Using Genome Annotation Projects to Teach Eukaryotic Gene Structure and to Engage Students in Genomics Research

Maria S. Santisteban¹, James Godde², Anya L. Goodman³, Charles R. Hauser⁴, Don Paetkau⁵, Catherine Reinke⁶, Wilson Leung⁷, Cindy Arrigo⁸, Nathan T. Mortimer⁹, and Laura K. Reed¹⁰

¹ University of North Carolina at Pembroke, Department of Biology, PO Box 1510, Pembroke NC 28352 USA
² Monmouth College, 700 E Broadway, Monmouth, IL 61462
³ California Polytechnic State University, Chemistry and Biochemistry Dept., San Luis Obispo CA 93407 USA
⁴ St. Edward’s University, School of Natural Science, 3001 South Congress, Austin, TX 78704-6489 USA
⁵ Saint Mary’s College, Notre Dame, IN 46556
⁶ Linfield College, Biology Department, 900 SE Baker Street, McMinnville OR 97128 USA
⁷ Washington University in St. Louis, Department of Biology, Campus Box 1137, One Brookings Drive, St. Louis MO 63130-4899 USA
⁸ New Jersey City University, Department of Biology, 2039 Kennedy Blvd., Jersey City, NJ 07305
⁹ Illinois State University, School of Biological Sciences, Box 4120, Normal IL 61790-4120 USA
¹⁰ University of Alabama Tuscaloosa, Department of Biological Science, Box 870344, Tuscaloosa, AL 35487-0344 USA

(maria.santisteban@uncp.edu; jgodde@monm.edu; agoodman@calpoly.edu; charlesh@stedwards.edu; dpaetkau@ saintmarys.edu; creinke@linfield.edu; wleung@wustl.edu; carrigo@njcu.edu; ntmorti@ilstu.edu; lreed1@ua.edu)
The Genomics Education Partnership (GEP; https://thegep.org) began as a consortium of 16 faculty in 2006 with a goal of providing students with Course-based Undergraduate Research Experiences (CUREs) in genomics. Today, GEP has over 200 faculty from more than 180 institutions and engages more than 3,900 undergraduates in authentic genomics research annually. These faculty joined and continued to participate in the GEP for many reasons, including the collaborative nature of the research, the well-established infrastructure, and the supportive network of like-minded colleagues. Faculty implement GEP materials in diverse settings — ranging from short modules (2–8 weeks) within a course, to a standalone full-semester course, to independent student research. GEP students show significant gains in scientific knowledge and attitudes toward science. In addition to improving their understanding of the research process and how new knowledge is created in the field, GEP students acquire desirable and transferable skills essential for future participation in the workforce, such as problem solving, independence, application of knowledge, team-work, and collaboration. Students also gain competence in the use of computational algorithms to analyze large biological datasets — thereby preparing students for a growing need of a workforce trained at applying statistics and computational tools to analyze large datasets. In addition, GEP students and their faculty mentors are eligible to be co-authors on the scientific publications that are based on their work. In this workshop, we will provide an overview of the GEP community, a hands-on guided tour of our introductory curriculum aimed to teach gene structure, transcription, translation, and processing, and a step-by-step walkthrough that illustrates the protocol for annotating a protein-coding gene in *Drosophila*. Participants will receive information on how to join the GEP community and receive training and resources to enable their implementations.

Keywords: Genomics Education Partnership (GEP), bioinformatics, undergraduate research, eukaryotic gene structure, gene annotation, genome browser, mRNA processing, transcription, translation

Introduction

**GEP’s mission and trajectory**

The Genomics Education Partnership (GEP; https://thegep.org; Elgin et al. 2017) began as a consortium of 16 faculty in 2006 with a goal of providing students with Course-based Undergraduate Research Experiences (CUREs) in genomics. This CURE is organized around teaching students how to analyze data to annotate genes and genomic regions to support comparative genomics investigations. Today, GEP has over 200 faculty and engages more than 3,900 undergraduates in authentic genomics research annually (Figure 1a).

Genomics has several advantages compared to wet lab research that makes it ideal for engaging large numbers of undergraduate students in research during the academic year. Bringing genomics research into the classroom is cost effective because only computers and access to the Internet are required. Large genomics datasets are publicly available, and students can use freely available bioinformatics tools and databases (e.g., genome browsers, NCBI BLAST) to analyze these datasets. Due to the lack of safety issues, restricting access to laboratory installations is not a concern. Students can quickly acquire the skills needed to help each other, which results in peer teaching and enables faculty to support larger numbers of students. Because mistakes are relatively inexpensive, students can iteratively attempt different approaches to solve a challenging research problem, thereby becoming more familiar with the research process and gaining deeper insights into the research problem. Genomics also introduces students to the use of computational algorithms to analyze large biomedical datasets — preparing students for a growing need of a workforce trained at applying statistics (e.g., machine learning) and computational tools to analyze large datasets. Many students present their work in local, regional, or national conferences. Students and their faculty mentors are eligible to be co-authors on the scientific publications that are based on their work.

Faculty joined and continued to participate in the GEP for many reasons, including the collaborative nature of the research, the well-established infrastructure, and the supportive network of like-minded colleagues (Shaffer et al. 2010; Lopatto et al. 2014). Developing
and maintaining quality CUREs that challenge and engage students is demanding and requires a great deal of dedication and expertise. The GEP community meets these challenges by hosting online and in-person training workshops for new members to prepare them to teach CUREs, by leveraging resources among participating institutions not available outside R1 institutions (i.e. McDonnell Genome Institute at Washington University), and by providing up-to-date curriculum materials as well as technical and pedagogical support during the academic year. GEP also hosts an annual Alumni Workshop where GEP faculty can come together and share best practices, collaboratively develop new curricula and tools, and work on grants and papers, etc.

In the last five years, the GEP organization has transitioned from a centralized model to distributed leadership. From 2006–2017, the GEP was led by Professor Sarah C.R. Elgin at Washington University in St. Louis. The central organization hosted the new Faculty/TA Training Workshops and Faculty Alumni Workshops, maintained the GEP curriculum materials, provided technical and teaching support, managed the assessments and evaluations of GEP faculty and students, as well as the creation, claiming, submission, and reconciliation of student projects. The central organization also managed the data analysis and drafting of the scientific (e.g., Leung et al. 2017) and education research (e.g., Lopatto et al. 2014; Shaffer et al. 2014; Lopatto et al. 2020) manuscripts.

Under the new distributed leadership model, Professor Laura K. Reed at The University of Alabama serves as the GEP director, but the different responsibilities of the GEP are distributed among multiple committees led by other GEP faculty members (https://thegep.org/staffandleadership/). The current GEP committees include: Steering; Assessment; Curriculum; Diversity, Equity, and Inclusion; Professional Development and Mentoring; and Science/IT. Each GEP faculty is affiliated with at least one committee, and the leadership for each committee (e.g., chair, vice chair, members) rotates every 2–3 years.

The community is also exploring and implementing new ways to recruit and train new members. During the past two years, GEP faculty and staff have organized online training workshops with experienced GEP faculty serving as trainers and mentors for new members (https://thegep.org/new-member-trainings/). Most of the new member training workshops were held during the summer — GEP faculty organized six online training workshops in Summer 2020 and three online training workshops in Summer 2021 for new members. During the 2020–2021 academic year, GEP faculty organized three online training workshops and one Regional Node Training workshop to train new members. These online training opportunities enabled faculty to incorporate the web-based GEP curriculum into existing courses, and allowed students to engage in research despite the lack of access to the physical lab environment during the COVID-19 pandemic. The pandemic, despite forcing new member training to be conducted online, rather than the traditional in-person workshops, actually led to a significant increase in new GEP members, as well as interest in teaching using modules that could easily be adapted to an online format. Faculty that had previously been hesitant about using in-silico labs were suddenly interested in using the materials that the GEP had already developed.

One outcome of our virtual training and the rapid growth of our community has been a discovery of the need to connect individual faculty members with each other to offer easy access to experienced colleagues. Thus, volunteer mentors have been matched with newly trained faculty based on geography and if possible — based on similar instructional implementation. In addition to the virtual community, we have a network of regional nodes (Figure 1b) that during non-pandemic times hold in-person activities, such as student research symposiums, faculty meetings, and TA trainings. During pandemic times these nodes have held virtual trainings, virtual faculty happy hours, and virtual research symposiums highlighting student work as a way to keep faculty and students connected and supported while the organization grows.

**Collaborative Science Projects**

GEP has diversified the science projects pursued by the community. In the past, the GEP research project has focused on investigating the properties and evolution of the Muller F element (also known as the dot chromosome) in multiple *Drosophila* species. This domain is unusual in the *Drosophila* genome because it exhibits both heterochromatic and euchromatic characteristics. The distal 1.3 Mb of the *D. melanogaster* F element is primarily heterochromatic by many criteria (e.g., high repeat
density, low rates of recombination, enriched in the histone modifications H3K9me2/3 and the chromosomal protein HP1a). While heterochromatin is generally associated with gene silencing, the F element also contains ~80 protein-coding genes that exhibit a similar range of expression compared to genes that reside in the euchromatic portions of the genome. In order to use comparative genomics to characterize the evolution of genes and repeats in this unusual domain, GEP students participated in improving the genome sequence and annotating the small, mostly heterochromatic F element and a comparable portion of the euchromatic D element in different Drosophila species. Their work contributed to three publications on the evolution of the F element, with all contributing students as co-authors (Leung et al. 2010; Leung et al. 2015; Leung et al. 2017).

The most recent F element research project engages GEP faculty and students in improving the assembly and annotating F elements from a group of species that shared a common ancestor with D. melanogaster about 10–15 million years ago. These species are at the ideal evolutionary distances from D. melanogaster for the identification of conserved regulatory motifs via comparative genomics. Each student (or team of 2–3 students) took on the challenge of finishing ~100 kb and/or annotating a ~40 kb project from one of these species, either from the F element or a comparison region on the D element. In addition to the coding regions, GEP students also annotated the promoter regions of these genes to decipher the characteristics of the transcription start sites. Ongoing student annotation reconciliation and collaboration with Dr. Jeremy Buhler from the Washington University in St. Louis Computer Science & Engineering Department will utilize the Magma program (Ihuegbu et al. 2012) to identify conserved regulatory motifs via phylogenetic footprinting. The analysis could tell us much more about this interesting chromosome and the genes that reside there, potentially giving additional insights into the regulatory factors that enable F element genes to be expressed in a heterochromatic environment.

Three new scientific projects have been incorporated to the GEP community: the expanded F element project, the parasitoid wasp venom gene annotation project, and the Drosophila pathways project. These new opportunities expand both the range of scientific fields in which faculty and students may find particular interests, as well as the possibilities of pairing annotation projects with wet lab investigations.

### The Expanded F Element Project

The expanded F element project aims to identify factors that contribute to changes in chromosome size and to assess the impact of chromosome size on gene characteristics. The assembled portion of the D. ananassae F element (~19.1 Mb) is substantially larger than the D. melanogaster F element (~1.3 Mb). Previous analysis of a 1.4 Mb region of the D. ananassae F element suggests this expansion can primarily be attributed to the increase in transposon density (Leung et al. 2017). Preliminary analyses of the Drosophila genomes sequenced by the modENCODE project indicates that the F elements in D. bipectinata, D. kikkawai, and D. takahashii are also larger than the D. melanogaster F element. However, the extent of the F element expansion in these three species are unclear due to gaps and misassemblies in the published genome assemblies.

