 5 legend confuses Akt-PC and Akt-PE.*
    2. *Link for reference 40 goes to a retracted paper.*

The two minor corrections that have been suggested have been corrected. We have made these corrections pertaining to the Figure 5 legend confusing Akt-PC and Akt-PE , and also updated reference 40 to the correct final publication (note the initial retraction and republication was for a technical issue in the publication process, not a scientific issue) .

- *The initial project-claiming step includes an estimated difficulty level; this is potentially*

very useful, but the current assessment appears to be based on the overly simplistic criteria of the number of isoforms, number of coding exons, and evolutionary distance from the reference genome. If this difficulty assessment could be expanded to include atypical phenomena such as non-canonical splicing and non-canonical start codons, it would be much more useful.

This is a very good suggestion, and we will definitely consider incorporating this kind of information in our claim pipeline in the future. The primary intent for including the difficult score is for faculty to gauge the approximate challenge of the gene to determine if it might be appropriate to their class, and is not a piece of information intended to be shared with the student. For this purpose, we have found that the current definition of the difficulty (including the number of polypeptides, isoforms, and species divergence) seems to be adequate. From the student perspective, the labor required to complete their project generally scales with the number of unique polypeptides and exons. .

- *Time permitting, it would be beneficial for a student to receive two genes to annotate: one simple case and a second more complicated case.*

It would definitely be beneficial to have students annotate (at least) two genes, one simple, and the second more complicated. We recommend to our faculty that they do this if possible, but, this decision ultimately lies with the faculty member who is implementing GEP curriculum within their course, and their pedagogical goals. Further, faculty are welcome to perform their own screening on the various potential projects before claiming them, to determine whether they feel a model is at the right scale for their students.

- *Were students asked to complete the gene model for the described case of an incomplete genome assembly (split scaffold, Figure 11)? I should think these would fall in the very difficult category and should be screened out.*

Instances where the assembly is idiosyncratic are prime examples where manual annotation of models by students is beneficial. Cases this extreme are rare. However, these kinds of cases are what make students most excited and proud of their work when they solve them. We do not recommend that a student's first project be this difficult but once they understand the basics, many students are able to tackle these more complex scenarios. We also provide support to the faculty and students through virtual office hours and a Slack team with the project leaders for the faculty, virtual TAs for the students. We describe and analyze the effectiveness of our pedagogical approaches in other manuscripts either published and or in preparation. We have added some of those citations to the manuscript discussion.

- *No ortholog identified by tBLASTn: Can you give an example of successful annotation of a model that was located based on synteny and pieced together from 'additional features' in the syntenic region?*

Thank you for pointing out that we did not include an example. We apologize for this oversight. We have added an explanation of how to resolve an error like this by citing *Ilp3* in *D. grimshawi*. We have also added a figure to show this

- *"Correct" naming of protein isoforms is mentioned. The policy appears to be that an isoform should be named with the same suffix as the analogous isoform for the orthologous D. melanogaster gene. This is perhaps helpful, but it is not strictly necessary.*

Correct naming of protein isoforms is mentioned in order to maintain consistency across the entire project. Though this is not explicitly necessary, it is beneficial to

assess and predict putative function of each transcript in the target species.

- *Results using a group of 310 students who submitted gene models are very briefly described (Figure 12); it would be informative if this assessment were expanded. For example:*
  - *How effective was the strategy to have two independent students annotate a given gene? (Was an incorrect submission usually accompanied by a correct submission?)*
  - *What percentage of incorrect submissions were in the more difficult category?*
  - *What percentage of incorrect submissions involved atypical phenomena?*

In general, we do see more errors in models classified as being more difficult, but understanding a more discrete error profile is out of the scope of this paper. The primary goal of this manuscript is to provide the description of the standard protocol used by the student annotators part of the GEP to annotate a gene within the target species. Once more data is available, we can make a more detailed assessment of the quality of student models relative to the nature of the gene characteristics.