The advent of sequencing technologies that produce long reads (e.g., Pacific Biosciences, Nanopore) provide the opportunity to improve the quality of the F element assemblies for these four Drosophila species. The GEP has constructed new genome assemblies for D. bipectinata, D. kikkawai, and D. takahashii based on sequencing data generated by the Pacific Biosciences, Nanopore, and illumina sequencers. GEP students will analyze the improved genome assemblies and annotate the coding regions on the expanded F elements as well as genes on euchromatic reference regions near the base of the Muller D elements. In collaboration with Dr. Thomas Gingeras at Cold Spring Harbor Laboratory, the GEP has generated RAMPAGE (Batut et al. 2013) data for these four species. GEP students can utilize the RAMPAGE data to annotate the transcription start sites and to characterize the promoter architecture of F element genes in the four species. The gene annotations for these four Drosophila species will be used in a comparative analysis to identify the major contributors to the F element expansion and to assess the impact of the expansion on gene characteristics (e.g., promoter shape, gene size, codon bias, etc.). The comparative analysis might also identify factors that enable genes to function in a heterochromatic region with high repeat density. This project is led by Dr. Cindy J. Arrigo at New Jersey City University.
The Parasitoid Wasp Venom Gene Annotation Project

The parasitoid wasp venom gene annotation project aims to provide further insights into the function and evolution of parasitoid venoms. Current research has hinted at putative mechanisms of action, and it is expected that improved annotation of venom genes will help to characterize the putative functions of venom proteins and to identify proteins with potential dominant negative functions. Similarly, understanding the genomic context of venom genes should help resolve questions concerning the evolution of venom genes. Initially, GEP students will focus their annotation efforts on the *Ganaspis* sp. 1 venom-encoding genes, and then expand to *Leptopilina boulardi* and *Leptopilina heterotoma*. These gene annotations will utilize *Nasonia vitripennis* and *D. melanogaster* as the informant genomes, and leverages experimental data from RNA-Seq and protein mass spectrometry to facilitate the annotation of novel genes. The gene models produced by GEP students will allow us to explore the conservation of evolutionary and functional mechanisms across parasitoid species. This project is led by Dr. Nathan T. Mortimer at Illinois State University.

The Drosophila Pathways Project

The *Drosophila* pathways project uses network analysis approaches to elucidate the evolution and function of biological pathways, initially focusing on the insulin signaling pathway. Past studies have shown that the rates of evolution for the gene’s protein-coding regions are correlated with its position within the network, the number of physical interactions, expression levels, and the existence of closely-related paralogs. GEP students will produce gene annotations for the putative orthologs and paralogs of *D. melanogaster* genes involved in the insulin signaling pathway in 27 *Drosophila* species. These improved gene models will be used to ascertain if the rates of evolution for the genes’ regulatory regions parallels that of their protein-coding regions. GEP students will also draft microPublication (μPub) articles describing the annotation of each gene (e.g., Lose et al. 2021). This project is led by Dr. Laura K. Reed at The University of Alabama.

<table>
<thead>
<tr>
<th>Research Goals →</th>
<th>Research Goal 1: How did the Muller F element of <em>Drosophila</em> evolve, and how does its evolution impact the function of the resident genes?</th>
<th>Research Goal 2: In the context of a pathway, how have the genes and the regulation of those genes evolved across species (e.g., in <em>Drosophila</em>, parasitoid wasps)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Goals ↓</td>
<td>Students will integrate genomic data from other <em>Drosophila</em> species with experimental data for the target species (e.g., RNA-Seq) to develop a hypothesis for the gene model in their assigned species.</td>
<td>Students will integrate genomic data from other <em>Drosophila</em> and wasp species with experimental data (e.g., RNA-Seq, protein mass spectrometry) to develop a hypothesis for the gene model in their assigned species.</td>
</tr>
<tr>
<td>Student Goal 1: Describe and conduct comparative genomics analysis</td>
<td>Students will use tools to verify that the proposed gene model satisfies basic biological constraints.</td>
<td>Students will use tools to verify that the proposed gene model satisfies basic biological constraints.</td>
</tr>
<tr>
<td>Student Goal 2: Correctly build a eukaryotic gene model</td>
<td>Students prepare a report to defend their gene model given the available evidence. This report along with their model are used in the compilation and reconciliation of the full data set for the comparative analysis of the Muller F element.</td>
<td>Students prepare a report to defend their gene model given the available evidence. This report along with their model are used in the compilation and reconciliation of the full data set for the network analysis.</td>
</tr>
<tr>
<td>Student Goal 3: Participate productively in the science research process</td>
<td>Table 1. The table shows how tasks align student and research goals for all three science areas. Adapted from CUREnet (<a href="https://serc.carleton.edu/curenet/collection/215335.html">https://serc.carleton.edu/curenet/collection/215335.html</a>)</td>
<td></td>
</tr>
</tbody>
</table>
Models for Implementation

One of the factors which makes the resources provided by the GEP a valuable and attractive instrument for teaching and research is the potential for flexibility and diversity in the implementation. GEP faculty are from diverse institutions that vary in size, selectivity, and student characteristics. There are currently about 214 active GEP faculty members representing 2-year colleges, 4-year colleges with a primary focus in teaching, institutions offering masters and PhDs, and Minority-serving institutions, including Historically Black Colleges and Universities (HBCUs) and tribal colleges, etc. (Figure 1; https://thegep.org/directory/). The commitment of the GEP to broadening participation has provided access to research for students who otherwise would not be able to access mentored research. GEP has made it a priority to make these students authors in publications. Efforts toward greater inclusivity have also challenged us to improve the curriculum so that they conform to the ADA Standards for Accessible Design (https://www.ada.gov/2010ADASTANDARDS_INDEX.HTM); we also currently provide some of our introductory curriculum in Spanish.

Figure 1. Map of current GEP members (a) and regional nodes (b)

Faculty have different pedagogical goals which result in differences in implementation. The courses that utilize GEP materials range from Introductory Biology to Genetics, Biotechnology, Molecular Biology, Cell Biology, Genomics and Bioinformatics courses. For some faculty, providing students with a research experience is an essential goal, whereas others are looking to teach students the concepts of genomics, bioinformatics, and/or genetics with a hands-on approach that uses available genomics data organized in a genome browser. Some faculty implement GEP activities as a module within a course, devoting between 2 to 8 weeks to the GEP materials. Some faculty implement GEP as a standalone laboratory course for the entire semester, while others use it for independent research experiences. We have found that students at all institution types receive the same significant benefit from their participation in GEP. The one factor that scales with the degree of benefit is the amount of class time devoted to GEP activities — more instructional time invested in the project leads to better outcomes (Shaffer et al. 2014).

Workflow for GEP Scientific Projects

The GEP uses a “divide and conquer” strategy to facilitate student contributions to the GEP scientific projects (Figure 2). The lead scientists for each scientific project define the list of genomic regions or a collection of genes that require manual curation based on their scientific objectives. Visualization of the publicly available genome assemblies in conjunction with computational (e.g., protein sequence alignments, gene predictions) and experimental data (e.g., RNA-Seq, RAMPAGE) within
a single workspace can assist the lead scientists in identifying the regions of interests and in generating the student projects. To enable faculty with limited technical expertise to construct genome browsers, the GEP has partnered with the Galaxy Project (https://galaxyproject.org) to create G-OnRamp (https://g-onramp.org), a web-based platform for constructing UCSC Assembly Hubs and JBrowse/Apollo genome browsers for eukaryotic genomes (Liu et al. 2019). G-OnRamp was used to construct genome browsers for four wasp species as part of the GEP parasitoid wasp venom gene annotation project (Sargent et al. 2020).

GEP faculty claim the annotation projects, and each annotation project is completed by at least two students working independently at two different institutions. The completed annotation projects are submitted to the lead scientist and then reconciled by experienced students to create the reconciled gene models set for downstream analyses. The analysis results are reported in scientific publications and the associated datasets are submitted to public databases. GEP students who complete an annotation project and their faculty mentors are eligible to be co-authors on the resulting publication, provided that they read, critique, and approve the manuscript. GEP science publications have many student co-authors. For example, the manuscript analyzing the evolution of the F element in four Drosophila species has 1014 co-authors; 940 of whom participated as students (Leung et al. 2015). For the Drosophila pathways project, GEP students and faculty also create and submit articles describing the annotation of each gene model to the microPublication Biology journal (Rele et al. 2021; Lose et al. 2021).

Figure 2. Overview of the annotation workflow for GEP scientific research projects
Curriculum to Prepare Students for Genome Annotation

Many faculty find that it is beneficial for students to review the organization of eukaryotic genes/genomes prior to engaging in the GEP research projects. GEP faculty members have developed and published six "Understanding Eukaryotic Genes" (UEG) modules that provide a review of eukaryotic gene structure, transcription, and translation (Laakso et al. 2017; https://thegep.org/ueg/) as well as four Transcription Start Sites modules that provide an overview of the strategies for annotating the 5’UTRs of Drosophila genes (https://thegep.org/tss/). In addition to these introductory materials, GEP faculty have also developed presentations, step-by-step walkthroughs, workflows, and exercises for the GEP research projects. These resources are freely available for faculty and student use through the “Curriculum” search page of the GEP website (https://thegep.org/curriculumsearch).

For this workshop, we will introduce a subset of the UEG modules and complete a walkthrough that illustrates the protocol for annotating a Drosophila gene.

Understanding Eukaryotic Genes

The six UEG modules were created to help students understand eukaryotic gene structure and functionality via an active learning approach instead of traditional lectures. These modules leverage the features provided by the genome browser to enable students to visualize and explore eukaryotic genes/genomes at multiple scales (from a single nucleotide to a genomic region with multiple isoforms and genes).

These six modules now, also available in Spanish, are:

1. Introduction to the Genome Browser: What is a gene?
2. Transcription, Part I: From DNA sequence to transcription unit
3. Transcription, Part II: What happens to the initial transcript made by RNA pol II?
4. Removal of introns from pre-mRNA by splicing
5. Translation: The need for an Open Reading Frame
6. Alternative splicing

The UEG modules can be used in introductory classes by themselves or can be used as a preamble to the annotation process. If time is a concern, instructors can assign the modules as an outside-of-class independent project, in which students work through each Module independently, watching the videos that accompany each Module on the GEP YouTube channel (https://thegep.org/videos), and answering the questions in each Module for submission. Most students can complete the UEG modules in eight to ten 1-hour class periods.

Instructors can use multiple active learning approaches with these modules and deploy them in many different contexts. All or a subset of UEG modules can be used by an instructor, as each module is a stand-alone exercise. In most of our implementations, students work in groups of two or three using their own computers or computers in a computer lab. The instructor gives a brief introduction to the Module, and then students discuss each part of the lesson within their small groups while they explore the genome browser. We also often use large group discussions or peer instruction to address student questions and challenging concepts. Typically, students work together to complete each Module and produce a gene model by the end of the Module 6. Students submit their answers to the questions associated with each Module to their instructor. As students discuss the exercises within each Module, they gather evidence, evaluate potential answers to each question, and occasionally resolve contradictions in the evidence. Using a group format has some risks of uneven participation, but generally stimulates useful dialogue.
Annotation Walkthrough

To prepare students to participate in one of science research projects, instructors usually start with a presentation which provides the scientific background of the research project, and describe how the data they are analyzing was generated. Instructors then typically ask their students to complete a step-by-step annotation walkthrough which illustrates the annotation protocol using one or more example genes. For this workshop, the student handouts include a modified version of the annotation walkthrough for the *Drosophila* F element research project (the original version is available through the F Element Project page on the GEP website; https://thegep.org/felement/). The walkthrough for the *Drosophila* pathways project is available on the Pathways Project page on the GEP website (https://thegep.org/pathways/). (The walkthrough for the parasitoid wasp venom gene annotation project is currently under development.)

Once the students have completed the annotation training, instructors can claim annotation projects and assign them to students either independently or in groups of two or three. When the students have completed their project, the instructor submits the project back to the GEP where it will be combined with other student projects to produce the datasets for the downstream analyses and eventual publication. Most independent annotation projects take approximately twelve hours of class time.

Assessment of Participating Students and Faculty

To enable the GEP to learn and continue to optimize the student experience, participating students are encouraged to take a pre-course quiz and survey before starting to work with GEP materials (available through the Assessment page on the GEP website; https://thegep.org/assessment/). At the end of the course, students will take the post-course quiz and survey. To mitigate test sensitization, there are two versions of the knowledge quiz that cover the same concepts. Students are randomly assigned to one version of the quiz for the pre-course assessment, and assigned to the other version of the quiz for the post-course assessment. A separate set of quizzes for the "Understanding Eukaryotic Genes" curriculum has been developed that aims to assess students' knowledge gains.

Student/faculty gains and testimonials

Besides the gain in knowledge about eukaryotic genes and genomes captured in the quizzes, faculty have also observed many positive impacts for their students in the areas of problem solving, independence, application of knowledge, involvement in peer-to-peer teaching, team-work and collaboration, and developing a sense of ownership (Shaffer et al. 2010). The GEP approach facilitates an introduction to the process of research and helps to develop an understanding of how new knowledge is created in the field. Indeed, students in GEP-associated courses gained many of the skills and attitudes normally associated with a wet lab summer undergraduate research experience (Lopatto et al. 2008).

The two tables below (adapted from Lopatto et al. in preparation) list sample comments from students that reflect their perception of "research emphasis practices" utilized by their faculty.
Table 2. Comments from GEP student focus groups that reflect strategies to emphasize active research participation within the classroom.

<table>
<thead>
<tr>
<th>Topic from Faculty Log</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better understood genes, annotations, genomics</td>
<td>I feel like the goals were to understand a lot of the like nitty-gritty ways that genes work like start [codons] and stop [codons] and specific things to look for in like a sequence that you know is or possibly could be a gene. But also to understand on like a larger scale how to use the program to do that for you and how to use, like we had like four or five programs that we used on a regular basis to like check with each other and how to know when you needed more information or when you had the information you needed because going into it you don’t know.</td>
<td>We also learned a lot of tricks and gimmicks as to how to annotate a gene or just how to look at a genome because the first one we looked at it was like what is this. And then at the end of the course we were like, oh, I recognize this.</td>
</tr>
<tr>
<td>Better understood how science was performed</td>
<td>Science isn’t only being in the lab doing really cool things, you might say; science is most importantly the part in which we are analyzing data. All of the models that make up science as we know it now come from that.</td>
<td>It really is because I think one thing I learned is not everything will go according to what’s been predicted or planned. It’s going to keep on changing, so I think it, I mean, helps you to your mind towards it like - okay, not everything is going to work the way it should. So it… you know, it makes it a bit more adaptive and flexible in trying out different methods or techniques to find out a solution.</td>
</tr>
<tr>
<td>Discussed poster, paper, oral presentations</td>
<td>We also did presentations throughout the year so I felt like every week or couple weeks we would go up with our partner and present where we were at which was really helpful because it was like at that point not just working with your partner but having people be like well, you could try this or that next.</td>
<td>And I also wanted to say that we have a presentation at the end of the course, so I think that that presents our results to other teammates and basically from there we can infer different things and have different conclusions about our different contigs that we annotated. So I feel like it was a research team.</td>
</tr>
<tr>
<td>No right answer, construct an argument without knowing the right answer.</td>
<td>Well I always really like the feeling, just the… of accomplishment after you solve a particularly complex problem. Especially one that hasn’t been addressed before like this project where there’s really no… no one knows the right answer yet. And so I think it’s… it’s a real good… it’s a real strong sense of accomplishment. You feel like you’ve done something useful for the world, you know?</td>
<td>Also with scientific knowledge, it’s never really 100% a fact because we’re always… it’s always one of those things where we’re just waiting until something disproves it and/or strengthens it, you know? And it’s always a theory, never 100% a fact. And I think with scientific knowledge, it just strengthens… it gets to a point where it strengthens and strengthens and strengthens until something challenges it and then that challenge branches off to either be a part of that already existing theory or completely turns it around and disproves something we’ve known for so</td>
</tr>
</tbody>
</table>
long and we kind of follow a new branch and try to explore that. It’s always an explorable knowledge. It’s never an indefinite type of thing.

Novel Research Problem (Real Research)  
Okay. I think that it was definitely a very different experience than any other lab-based class you could have, since you’re actually dealing with novel research. But I think that between the integration of like actually doing the novel research and understanding — like learning the principles behind them, it was definitely like a very positive and like much more learning-based experience than previous lab-type classes.

I really liked the dynamic of integrating current research and that all the data that we are providing are going to benefit other people and that we are contributing to something real, not like the usual lab work that is just for us to observe the application of what we’ve learned in theory.

Table 3. Example comments from student surveys that reflect the GEP faculty use of strategies to emphasize active research participation within the classroom.

<table>
<thead>
<tr>
<th>Topic from Faculty Log</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel Research Problem (Real Research)</td>
<td>I think the effort of integrating research was a great strength to this course. We have done so many labs where our work was not important and it was getting old. It was special that we felt important and our work was valuable, it was not something that would be just thrown away.</td>
<td>I think it was cool that we got to actually conduct an experiment that was real. It made things seem more important than if we were just doing an experiment to do it.</td>
</tr>
<tr>
<td>Better understood genes, annotations, genomics</td>
<td>I thought this was a great class for anyone interested in genomics or even genetics in general. It goes in depth about the structure of genes, and gives students hands-on experience with many of the basic tools used by scientists working in genomics.</td>
<td>I think that despite the difficulty learning and understanding the research this course was valuable in teaching genomics. It is more beneficial to me to learn by doing the work than just by lecture.</td>
</tr>
<tr>
<td>Better understood how science was performed</td>
<td>This course was the most challenging course I had ever taken in my life. I have learned so much about how to conduct experiments, analyze data, and write scientific papers.</td>
<td>What was more valuable to me was being able to learn how researching works in general. Considering this is my first time, this gives me an opportunity to understand the difficulty of research a certain topic. I’ll be more prepared and know what to expect when I do any future research instead of guessing what researching might be like.</td>
</tr>
<tr>
<td>No right answer, construct an argument without knowing the right answer.</td>
<td>Understand that sometimes the answer is not always clear and to try many different approaches to try and figure something out. This was challenging but very useful and our professor was so helpful.</td>
<td>One of the strengths was that [students can] engage their knowledge with real scientific problems, but the fact that nobody knows if they are right was scary.</td>
</tr>
</tbody>
</table>
Student Outline

Understanding Eukaryotic Genes

These modules are designed to facilitate student learning of gene structure, transcription, translation, and splicing using the genome browser. Each Module begins with a lesson plan that describes the objectives, pre-requisites, homework, and class instructions for the Module, followed by short readings to explain key terms and concepts. Each Module also contains multiple questions to help assess student understanding of the biological concepts. This set of curriculum also includes a Glossary which explains the key terms used in the UEG modules.

The modules require you to use the genome browser to visually explore DNA sequence and other genomic features surrounding the tra gene in Drosophila melanogaster. The genome browser is an important bioinformatics research tool, that is being used in these exercises as a pedagogical tool. The official UCSC Genome Browser (https://genome.ucsc.edu/) and the GEP UCSC Genome Browser mirror (https://gander.wustl.edu/) are publicly available online. There are also other project-specific mirrors of the UCSC Genome Browser and UCSC Assembly Hubs available online [e.g., UCSC Genome Archive (https://hgdownload.soe.ucsc.edu/hubs/), EMBL-EBI Track Hub Registry (https://trackhubregistry.org/)].

Module 1 will provide you with an introduction to the genome browser and its navigation. The rest of the modules will provide further training in the use of this freely available tool to understand complex biological phenomena. It is recommended that you watch the supporting videos that explain concepts such as splicing and phase, RNA-Seq data and the TopHat algorithm, or specific functionalities (e.g., “Short Match”) of the genome browser. The collection of videos is available through the “Understanding Eukaryotic Genes (UEG)” playlist on the GEP YouTube channel.

As determined by your instructor, you will be working either independently or cooperatively to understand eukaryotic gene structure and produce a gene model by the end of Module 6. You will be required to work through the questions embedded in the modules and submit the answers to your instructor. Upon completion of the modules, you will be better equipped to participate in a course-based research experience involving gene annotation.

{UNDERSTANDING EUKARYOTIC GENES MODULES for students}

Annotation of a Drosophila gene walkthrough

This walkthrough uses the annotation of the CG31997 gene on the D. biamipes Muller F element to illustrate the comparative annotation protocol for the F element project. This document illustrates how you can investigate a portion of a genomic region in an annotation project using FlyBase, the Gene Record Finder, and the gene prediction and RNA-Seq evidence tracks on the GEP UCSC Genome Browser. The walkthrough starts by identifying a genomic feature of interest in a D. biamipes project based on the gene predictions tracks. FlyBase blastp search is used to compare the predicted protein sequence for this D. biamipes feature with a database of proteins from the informant genome (D. melanogaster) to identify the putative ortholog. Once a putative ortholog has been identified the walkthrough demonstrates how to identify the precise coordinates of each coding exon using NCBI BLAST searches and RNA-Seq data. The walkthrough also describes a consideration of splicing and the need for compatible splice donor and acceptor splice sites between adjacent coding exons in order to maintain the open reading frame of annotated splice isoforms. The walkthrough concludes by verifying aspects of the proposed gene model using the Gene Model Checker and it includes a sample F Element Annotation Report demonstrating the submission of an annotated gene model report.

Annotation Workflows

GEP faculty and students have developed workflows that summarize the key analysis steps and program parameters in the GEP annotation protocol. Many students find these workflows useful as a quick reference when
they are working on their own annotation projects. These workflows are also provided as part of the student handouts for your reference:

1. The “GEP Annotation Workflow” provides an overview of the key analysis steps and bioinformatics tools used for annotation in the F element project.