- *What percentage of students were unable to create a gene model on their own (initially)? What type of assistance was provided to these students?*

We really cannot quantify what percentage of students were unable to create their own model initially for a variety of technical and semantic reasons, primarily since the reconciliation process only sees the summative product of the student work, not the formative steps the students experienced. We do know however that the formative frustration of the challenges of generating a model, followed by ultimate success is essential for driving the positive impact of the research experience (Lopatto et al., 2020; PMID: 32148609). The GEP provides a variety of support for students and their faculty mentors including extensive curriculum resources, video tutorials, virtual teaching assistants, and virtual office hours with project leaders.

- *Has feedback from students precipitated any changes?*

Student feedback is constantly considered as the curriculum and project protocols evolves to keep pace with changes in genomic data resources, with semi-annual updates. Student comments, especially in the pilot stages for new curriculum, are actively solicited by the project leaders. Further, the faculty and virtual TAs observations of how students may struggle with new concepts, inspire new or improved educational resources.

Thank you again for all your input to help us make this manuscript more complete and robust.

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 January 2023

https://doi.org/10.5256/f1000research.139289.r158890

© 2023 Murphy T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

?

Terence D. Murphy
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

This paper describes a protocol for manual annotation of genes involved in insulin signaling pathway (ISP) in Drosophila species using a platform developed by Genome Education Partnership (GEP) as a tool to educate undergraduate students and get them interested in scientific research. The platform also helps generate valuable manually curated annotation in these species where there isn't much evidence data available, by leveraging data from D. melanogaster. The protocol is well conceived to meet the needs of an educational program. The paper is generally sound, but would benefit from a few clarifications and corrections as described below.

1. While this is a 'Methods" paper, the authors should perhaps provide some context in the Introduction about what GEP is and what its broader goals are (citing the latest publication on GEP) for the benefit of the naïve reader before diving into the details of the manual annotation protocol.

2. In figure 1, the Y-axis is labeled as "number of predicted protein-coding genes", but the legend includes the following statement that implies it's actually number of isoforms: "*Prediction differences can partly be attributed to some algorithms predicting a single isoform in a genomic region (e.g., GeneID), while others predict multiple isoforms per genomic region (e.g., genBlastG, Spaln).*" Furthermore, the values for NCBI RefSeq Genes are all in the 20-45k range, when NCBI reports most Drosophila species as having 13-15k protein-coding genes. The authors need to check the data for all annotation programs provided in figure 1 and correct the figure.

3. In table 1, the authors list "UCSC Assembly" with names that largely aren't shown in UCSC and don't correspond to the published assembly names. For example, the D. miranda assembly GCA_000269505.2 is named DroMir_2.2. Where did the name DmirGB2 come from? The authors should clarify the meaning of the first column, and add an additional column with the official name for each assembly since it's difficult to recognize assembly accessions.

4. Many of the assemblies being used by the GEP project have been superseded by newer and higher quality assemblies. Do the authors periodically upgrade to new assemblies, or is this viewed as unnecessary effort for the educational goals of the project?

5. To support the statement *"As of August 2022, GEP students from 79 institutions have used this annotation protocol to construct 2,101 gene models across 27 Drosophila species."*, it would be useful to provide a supplemental table with the list of gene models curated for this project so far. Additional data on conformance of student models to the final consolidated

annotations would also be of interest if readily available.

6. In order to aid access to the final gene models, in addition to the TPA submissions the authors could construct a genome browser track available through the remote Track Hub system. This could be updated periodically (e.g. at the end of each semester or year) with additional gene models, and made available through the Track Hub Registry system. While it would be useful to include the availability of this track in the paper if the authors have the resources to generate it, it's not a requirement for publication.

7. Figure 5: It seems that the isoform names are reversed in the legend. Looking at the 'User-created Annotations' track, Akt-PC is the isoform with the longer coding region and not Akt-PE as stated in the legend.

8. Assembly errors: For cases where the student identifiers an error in a target assembly, how is that information conveyed in the TPA records? Is the TPA sequence corrected via a VCF file? What is done with a case like D. pseudoobscura chico from Figure 11?