2. The “Identify D. melanogaster Ortholog” decision tree illustrates the list of criteria that you can use to identify the putative D. melanogaster ortholog of a predicted gene.

3. The “Annotating Splice Sites” workflow describes the logic and the different lines of evidence that could be used to identify the splice donor and acceptor sites of coding exons.

(ANNOTATION OF DROSOPHILA GENE and WORKFLOWS HANDOUTS for students)

Materials

IT and Audio-Visual Support & Classroom Requirements

These exercises require a classroom or computer lab with a computer lectern with easy-to-use controls connected to a high-resolution ceiling-mounted projector and a sound system. Students would also need a fast Ethernet or wireless connection to the Internet. If students use their own laptops or tablets, well-distributed power outlets and large work surfaces are important. Students devices should have access to an Office Suite (e.g., Microsoft Office, Office 365, Google Docs), a simple text editor (e.g., WordPad, TextEdit), and a modern web browser (e.g., Google Chrome, Mozilla Firefox). If students use computers in a computer lab, instructors may need to consult with their local IT administrators to adjust the security settings on the computers (e.g., enable JavaScript, cookies, pop-up windows).

Online Documents

Materials developed for use by GEP members and other educators are freely available from the GEP website (https://thegep.org/; under “Curriculum”).

The “Understanding Eukaryotic Genes” curriculum was developed by Margaret Laakso, Carina Howell, Cathy Silver Key, Leocadia Paliulis, Maria Santisteban, Chiayedza Small, Joyce Stamm, and Elena Gracheva (https://thegep.org/ueg/). Members of the GEP assisted by reviewing and revising these materials. All the modules are available with answer keys in a single package that uses the GEP name/password. Dr. Leocadia Paliulis (Bucknell University) has developed multiple videos that accompanied this curriculum and the videos are available on the GEP YouTube channel (https://www.youtube.com/c/GenomicsEducationPartnership).

The Annotation of the Drosophila Gene walkthrough was developed by Wilson Leung at Washington University in St. Louis (https://thegep.org/lessons/wleung-walkthrough-annotation_drosophila_gene/).

The three workflows included in the student handouts that provide a quick guide to the annotation protocol for the F Element project are available through the “Workflow” Curriculum Type tag on the GEP website (https://thegep.org/lesson-types/workflow/).

Tools for the GEP annotation projects are available through the “Projects” section of the GEP website (https://thegep.org/projects/). Each GEP science projects require a different set of annotation tools that are listed under the “Resources & Tools” section of the project page on the GEP website. For example, the Annotation of the Drosophila Gene walkthrough uses the annotation tools listed on the F Element project page (https://thegep.org/felement/).

Finally, PowerPoint presentations used in this workshop as well as the answer keys can be found in the Appendices.

Annotation Tools

Gene Model Checker (https://thegep.org/checker)

This tool tests whether the proposed model satisfies basic biological constraints (e.g., contains a start codon, stop codon, canonical splice sites), and
to visualize a comparison of the proposed gene model against the putative ortholog in *D. melanogaster*. This tool can help students identify and correct errors in the proposed gene models. The Gene Model Checker can also be used as a teaching tool to allow students to propose an initial gene model, receive feedback from their instructors, and then iteratively refine the gene model.

**Gene Record Finder** ([https://thegep.org/finder](https://thegep.org/finder))

This tool provides the gene structure information for the ortholog within *D. melanogaster* (i.e., the informant genome). The information is presented in graphical and table formats. Students can use this tool to examine the coding exon usage in the different isoforms of a *D. melanogaster* gene, and to retrieve the amino acid sequences for each coding exon to facilitate the mapping of coding exons against the target genome.

**Annotations File Merger** ([https://thegep.org/merger](https://thegep.org/merger))

This tool combines the annotation files for individual isoforms (produced by the Gene Model Checker) into a single file for project submission.

**Notes for the Instructor**

The CURE is introduced to students through faculty that have been trained in the tools, curriculum, and scientific background of the project (via in-person or online workshops). In some implementations, teaching assistants (TAs) also facilitate the learning process. TAs can be students who have previously taken the course and did well, or have successfully participated in an independent study project that utilizes GEP materials. Some of the TAs may have also received training from the GEP community at various regional node workshops. The GEP also has a group of virtual TAs that provide students and faculty with real time support in GEP activities ([https://thegep.org/taschedule/](https://thegep.org/taschedule/)).

TAs and peers are extremely valuable in CUREs because of the hands-on nature of this teaching. A faculty member might be able to help 5–6 groups of students and have insightful discussions about their project, approaches, assumptions, and conclusions. However, it becomes challenging to provide quality feedback to more groups. TAs can often help students who have questions on how to interpret analysis results (e.g., BLAST searches, Gene Model Checker checklists) and to resolve technical issues (e.g., with web tools, web browsers, operating systems, etc.). Hence TAs can help reduce the amount of time that a student would need to wait for help, and mitigate frustrations caused by technical issues. We have found that TAs are essential in guiding students in large sections with 40 or more students.

However, we caution faculty about providing students with too much help too early in the analysis process, as this can rob them of the aha moments and of the gains that come from struggling through the material. We have observed that students often struggle with annotation, yet find value in that struggle. Analysis of the faculty log, knowledge quizzes, and surveys from over a hundred faculty and their students supports the idea that frustration is “formative” for student learning (Lopatto et al. 2020). As mentioned above, iterations in genomics research can be done quickly and inexpensively. We therefore recommend creating a supportive environment of faculty, TAs, and peers that allows for formative frustration and iteration to enhance student learning.

We suggest communicating to students early on (and with periodic reminders) that they are participating in authentic research. Indeed, we have determined that making students aware that their work contributes to scientific knowledge improves both learning about genes and genomes and learning about the nature of science (Lopatto et al. in preparation). Further, in our recent study, faculty used up to nine different “research emphasis practices” in their course implementations. We find that the use of these active learning practices shows positive correlations with the average student benefit reported in the post-course student survey and with the score on the post-course annotation quiz. However, we did not find any single active learning practice that is more effective than others in affecting student performance.

We would also recommend performing frequent progress checks as students work on their research project. This approach breaks the whole project into smaller parts with well-defined goals that would avoid overwhelming the students. Having students turn in results at designated points along the process also helps catch fundamental problems early and also keeps students on track. For example, the
Pathways project developed an Annotation Notebook to help students keep track of their work (https://thegep.org/lessons/pathways-project-annotation-notebook). For annotation projects with multiple genes, it is a good idea to ask students to submit the annotation for one of the genes before they complete the entire project. This would help detect misconceptions or issues with the analysis approach early in the process, and avoid propagating the errors to all the gene models in a project.

As mentioned above, the UEG modules can be implemented by themselves and they are designed for students at the introductory undergraduate level. If teaching a junior/senior level course, students can spend two weeks working through a subset of the UEG modules (e.g., 1, 5, and 6), followed by one week on the Annotation of a Drosophila gene walkthrough, and one week on an annotation example. They would then be ready to work on their annotation projects for about 4–6 weeks.

If the implementation is a standalone course, students can be given a deeper introduction to BLAST in the beginning and work through an exercise (https://thegep.org/curriculumsearch; Curriculum Type: "Lesson with Exercises"). The GEP has also developed curriculum materials which focus on the analysis of RNA-Seq data, annotation of transcription start sites, and using de novo motif discovery tools (e.g., MEME Suite) for motif finding. To help GEP students better understand the limitations of bioinformatics tools, the GEP has developed curriculum materials on the algorithms used for sequence alignments (e.g., dynamic programming) and computational gene predictions (e.g., hidden Markov models). Depending on the pedagogical goals of the course, students can also perform activities beyond annotation, such as performing wet lab experiments to test hypotheses made during their annotation, and using multiple sequence alignments to examine the evolution of a gene across 27 Drosophila species.

Most faculty implementations require students to present their work either to the class at the end of the course or to the larger community, sometimes as a poster within an institution-wide research symposium. Many faculty also require their students to submit a written report that summarizes their analysis. For students participating in the research project, they will also need to complete the GEP Annotation Report form, and to prepare the annotation files for submission.

Acknowledgments

GEP is currently recruiting new members (https://thegep.org/contact/).

The GEP is supported by NSF IUSE-1431407 to Sarah C.R. Elgin, NIH R35GM133760 to Nathan T Mortimer, NSF IUSE-1915544 to Laura K. Reed, and NIH IPERT R25GM130517 to Laura K. Reed.
Cited References


About the Author(s)

Maria S. Santisteban
Department of Biology, UNC Pembroke, Pembroke, NC 28352

Dr. Maria Santisteban is a Professor of Biology at the University of North Carolina at Pembroke. Her research interests focus on the role that the H2A.Z histone variant plays in gene
regulation, which her lab investigates using the model organism *Saccharomyces cerevisiae*. She teaches Cell Biology (BIO 3710), Molecular Biology (BIO 4130), and Microbiology (BIO 3150) and has integrated GEP materials into the first two courses, as well as independent research student projects. Dr. Santisteban joined the GEP in 2012 and implemented GEP material for the first time in 2013. Currently she serves as chair of the GEP Curriculum Committee and NC node leader.

James Godde  
Department of Biology, Monmouth College, Monmouth, IL 61462

Dr. James Godde is the McGrath Professor of Biology at Monmouth College in Illinois, where he has taught for the past 20 years. His research focuses on repetitive elements of DNA such as trinucleotide repeats in eukaryotes and CRISPRs in prokaryotes. Dr. Godde teaches Bioinformatics (BIOC 300), Genetics (BIOL202), Microbiology (BIOL302), and Molecular Biology (BIOL354). It is in this final course that he has implemented GEP material. Dr. Godde joined the GEP in 2019 and taught students using the Understanding Eukaryotic Genes modules for the first time this same year. He is currently serving as Vice-chair of the GEP Professional Development and Mentoring Committee.

Wilson Leung  
Department of Biology, Washington University in St. Louis, St. Louis, MO 63130

Wilson Leung serves as the IT director for the Genomics Education Partnership (GEP). He helps maintain the GEP web framework and infrastructure. He creates the student projects and the evidence tracks for the *Drosophila* genome assemblies available on the GEP UCSC Genome Browser, and is the primary developer of the annotation tools (e.g., Gene Model Checker and the GEP Project Management System) used by GEP faculty and students. He also contributes to the development of GEP curriculum materials, and serves as an instructor for the GEP Faculty Training Workshops. He is the Program Manager for the G-OnRamp project, and a member of the Genomics Education Alliance (GEA) IT/Infrastructure Committee. His research interests include chromatin packaging and the evolution of the *Drosophila* Muller F element. He is a lab instructor for the Bio 4342 Research Explorations in Genomics course at Washington University in St. Louis — an upper-level laboratory course that implements GEP materials.

Catherine Reinke  
Department of Biology, Linfield University, McMinnville, OR 97128

Dr. Catherine Reinke is an Associate Professor of Biology at Linfield University in McMinnville, OR, where she has taught for the past 10 years. Her research interests focus on the biogenesis and regulation of microRNAs, which her lab investigates using the model organism *Drosophila melanogaster*. She teaches Principles of Biology (BIOL 210), Genetics (BIOL 270), Molecular and Cellular Biology (BIOL 400), and Animal Development (BIOL 420) and has used GEP materials to teach sophomores, juniors and seniors in Genetics, Molecular Cell Biology, and Animal Development, as well as independent research student projects. Dr. Reinke joined the GEP in 2012 and implemented GEP material for the first time in 2012. Currently she serves in the GEP Professional Development and Mentorship Committee as the Director of New Member Training.

---

**Appendices**

- PowerPoint presentations
- Answer keys
- COMPLETE STUDENT PACKAGE HANDOUT
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>“3 prime”; Refers to carbon 3 of the nucleic acid sugar component (either ribose in RNA or deoxyribose in DNA) to which additional nucleotides may be added by polymerase, often used to refer to that end of a single-stranded DNA or RNA molecule where the 3’ carbon retains its hydroxyl group (-OH) and no further nucleotides are bonded.</td>
</tr>
<tr>
<td>5'</td>
<td>“5 prime”; Refers to carbon 5 of the nucleic acid sugar component (either ribose in RNA or deoxyribose in DNA), to which the triphosphate is attached in a nucleotide triphosphate, often used to refer to that end of a single-stranded DNA or RNA molecule where the 5' carbon's phosphate group(s) is/are unattached to a preceding nucleotide.</td>
</tr>
<tr>
<td>alternative splicing</td>
<td>The inclusion or exclusion of certain exons in the splicing reactions that determine the sequences included in the final mRNA product. This mechanism is utilized to generate a series of closely related protein isoforms, which differ by the inclusion or exclusion of the particular protein regions encoded by those exons. Alternative splicing is directed by RNA-binding proteins that may block, or stimulate, utilization of a particular splice site.</td>
</tr>
<tr>
<td>amino acid</td>
<td>The basic building block of proteins, a small molecule with a -C-C- core, an amine group (-NH₂) at one end and a carboxylic acid group (-COOH) at the other end. The general structure can be represented as NH₂-CHR-COOH, where R can be any of 20 different functional groups of acidic, basic, or nonpolar character.</td>
</tr>
<tr>
<td>annotation</td>
<td>Gene annotation is the process of notating the location, structure, and identity of genes in a genome. As initial attempts may be based on incomplete information, gene annotations are constantly changing as further data becomes available. Gene annotation databases are updated regularly, and different databases may refer to the same gene/protein by different names, reflecting new knowledge and improved understanding of protein function.</td>
</tr>
<tr>
<td>base</td>
<td>Although formally incorrect (the nitrogenous base that defines A, C, G, T and U is only part of the whole nucleotide), this is often used as a synonym for &quot;nucleotide&quot; in referring to the A, C, G, T, and U components of DNA and RNA.</td>
</tr>
<tr>
<td>base pair/base pairing</td>
<td>The hydrogen bonding of one of the bases (A, C, G, T, U) with another, as dictated by optimal hydrogen bond formation in DNA (A-T and C-G) or in RNA (A-U and C-G). Two polynucleotide strands, or regions thereof, in which all the nucleotides form such base pairs are said to be complementary. In achieving complementarity, each strand of DNA can serve as a template for synthesis of its partner strand - the secret of DNA replication's extremely high accuracy and thereby of inheritance.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>canonical</td>
<td>In agreement with existing principles and standards generated from data and evidence. For example, the canonical <strong>splice donor sequence</strong> is GT. In rare instances, however, the sequence GC is used instead; since GC is not the “standard”, it would be referred to as <strong>non-canonical</strong>.</td>
</tr>
<tr>
<td>cDNA</td>
<td>&quot;complementary DNA&quot;; a double-stranded DNA molecule prepared in vitro (&quot;outside of the body&quot;; i.e., in a test tube) by employing an RNA molecule as a template to synthesize DNA using reverse transcriptase. The RNA component of the resulting RNA-DNA hybrid is enzymatically degraded, and the complementary strand then synthesized by DNA polymerase. The resulting double-stranded DNA can be used for cloning and analysis.</td>
</tr>
<tr>
<td>CDS</td>
<td>&quot;coding sequence&quot;; that part of the DNA sequence of a gene that is translated into protein.</td>
</tr>
<tr>
<td>coding exon</td>
<td>In a gene, any exon that contains some part of the CDS; in contrast, an exon that has no part translated into protein is called a <strong>&quot;non-coding exon.&quot;</strong></td>
</tr>
<tr>
<td>coding strand/positive strand</td>
<td>In a gene, the DNA strand that has the sequence found in the RNA molecule. Also called the sense, positive, or non-template strand.</td>
</tr>
<tr>
<td>codon</td>
<td>The sequence of three nucleotides in DNA or RNA that specifies a particular <strong>amino acid</strong>.</td>
</tr>
<tr>
<td>coordinate</td>
<td>Numerical position within a biological sequence; for example, the first base in a DNA sequence would have the coordinate “1”.</td>
</tr>
<tr>
<td>downstream</td>
<td>Refers to the genomic region that comes after the feature being examined.</td>
</tr>
<tr>
<td>exon</td>
<td>In eukaryotes, a contiguous segment of DNA that corresponds to a portion of the mature (processed) RNA product of that gene. Exons in eukaryotic genomes are often, but not always, separated by <strong>introns</strong>. Although exons are transcribed with the introns, the latter are spliced out during RNA processing and degraded.</td>
</tr>
<tr>
<td>feature</td>
<td>Any region of defined structure/sequence in a genomic fragment of DNA. Inherent features would include genes, pseudogenes, and repetitive elements. A feature may also be predicted by computational algorithms, such as those aimed at identifying protein-coding genes.</td>
</tr>
<tr>
<td>intron</td>
<td>Non-coding section of a eukaryotic nucleic acid sequence found between exons. Introns are removed (&quot;spliced out&quot;) from the primary transcript/pre-mRNA after transcription and before the molecule is exported to the cytoplasm for translation.</td>
</tr>
<tr>
<td>isoforms</td>
<td>Potentially different versions of a protein encoded by a single gene. Isoforms result from alternative splicing of a particular <strong>pre-mRNA</strong>, and/or the use of a different transcription start site.</td>
</tr>
<tr>
<td>mRNA</td>
<td>Mature messenger RNA that has been completely processed and is ready for translation; it has a 7-methylguanosine cap at its 5' end, a <strong>poly(A) tail</strong> at its 3' end, and has all its <strong>introns</strong> spliced out.</td>
</tr>
<tr>
<td>term</td>
<td>description</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>non-coding strand/ negative strand</td>
<td>Also called the anti-sense, template, or non-coding strand. This strand of the DNA sequence of a single gene is the complement of the 5' to 3' DNA strand known as the positive, sense, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.</td>
</tr>
<tr>
<td>nucleotide</td>
<td>The basic building block of DNA (A, C, G, T) and RNA (A, C, G, U). Nucleotides consist of a nitrogenous base, a 5-carbon sugar (either ribose in RNA or deoxyribose in DNA), and phosphate group(s).</td>
</tr>
<tr>
<td>ORF</td>
<td>&quot;Open reading frame&quot;; a long stretch of codons in the same reading frame uninterrupted by termination codons; an ORF may reflect the presence of a gene.</td>
</tr>
<tr>
<td>phase</td>
<td>The phase describes the number of bases between the end of the exon (defined by the splice site) and the full codon nearest that splice site. The number of bases between the adjacent full codon and an exon/splice site can be 0, 1 or 2. The phase of an upstream exon will determine which frame is translated in the downstream exon by indicating how many bases after the splice acceptor site are needed to create a full codon of 3 bases.</td>
</tr>
<tr>
<td>poly(A) tail</td>
<td>About 250 adenine nucleotides that are post-transcriptionally added by poly(A) polymerase to the 3' end of eukaryotic transcripts, following cleavage of the newly synthesized RNA ~20 nucleotides downstream of an AAUAAA polyadenylation signal sequence.</td>
</tr>
<tr>
<td>pre-mRNA (primary transcript)</td>
<td>The initial transcript from a protein-coding gene that contains both introns and exons. Pre-mRNA requires the addition of a 5' cap and 3' poly (A) tail and the removal of introns to produce the final mRNA molecule containing joined exons.</td>
</tr>
<tr>
<td>promoter</td>
<td>A segment of DNA to which RNA polymerase binds to initiate transcription of the downstream gene(s).</td>
</tr>
<tr>
<td>putative</td>
<td>Something that may be predicted or inferred but that requires more evidence to confirm or refute.</td>
</tr>
<tr>
<td>read</td>
<td>A raw DNA sequence.</td>
</tr>
<tr>
<td>reading frame/frame</td>
<td>A frame is a single series of adjacent nucleotide triplets in DNA or RNA: one frame would have bases at positions 1, 4, 7, etc. as the first base of sequential codons. There are three possible reading frames in an mRNA strand and six in a double stranded DNA molecule due to the two strands from which transcription is possible. Different computer programs number these frames differently, so care should be taken when comparing designated frames from different programs. One common way is to refer to the three possible left-to-right reading frames as +1, +2, and +3 and the three possible right-to-left reading frames as -1, -2, and -3.</td>
</tr>
<tr>
<td>splicing</td>
<td>The process by which introns are removed and exons are joined to produce a mature, functional RNA (mRNA) from a primary transcript. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>splice acceptor site</td>
<td>The <strong>splicing</strong> site at the 3’ end of an <strong>intron</strong>, at the boundary between an intron and the <strong>exon</strong> immediately <strong>downstream</strong>. The canonical splice acceptor site dinucleotide sequence is AG.</td>
</tr>
<tr>
<td>splice donor site</td>
<td>The splicing site at the 5’ end of an <strong>intron</strong>, at the boundary between an intron and the <strong>exon</strong> immediately <strong>upstream</strong>. The canonical splice donor site dinucleotide sequence is GT; in rare cases, the non-canonical sequence GC is used instead.</td>
</tr>
<tr>
<td>splice junction</td>
<td>Either a <strong>splice acceptor site</strong> or a <strong>splice donor site</strong>.</td>
</tr>
<tr>
<td>start codon (initiation codon)</td>
<td>The first codon of a <strong>CDS</strong>. In eukaryotes this is almost always ATG, which codes for methionine (one of the 20 <strong>amino acids</strong>).</td>
</tr>
<tr>
<td>stop codon (termination codon)</td>
<td>A codon that specifies the termination of protein synthesis; sometimes called a &quot;nonsense codon&quot; since it does not specify an <strong>amino acid</strong>.</td>
</tr>
<tr>
<td>transcription</td>
<td>The process of copying one strand of a DNA double helix by RNA polymerase, creating a complementary strand of RNA called the transcript.</td>
</tr>
<tr>
<td>translation</td>
<td>The process by which codons in an mRNA are “read” by the ribosome and tRNAs to direct protein synthesis.</td>
</tr>
<tr>
<td>TSS (transcription start site)</td>
<td>The location in DNA, generally <strong>upstream</strong> of a gene’s coding sequence, where RNA polymerase begins <strong>transcription</strong>.</td>
</tr>
<tr>
<td>UTR</td>
<td>&quot;Untranslated region&quot;; a segment of DNA (or RNA) that is transcribed and present in the mature <strong>mRNA</strong> but is not translated into protein. UTRs may be found at either or both of the 5’ and 3’ ends of a gene or transcript.</td>
</tr>
<tr>
<td>upstream</td>
<td>Refers to the genomic region prior to the <strong>feature</strong> being examined.</td>
</tr>
</tbody>
</table>
Module 1: Introduction to the Genome Browser: What is a gene?