9. For the reconciliation process, how would you determine that an isoform is 'mislabeled'? Is it in comparison to D. melanogaster isoform name? In general, what are the rules to name isoforms in target species?

10. "*The annotation effort of the student annotators is of value because the students can outperform computational gene predictors*." - This statement seems a bit misleading unless there is a mention of scale. Computational annotation methods can predict gene and transcript models on a genomic scale, which is not possible by manual annotators (in a reasonable timeframe). It would be reasonable to say that manual annotators can correct errors found in computational predictions and can fill in gaps, as in the example cited in the paper where computational methods could not detect an additional protein isoform due to the use of an upstream non-AUG start codon. If the authors wish to keep the statement, a suggestion is to add something that conveys "at the level of an individual gene or a small set of genes" at the end. It would also be helpful to add some statistics on how often students identify errors or additions to automated gene sets.

11. The report is primarily focused on describing aspects of the annotation protocol. However, the key aspect of this work is its use as an educational tool to engage students. It would be helpful to add some data or at least anecdotal statements to the discussion about the success of the educational goals. For example, do instructors joining the project tend to stay with it in successive years? Is there survey or other data from the students or TAs about whether it's a valuable aid? Adding some text to the discussion, or results if available, would help other instructors decide if this would be a valuable addition to their curricula.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use

by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Expertise in manual and automated gene annotation techniques and supporting datasets

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jun 2023

Chinmay Rele

Thank you for your time and detailed descriptions of the ways you think our methods article could be better improved. We have made most of the changes that you have suggested, and have detailed them below.

- *While this is a 'Methods" paper, the authors should perhaps provide some context in the Introduction about what GEP is and what its broader goals are (citing the latest publication on GEP) for the benefit of the naïve reader before diving into the details of the manual annotation protocol.*You suggested to add some context about what the GEP is, and what its broader goals of the GEP are.

This suggestion is very helpful, and more information about the GEP, its mission statement, and larger scientific publications has been added after the third paragraph of the Introduction.

- *In figure 1, the Y-axis is labeled as "number of predicted protein-coding genes", but the legend includes the following statement that implies it's actually number of isoforms: "Prediction differences can partly be attributed to some algorithms predicting a single isoform in a genomic region (e.g., GeneID), while others predict multiple isoforms per genomic region (e.g., genBlastG, Spaln)." Furthermore, the values for NCBI RefSeq Genes are all in the 20-45k range, when NCBI reports most Drosophila species as having 13-15k protein-coding genes. The authors need to check the data for all annotation programs provided in figure 1 and correct the figure*.

We have made clear that some predictors predict single transcripts or multiple transcripts per genomic region while providing examples of such predictors in the caption of Figure 1.

- *In table 1, the authors list "UCSC Assembly" with names that largely aren't shown in UCSC*

and don't correspond to the published assembly names. For example, the D. miranda assembly GCA_000269505.2 is named DroMir_2.2. Where did the name DmirGB2 come from? The authors should clarify the meaning of the first column, and add an additional column with the official name for each assembly since it's difficult to recognize assembly accessions.

Thank you for this suggestion and we apologize for the confusion. We have added the canonical assembly names from NCBI under the "Assembly Name" column, and the new assembly name within the GEP framework has been clarified in the caption of this table.

- *Many of the assemblies being used by the GEP project have been superseded by newer and higher quality assemblies. Do the authors periodically upgrade to new assemblies, or is this viewed as unnecessary effort for the educational goals of the project?*

We do intend to use newer assemblies as they become available. Our rate of adoption is delayed by the need to synchronize all curricular materials with updated genome assemblies, but we do this as we are able. We have added this at the end of the third paragraph of Introduction.