Joyce Stamm

Objectives

- Demonstrate basic skills in using the UCSC Genome Browser to navigate to a genomic region and to control the display settings for different evidence tracks.
- Explain the relationships among DNA, pre-mRNA, mRNA, and protein.

Prerequisites

- Knowledge of:
  - DNA structure (base composition, anti-parallel double-stranded helix, base-pairing properties)
  - Chromosome structure (a chromosome is a continuous DNA molecule, basic understanding of chromosome arms)
  - Protein structure (proteins are made up of amino acids)

Class Instruction

- Discuss the question: What is a gene? (Discuss with a partner, then as a class.) Emphasize the function of a gene; consider how the structure of the gene is related to its function.
- Work through the genome browser investigation, with pauses to discuss the answers to the questions.
- Conclude with an emphasis on the main points:
  - Genes may run in either direction on a chromosome
  - Genes are represented on the genome browser as blocks connected by lines
  - Eukaryotic genes are made up of protein-coding exons (the blocks) connected by introns
  - Proteins usually begin with a Methionine (M) and end at a stop codon (*)

Associated Videos

- Genome Browser Video: https://youtu.be/6Ho6lxxmEec
- Tracks Video: https://youtu.be/zRm1Vr-lY8I
Table of Contents

Introduction to the Genome Browser ................................................................. 2
Genes are composed of exons and introns ......................................................... 7
Genes provide the information to make proteins ............................................... 8
Genes have directionality .................................................................................... 10
Coding exons are translated in a single reading frame ........................................ 11
Conclusion: ........................................................................................................ 15

Introduction to the Genome Browser

Genes encode information that our cells use to carry out their functions. In particular, protein-coding genes provide the cell with the information to make messenger RNAs (mRNAs), which are then used to make proteins. In this module, we will use a web-based visualization tool called a Genome Browser to explore the structure of a eukaryotic gene and obtain a basic understanding of how this information is stored and used. In subsequent modules, you will learn more about the details of these biological processes and use the Genome Browser to examine the experimental data that provide evidence for a detailed gene structure. The protein-coding genes in eukaryotes (higher organisms, with a cell nucleus) are much more complex than the protein-coding genes in prokaryotes (bacteria, organisms without a nucleus). We are still trying to figure out all of the details!

1. Start by watching the Genome Browser video.
2. Open a web browser and navigate to a custom version of the Genome Browser. The browser was developed by the Genome Bioinformatics Group at the University of California Santa Cruz (UCSC). The custom version is at http://gander.wustl.edu. Click on the "Genome Browser" link on the left menu (Figure 1).

Figure 1  Access the Genome Browser gateway page using the "Genome Browser" link.
3. Change the following fields in the "Genome Browser Gateway" section (Figure 2):
   - Select “D. melanogaster” under the “REPRESENTED SPECIES” field. This will allow you to view the genome of the insect *Drosophila melanogaster*.
   - Confirm that "Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)" is in the "Assembly" field. This is the version of the *D. melanogaster* genome that you will view. The genome assembly is simply the genome sequence produced after chromosomes have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together. A genome assembly is updated when DNA has been sequenced that allows gaps to be filled. It may also be updated when a new assembling algorithm is released. The August 2014 Drosophila melanogaster (BDGP Release 6 + ISO1 MT/dm6) assembly was produced by the Berkeley Drosophila Genome Project (BDGP).
   - Enter "chr3L" into the "Position/Search Term" text box so that you can view the left (L) arm of chromosome 3 (chr3).

4. Click on the “GO” button.

5. The next screen can be divided into four major sections (Figure 3):
   - A top green toolbar is used to navigate to the different tools provided by the Browser.
   - Navigation Controls allow us to navigate or zoom to different parts of the genome.
   - A genomic features panel (the white area) shows the locations of the different genomic features within the portion of the genome (e.g., chr3L) specified by the label next to the “enter position or search terms” text box.
   - A Display Controls section may be used to manipulate how much detail is visible in the genomic features panel of the Genome Browser. To match the screenshot in Figure 3, scroll down in this section to the bar labeled “Mapping and Sequencing Tracks”, go to “Base Position”, and select “dense” from the drop-down menu. The scroll down to “Genes and Gene Prediction Tracks”, go to “FlyBase Genes” and select “squish” from the drop-down menu. Check that all other tracks are set to “hide”, and then click on any “refresh” button to update the genomic features panel.
Figure 3  The four major sections of the Genome Browser.

You can use the buttons in the "Navigation control" section to navigate to different parts of the genome. You can zoom in to a region by clicking on one of the buttons next to the "zoom in" label (i.e., 1.5x, 3x, 10x, base). Similarly, you can zoom out by clicking on the buttons next to the "zoom out" label. Alternatively, you can enter the genome coordinates into the "enter position or search terms" field and then click on the "go" button to navigate to a specific region in the genome assembly.

The "size" field next to the "enter position or search terms" text box (red arrow in Figure 4) shows the total size of the genomic region that you are viewing. In this case, the "size" field shows that chr3L (i.e., the left arm of chromosome 3) in *Drosophila melanogaster* has a total length of ~28 million base pairs (bp). We will learn more about the key functionalities of the Genome Browser in subsequent modules. For now, we will focus on the large white rectangle shown on this page; this contains a graphical representation of the genomic features (e.g., protein coding genes, percent GC content) of chr3L mapped against the DNA sequence, which is embedded in the top line of the white box.

The different types of features (also known as “tracks” or “evidence tracks”) are separated by a title and are often shown in different colors. What types and how many tracks are shown in the view of genomic features is controlled by the display controls at the bottom. The view shown on Figure 4 displays only some of tracks in the “Gene and Gene Prediction tracks”, and all the other tracks in other sections (transgenic insertions, chromatin domains, ChIP seq tracks, Expression and Regulation, etc.) are “hidden”. More information about evidence tracks is available in the Tracks video.

We can examine the region under the blue title labeled “FlyBase Protein-Coding Genes” to estimate the number of protein-coding genes on chr3L. In this track each gene is represented by a set of blue boxes connected by thin blue lines. There are clearly fewer blue boxes at the right side of the image compared to the left, which suggests that genes are not uniformly distributed along the chromosome (Figure 4).
In the genome browser, each chromosome may be organized into smaller projects called contigs (for contiguous sequences). In this next part, we will examine contig1, a much shorter region in the left arm of chromosome 3.

6. Click on the "Genomes" link on the top toolbar to return to the Genome Browser Gateway page.
7. Change the assembly option to "July 2014 (Gene)" and verify that the "position" field has been set to "contig1" (Figure 5).
8. Click on the "GO" button.

The "size" field now has the value "**size 11,000 bp**", which means that contig1 has a total length of 11,000 bp.

To further explore the features on contig1, we will examine the results from two of the available tracks.
9. Scroll down to the "Display controls" section (i.e., green bars) to the bar labeled "Mapping and Sequencing Tracks" and verify that the display mode under the "Base Position" track is set to "dense" and the "FlyBase Genes" track is set to "pack".

10. The display mode for all other evidence tracks should be set to "hide" (Figure 6).

11. Click on any "refresh" button to update the Genome Browser image.

![UCSC Genome Browser on D. melanogaster July 2014 (Gene) Assembly (dm3gene)](image)

**Figure 6** Verify the display settings for the "July 2014 (Gene)" assembly.

Explore the contig1 genomic region using these tracks on the Genome Browser. You will observe distinct groups of connected boxes. These connected boxes and lines are genes, and their names are indicated on the left. Connected boxes and lines that are stacked vertically represent alternative forms of a gene, called isoforms. Answer the following questions:

**Q1.** How many genes are there in contig1?

**Q2.** What are the names of these genes?

**Q3.** Which gene has the largest span (i.e., the largest distance between the start and end of the gene)?

12. Now let's examine the gene at the end of this contig more closely. Type "contig1:9,841-9,870" into the "enter position or search terms" text box and then click on “go". (Note that you don’t need to use commas when entering base positions). The Genome Browser image will update to show only bases 9,841 to 9,870 of contig1. Note the letters that appear just below the base position numbers. These letters correspond to the nucleotide at each position. For example, both forms of the tra gene, tra-RA and tra-RB, begin with a T at position 9,851 (Figure 7).
13. Look at the left end of the display, under the word “contig1”. The arrow here is pointing to the right. When you click on the arrow, the arrow will switch orientation and point to the left (Figure 8). In addition, the nucleotides in the "Base Position" track will also change from black to grey. Clicking on the arrow again will return it to its original orientation.

**Figure 7** The Base Position track shows the underlying genomic sequence for a region when you zoom in.

**Figure 8** Click on the arrow to change the nucleotides shown on the base position track.

Q4. What is the relationship between the bases displayed when the arrow is pointed to the left versus when it is pointed to the right?

Q5. Why do you think the bases are displayed in this way in the Genome Browser?

Both forms of the *tra* gene begin at 9,851 and they have the same prefix ("tra") but different suffixes ("-RB" and "-RA", respectively). The prefix corresponds to the name of the gene (tra) in *D. melanogaster* while the two suffixes indicate that there are two different versions (i.e., isoforms) of this gene. We will examine the differences between these two isoforms later. For now, we will focus our analysis on the A isoform of *tra* (tra-RA).

**Genes are composed of exons and introns**

14. To see the entire *tra* gene, type “contig1:9,800-10,860” in the "enter position or search terms" text box and click “go” (Figure 9). Alternatively, you can use the buttons next to the “zoom out” label and the arrows next to the “move” label to adjust the display.
15. Carefully examine the tra-RA isoform. Notice that the isoform consists of black blocks that are connected by lines. On the lines are arrowheads that point from left to right. The black blocks are the exons (expressed regions of the gene; Figure 10). To use the information stored in a gene, a cell uses the DNA sequence as a template to produce a molecule called a messenger RNA (mRNA). This process is called transcription. You will see in module 2 that while the initial transcript (product of transcription) is continuous, copying all the DNA, only exon sequences are retained in the processed mRNAs. The lines connecting the blocks are the introns (intervening regions of the gene). These sequences will be removed during the production of mature mRNAs. The arrows on the lines denote the direction of transcription (or orientation) of the gene.

Q6. How many exons does tra-RA contain?

Q7. How many introns does tra-RA contain?

**Genes provide the information to make proteins**

The mRNA sequence contains the information that the cell needs to make proteins. You will learn more about this process in module 5. Here we will use the Genome Browser to examine the basic features of a protein.

16. Return to the Genome Browser, and type "**contig1:9,850-9,875**" into the “enter position or search terms” text box.
17. Scroll down to the “Mapping and Sequencing Tracks” section and change the display mode for the "Base Position" track to "full" (Figure 11).
18. Click on the "refresh" button to update your display.
Proteins are made up of amino acids, and the mRNA provides the information for the amino acid sequence. This information is read by the cell in groups of three bases, with each three-base group (i.e., codon) specifying an amino acid. The Genome Browser uses single-letter abbreviations to represent each amino acid. These are shown on your Genome Browser as three new rows of information directly below the DNA sequence (Figure 12).

**Q8.** Why do you think it takes three lines to display the amino acid information? Hint: remember that a codon is specified by three bases, e.g., CCG = Proline (circled in Figure 12).

Module 5 will have more details about translation, the process of copying the information from mRNA into protein. For now, we will just identify the beginning and the end of the protein. You should see three codons that are highlighted in green (one in row 1 and two in row 3). These codons all correspond to the amino acid M (i.e., Methionine). This amino acid is almost always used to start a protein. There is only one codon that can code for Methionine: **ATG**.

The first M on the third row of amino acids (at 9,858-9,860) corresponds to the start of the protein for the A isoform of *tra*. The position of this Methionine also coincides with the transition of the thinner rectangle to the thicker rectangle. Hence the thick rectangles denote coding sequence – the parts of the exon that carry information about the protein sequence and are the translated parts – while the thin blocks indicate regions that are part of the exon but do not carry protein sequence information, or the untranslated parts (Figure 13).
Let’s examine the other end of the protein. There are three special codons (known as stop codons) that signal the end of the protein. These codons (TGA, TAA and TAG) are indicated by an asterisk “*” and are highlighted in red in the "Base Position" track.

19. Type “contig1:10,740-10,765” into the “enter position or search terms” text box and then click on the “go” button. Note the stop codon (*) at position 10,754-10,756, specified by the bases TGA, in the second row of amino acids (Figure 14). This is the last codon before the transition from the thick exon block to the thinner one. The Genome Browser therefore shows that a part of the mRNA extends beyond the end of the protein-coding region. This is a general property of mRNAs: they contain extra sequences both before and after the protein-coding sequence. These sequences, at the 5’ and 3’ end of the protein-coding sequences, are called the 5’ and 3’ UTRs (untranslated regions) respectively.

Genes have directionality

As you saw above, the sequence of the codons in the A isoform of tra are read from left to right relative to the orientation of contig1. This also means that the start of the protein is located toward the left of the end of the gene. However, recall that DNA is double-stranded, and that the two strands run in opposite directions to each other (i.e., they are antiparallel). It turns out that, like the tra gene here, some genes are read on the DNA strand conventionally termed the ‘top strand’ (from left to right), while other genes are read on the ‘bottom strand’ (from right to left). We will examine one such example next.

20. Type “contig1:5,350-5,375” into the “enter position or search terms” text box and then click on the “go” button. This region contains the start of the protein-coding region of the CG32165 gene. However, there are no Methionines (green boxes) at the transition point between the thin and thick rectangles (Figure 15, top). However, note that the arrows in the thicker part of the indicated exon point from right to left, indicating that this gene is read from the bottom strand.
21. Click on the arrow beneath the "contig1" label in the "Base Position" track so that it points in the same direction as indicated for the gene in this region. This will complement the sequence and allow you to read the bases of the ‘bottom’ strand of DNA. Remember that the codons on this strand must be read from **right to left**. Now you can see that there is a start codon in this region, the corresponding green M amino acid (at 5,365-5,367) in the third row (Figure 15, bottom).

![Figure 15](image1.png)

**Figure 15** Examine the start of the coding region for a gene on the minus strand.

**Coding exons are translated in a single reading frame**

The combination of the directionality (with two alternative directions) and the three rows in the "Base Position" track means that there are six different ways to translate a genomic region, (i.e., to determine the sequence of amino acids from a DNA sequence). These different ways to translate a genomic region are known as **reading frames**.

22. To illustrate this concept, change the “enter position or search terms” text box to "**contig1:1-12**" and then click “go” in order to zoom in to the first 12 nucleotides of the contig1 sequence.

23. Click on the arrow underneath the "contig1" label in the "Base Position" track so that it points to the right (Figure 16).

![Figure 16](image2.png)

**Figure 16** Examine the "Base Position" track for the first 12 bases of contig1 in the top strand.
The first row (frame +1) begins at the **first** nucleotide in contig1 and the first amino acid (P) is derived from the codon **CCC**. The second row (frame +2) begins at the **second** nucleotide in contig1 and the codon **CCG** also codes for the amino acid P. The third row (frame +3) begins at the **third** nucleotide in contig1 and the codon **CGG** corresponds to the amino acid R (Figure 17). Because a codon is comprised of 3 nucleotides, the codon beginning at the fourth nucleotide (GGT) is again in frame +1.

![Figure 17](image1.png)  
**Figure 17**  Interpreting the reading frame using the Base Position track.

Examination of the "Base Position" track at the beginning of the contig shows that the three positive reading frames are numbered relative to the start of the contig1 sequence. Similarly, the three reading frames on the bottom strand are numbered relative to the end of the contig1 sequence (i.e., the beginning of the reverse complement of the contig sequence). Because contig1 has a total length of 11,000bp, we will change the "enter position or search terms" field to "**contig1:10,989-11,000**" so that we can examine the last 12 nucleotides of this contig.

24. Click on the arrow underneath the "contig1" label so that it points to the left (Figure 18).

![Figure 18](image2.png)  
**Figure 18**  Examine the "Base Position" track for the last 12 nucleotides of contig1 in the bottom strand.

Because we are examining the reverse complement of the contig1 sequence, we need to read the nucleotide and amino acid sequences on the "Base Position" track from right to left. The first row (frame -1) begins at the last nucleotide (11,000) of contig1 and the codon **TGC** codes for the amino acid C. The second row (frame -2) begins at the penultimate nucleotide at 10,999 and the codon **GCA** codes for the amino acid A. The third row (frame -3) begins at 10,998 and the codon **CAT** corresponds to the amino acid H (Figure 19).
25. Now that we understand how to interpret the reading frame information using the "Base Position" track, we can investigate the coding regions of the tra gene more closely. Change the "enter position or search terms" field to "contig1:9,800-9,960" and then click on the “go” button.

26. Click on the arrow underneath the "contig1" label in the "Base Position" track so that we can examine the translations of the top strand (running left to right) (Figure 20).

Our previous analysis shows that there is a start codon (green rectangle in the “Base Position” track) in the third row that corresponds to the transition from the thin to the thick rectangles (Figure 20). Hence the coding part of the first exon of the A isoform of tra is said to be “in frame +3”. Notice that there is also an open reading frame (ORF – stretch of codons uninterrupted by stop codons) that overlaps with the thick box in the second row (frame +2) but there are no start codons that overlap with the thick box. In contrast, the first row (frame +1) contains a start codon, but the thick box also overlaps with a stop codon (red star). When we examine the region downstream of the black boxes, we find that there are stop codons in all three reading frames. However, these stop codons do not interrupt the open reading frame of the first exon because they occur in the region of the arrowed lines (i.e., the first intron, see blue arrows in Figure 21).
Stop codons (red stars) are found in all three reading frames in the first intron of tra-RA.

27. Change the "enter position or search terms" field to "contig1:10,100-10,600" so that we can examine the second coding exon of the A isoform of tra to determine its reading frame.

Q9. Based on the screenshot shown in Figure 22, which reading frame contains the amino acid sequence for the second coding exon of tra-RA?

![Figure 22](image) The genomic region surrounding the second coding exon of tra-RA.

28. Change the "enter position or search terms" field to "contig1:10,550-10,900" so that we can examine the region surrounding the last coding exon of the tra-RA isoform (Figure 23). Based on our previous analysis, we know that there is a stop codon in the second row that corresponds to the transition from the translated (thick rectangle) to the untranslated (thinner rectangle) region of the mRNA (Figure 23). Hence the last coding exon of tra-RA is in frame +2.

![Figure 23](image) The terminal coding exon of tra-RA is in frame +2.

Q10. Does frame +2 have an ORF in the coding region of this exon? What about frame +1 and frame +3?

Q11. Given that 3 of the 64 possible codons are stop codons, what is the chance of having a stop codon at any given position, assuming that the sequence is random?
You might have noticed that the initial coding exon of tra-RA is in frame +3 while the last coding exon is in frame +2. We will learn more about mRNA processing in subsequent modules that will explain this apparent discrepancy.

**Conclusion:**

In this lesson, you have learned to use the basic navigation features of the Genome Browser to examine the basic structure of a eukaryotic gene. To summarize:

- Genes provide the information to make proteins. This information is captured by transcribing the DNA to make RNA and is carried on the mRNA in the form of three-base groups called codons.
- Genes are composed of exons and introns. Exons are regions retained in the processed mRNA and are represented by black blocks in the browser, while introns are the regions that are removed during the process of creating the final mRNA and are represented by lines connecting the blocks.
- The codon ATG in DNA (AUG in mRNA) specifies the amino acid M (Methionine) and is highlighted in green on the "Base Position" track of the Genome Browser. The first Methionine provides the starting signal for protein synthesis.
- The codons TAA, TAG, and TGA in DNA (UAA, UAG, and UGA in mRNA) encode the stop codon (*) and are highlighted in red on the "Base Position" track of the Genome Browser. The stop codons provide the ending signal for protein synthesis.
- Genes may be read either from left to right (top strand of the DNA) or from right to left (bottom strand of the DNA). Arrows on a gene indicate its directionality.
- Each row in the "Base Position" track (set on full) corresponds to a different reading frame. Different coding exons for a transcript can be in different reading frames.

29. To practice using the browser and reinforce the above concepts, examine the third gene in this contig (spd-2-RA):

**Q12.** How many exons and introns are present in this gene?

**Q13.** What is the orientation of this gene relative to contig1? How do you know? Where are the start codon and the stop codon - give the base position numbers (coordinates) of the start and the stop codon?)

You have now completed module 1 and are ready to move on to module 2.

**Q14. Bonus:** Take a little time to explore some of the other evidence tracks on the browser. While looking at contig1 (size 11,000 bp), put the “GC Percent” track on **full**. What sort of pattern do you see, relative to the map of the genes? What can you conclude about gene structure?
Module 2: Transcription Part I: From DNA sequence to transcription unit

Maria S. Santisteban

Objectives

- Describe how a primary transcript (pre-mRNA) can be synthesized using a DNA molecule as the template.
- Explain the importance of the 5’ and 3’ regions of the gene for initiation and termination of transcription by RNA polymerase II
- Identify the beginning and end of a transcript using the capabilities of the genome browser (RNA-Seq, Short Match)

Prerequisites

- Understanding Eukaryotic Genes Modules: [https://thegep.org/ueg/](https://thegep.org/ueg/)

Class Instruction

- Discuss the questions: What is transcription? What cellular proteins are required for transcription? How does it work mechanistically? What is/are the products of transcription? (students discuss in pairs, then as a class)
- Work through the genome browser investigation, then identify where transcription starts and ends for the tra gene. How long is the pre-mRNA?
- Conclude by challenging students to think about these questions:
  - How important is it for RNA polymerase II to recognize the promoter sequence?
  - Do you think it is possible for a gene to have more than one transcription start site? How would RNA polymerase II know which one to choose? When would it make a difference in the protein product, and when not?