- *To support the statement "As of August 2022, GEP students from 79 institutions have used this annotation protocol to construct 2,101 gene models across 27 Drosophila species.", it would be useful to provide a supplemental table with the list of gene models curated for this project so far. Additional data on conformance of student models to the final consolidated annotations would also be of interest if readily available.*

We have updated the date, the number of models, and the number of *Drosophila* species. We have also included a heatmap showing the number of models submitted for each species/gene combination as a supplemental file. **The outcome of student gene model conformance to the final model is provided in Figure 12, note that many of the models listed in supplement 7 have not yet been reconciled.**

- *In order to aid access to the final gene models, in addition to the TPA submissions the authors could construct a genome browser track available through the remote Track Hub system. This could be updated periodically (e.g. at the end of each semester or year) with additional gene models, and made available through the Track Hub Registry system. While it would be useful to include the availability of this track in the paper if the authors have the resources to generate it, it's not a requirement for publication.*

Thank you for this suggestion. We have created a supplemental file (Supplement 8) with all the TrackHub links that have been generated as of March 24, 2023. We will be creating a TrackHub registry. We have added the reference to the supplemental file at the end of "Reconciliation Process" at the end of the Methods section.

- *Figure 5: It seems that the isoform names are reversed in the legend. Looking at the 'User-created Annotations' track, Akt-PC is the isoform with the longer coding region and not Akt-PE as stated in the legend.*

Thank you for pointing this error out – it has been corrected to mention that the PC isoform is the isoform with the longer coding region.

- *Assembly errors: For cases where the student identifiers an error in a target assembly, how is that information conveyed in the TPA records? Is the TPA sequence corrected via a VCF file? What is done with a case like D. pseudoobscura chico from Figure 11?*

Assembly errors are documented using a VCF file, which is then submitted to NCBI along with the TPA record for the specific model. This has been added to the end of "Assembly Errors" in the "Exceptions to the standard annotation workflow" section. **If**

the model is in two scaffolds in our assemblies, we opt to complete the reconciliation in a newer (for this instance, we would choose to reconcile the model in the DpseRefSeq1 (GCF_009870125.1)) assembly because our internal tools cannot adequately validate models that are present in two scaffolds.

- *For the reconciliation process, how would you determine that an isoform is 'mislabeled'? Is it in comparison to D. melanogaster isoform name? In general, what are the rules to name isoforms in target species?*

Good question. An explanation of how an isoform is defined/named has been added as a new section called Missing/Mislabeled Isoforms at the end of the Most Common Annotation Errors section.

- *(a)"The annotation effort of the student annotators is of value because the students can outperform computational gene predictors." - This statement seems a bit misleading unless there is a mention of scale. Computational annotation methods can predict gene and transcript models on a genomic scale, which is not possible by manual annotators (in a reasonable timeframe). It would be reasonable to say that manual annotators can correct errors found in computational predictions and can fill in gaps, as in the example cited in the paper where computational methods could not detect an additional protein isoform due to the use of an upstream non-AUG start codon. If the authors wish to keep the statement, a suggestion is to add something that conveys "at the level of an individual gene or a small set of genes" at the end.*

Thank you for pointing out this important distinction and lack of elaboration. We have now added the "individual gene level" qualifier to this at the end of the requested sentence. **We have also pointed out the computational gene prediction algorithms are very valuable at the genome scale.**

- *(b) It would also be helpful to add some statistics on how often students identify errors or additions to automated gene sets.*

This is an interesting question, and we hope to report on it in a future study, once we have enough data to be statistically robust.

- *The report is primarily focused on describing aspects of the annotation protocol. However, the key aspect of this work is its use as an educational tool to engage students. It would be helpful to add some data or at least anecdotal statements to the discussion about the success of the educational goals. For example, do instructors joining the project tend to stay with it in successive years? Is there survey or other data from the students or TAs about whether it's a valuable aid? Adding some text to the discussion, or results if available, would help other instructors decide if this would be a valuable addition to their curricula*.****

Studies on student learning outcomes showing the benefits to students have been performed by the GEP in separate publications, and a brief description of these findings have been added as the last paragraph of Discussion.

Thank you again for all your input to help us make this manuscript more complete and robust.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research