Associated Videos

- RNA-Seq and TopHat Video: [https://youtu.be/XD-egRcHYL4](https://youtu.be/XD-egRcHYL4)
- Short Match Video: [https://youtu.be/3I-CYhzPSCE](https://youtu.be/3I-CYhzPSCE)
Investigation 1: Identify the Transcription Unit

Introduction

This module will introduce you to the use of the Genome Browser to illustrate the process of transcription and help you identify regulatory elements, using the *Drosophila melanogaster* transformer (*tra*) gene as an example. You will use the UCSC Genome Browser Mirror developed by the Genomics Education Partnership (GEP), which contains RNA expression data, to identify the different parts of the gene that give rise to pre-mRNA through transcription.

Finding the transcript for tra-RA using the UCSC Genome Browser Mirror

1. Open a new web browser window and go to the UCSC Genome Browser Mirror site at [https://gander.wustl.edu/](https://gander.wustl.edu/). Follow the instructions given in Module 1 to navigate to the contig1 project in the *D. melanogaster* “July 2014 (Gene)” assembly.
2. To navigate to the genomic region surrounding the *tra* gene, enter “contig1:9,650-11,000” into the “enter position or search terms” field located just above the displayed tracks and then click on the “go” button. As you learned in the previous module, you can also use the buttons in the navigation controls section to zoom in, zoom out, and use the arrows to move to different parts of the contig. In addition, you can place your cursor on the “Scale” or the “Base Position” sections of the Genome Browser image and then drag your cursor from the initial position to the end position to zoom into a region of interest.
3. This region from 9,650-11,000 contains the entire *tra* (transformer) gene and the very end of the previous gene *spd-2* (spindle defective 2). As described in Module 1, the suffix (e.g., -RA) corresponds to the name of the isoform that is associated with the gene. Hence, *spd-2-RA* corresponds to the A isoform of the *spd-2* gene.
4. Because the Genome Browser remembers your previous display settings, we will hide all the evidence tracks and then enable only the subset of tracks that we need: Click on the “hide all” button located below the Genome Browser image. Then, configure the display modes as follows:

- Under “Mapping and Sequencing Tracks”
  - Base Position: full
- Under “Gene and Gene Prediction Tracks”
  - FlyBase Genes: pack
- Click on any of the “refresh” buttons to update the display (Figure 1)

Note: Depending on your screen resolution, you may need to zoom in further to see the nucleotides and amino acid translations even if you set the “Base Position” track to full.

![Image](UCSC_Genome_Browser_on_D._melanogaster_July_2014_Gene_Assembly_dm3gene.png)

**Figure 1** Configuring the display modes for the evidence tracks surrounding the tra gene.

**Identifying the transcription unit for the tra gene**

5. Now let’s investigate how the string of As, Ts, Cs, and Gs of the DNA sequence in this genomic region give rise to the mRNAs for the tra gene. The “FlyBase Genes” track shows the protein-coding genes that have been annotated by FlyBase. According to this track, there are actually two different mRNAs (tra-RA and tra-RB) made from the same DNA sequence (Figure 2). These represent two alternative forms known as isoforms of the transformer (tra) gene product.
Figure 2  FlyBase annotated isoforms A (blue arrow) and B (red arrow) of *tra* in *D. melanogaster*.

6. For the moment, we will focus only on the A isoform of *tra* (tra-RA). As you learned in Module 1, the black boxes represent the exons (the part of the transcript that makes up the mRNA); the thick black boxes represent the translated regions (i.e., the parts of the exons that contain information that codes for protein) while the thinner black boxes represent untranslated regions (i.e., the part of the exons that do not contain information that codes for protein). Lines that connect multiple boxes together represent introns, the parts of the transcript that are removed in the production of a mature mRNA. Collectively, they constitute the transcription unit, the part of the gene that is read by RNA polymerase II during transcription.

We use the name “transcription unit” rather than “gene” because genes also contain regulatory sequences (promoters and both positive and negative regulatory elements) that are not transcribed. In contrast to prokaryotes, where most of the transcript codes for protein in a single open reading frame (no introns!), in eukaryotes, the transcript contains a lot of extra nucleotides that are not used to form the protein.

Q1. What is the span — the start and end base positions — of the tra-RA transcription unit?

7. The Genome Browser contains tracks that we can use to visualize the regions of the DNA that are transcribed into RNA. For example, the “RNA Seq Tracks” section contains results from sequencing (mostly mature) mRNAs and then mapping the sequences found in the RNA-Seq reads back to the genome. Hence, regions with RNA-Seq read coverage usually correspond to regions in the genome that are being transcribed. To visualize the distribution of these RNA-Seq reads, scroll down to the bottom of the page and then click on the “RNA-Seq Coverage” link under the “RNA Seq Tracks” section header (Figure 3).

Figure 3  Click on the “RNA-Seq Coverage” to configure the display settings for this evidence track.
8. Using the controls in the “RNA-Seq Read Coverage” page that comes up when you click the “RNA-Seq Coverage” link, we will modify the display settings to the following (Figure 4):
   - Change the “Display mode” field to **full**
   - Select the “Data view scaling” field to **use vertical viewing range setting**
   - Change the “max” field under “Vertical viewing range” to **37**
   - Under the “List subtracks” section, unselect the “Adult Males” track
   - Click on the “Submit” button (“Display mode” line, near the top of the page)

By default, the RNA-Seq Coverage track will auto-scale based on the read depth (that is, the number of reads) in the viewing region. The settings above override this setting and manually define the scale to be from 1 to 37. The RNA-Seq Coverage track contains data from mRNA isolated from two separate samples, adult males and adult females. Here we unselect the “Adult Males” track so that the Genome Browser will only show the RNA-Seq read coverage from adult females. We will return to the “Adult Males” track in Module 6.

![Figure 4](https://example.com/figure4.png)  
Manually define the viewing range for the RNA-Seq Read Coverage track (red arrows) and select only the subtrack of interest (i.e., Adult Females, blue arrow).

9. The Genome Browser image now includes a track in blue with peaks and valleys, labeled “modENCODE RNA-Seq from *D. melanogaster* Whole Adult Females” (Figure 5). The y-axis corresponds to the number of RNA-Seq reads from whole adult females that have been mapped to each genomic position of this portion of contig1.
Q2. How do the peaks in the RNA-Seq Read Coverage track relate to mRNA abundance?

Q3. Most of the RNA-Seq reads come from mature (processed) RNA. Can you use this data to suggest where introns are located? Are there any regions that seem ambiguous?

In subsequent modules, we will learn more about the mRNA processing that occurs in the nucleus to remove introns prior to translation.

**Investigation 2: Identify the 5′ end of the transcription unit**

**Introduction**

Previous studies have identified sequence motifs that are enriched in the region surrounding a gene’s Transcription Start Site (TSS). This region is known as the core promoter. By convention, we designate the TSS as +1 and we specify the positions of the sequence motifs with respect to the TSS. For example, the initiator (Inr) motif is found at -2 relative to the TSS (2 bp upstream) while the TATA box motif is found at -31 or -30 relative to the TSS. Both of these motifs are in the same orientation as the transcript (Figure 6).
Figure 6  Motifs that are enriched near the transcription start sites of many eukaryotic genes (Juven-Gershon T and Kadonaga JT, 2010). Note that the motifs are often “degenerate,” N = any base, R = purine (either A or G), Y = pyrimidine (C or T), K = keto (T or G), M = amino (C or A), S = strong (G or C), W = weak (A or T), V = A/G/C (not T), D = A/G/T (not C).

In this module, we will review three lines of evidence to determine the TSS position(s) for the *tra* gene. Because RNA-Seq identifies regions of the genome that are being transcribed, we will use the RNA-Seq Coverage track to define the scope of the region to search. The start of the region with RNA-Seq read coverage is the 5’ end of the transcript and corresponds to the approximate TSS site, (i.e., the beginning of the transcription unit). RNA-Seq data hence becomes our first line of evidence to try to determine the location of the TSS. In other words, the information gathered from RNA-Seq will be used to support the choice of the TSS. To learn more about RNA-Seq, watch the RNA-Seq and TopHat video.

Q4. Examine the “RNA-Seq Coverage” and the “FlyBase Genes” tracks in the Genome Browser from left to right. At approximately which coordinate (base position) does the RNA-Seq data start for the *tra* gene? Remember that you can use the navigation controls at the top of the page to zoom in to the region of interest.

One of the first steps in mRNA processing is the addition of the 5’ cap at the beginning of the transcript (we will learn more about capping in the next module). There are experimental techniques that specifically isolate the sequences that are associated with the 5’ cap. These sequences or “reads” can then be mapped against the genomic assembly, and the TSSs will show higher read density than the rest of the genome. The modENCODE project summarizes these experimental data to produce a set of predicted TSSs; these predictions are shown in the “TSS Annotations” track. The TSS annotations predicted by modENCODE constitute our second line of evidence to determine the *tra* TSS location.

---

1. Before we turn on this track, we will zoom into the region between the end of the previous gene (\textit{spd-2}) and the region where we see RNA-Seq data for \textit{tra-RA}. Change the “enter position or search terms” field to “contig1:9,700-9,900” and then click “go”. We expect the RNA polymerase to bind and initiate transcription somewhere in this area. Scroll down to the “Gene and Gene Prediction Tracks” section and change the display mode for the “TSS Annotations” track to “pack”. Click on a “refresh” button (Figure 7).

![Image of genome browser](image_url)

\textbf{Figure 7} Annotated TSS’s in the region surrounding the start of the \textit{tra} transcripts.

\textbf{Q5.} How many TSS sites were identified using this technique?

\textbf{Q6.} Look at the labels next to each of the annotated TSSs. What are the labels for the TSS sites?

2. We will examine each of the annotated TSS’s separately to determine their precise coordinates. First, let’s zoom in on the feature labeled “TSS\_tra\_16584216” so you have about 11 nucleotides in view. Look at the ruler in the “Base Position” track to determine the coordinate for this TSS (Figure 8).

\textbf{Q7.} What is the coordinate for TSS\_tra\_16584216?
3. Now let’s zoom in to the second TSS site, TSS_tra_16584150 (Figure 9).

**Q8.** What is the coordinate for this TSS?

4. We will use the Genome Browser to gather additional evidence to identify the most likely TSS. First, let’s search for the Inr motif using the **Short Match** functionality under “Mapping and Sequencing Tracks”. Note that we expect this motif to overlap with the TSS (i.e., from -2 to +4 relative to the TSS). The presence of Inr motif in the 5’ region of the gene will be our third line of evidence to support the most likely TSS location. Change the “enter position or search terms” field to **contig1:9,700-9,900** and then click “go”.

5. To learn more about the Short Match functionality, watch the [Short Match video](#). Scroll down to the “Mapping and Sequencing Tracks” section and click on the “Short Match” link. Change the “Display mode” field to **pack** and the “Short (2-30 base) sequence” field to **TCAKTY** (Figure 10). Click on the “Submit” button.
   
   - **Note:** “TCAKTY” is the consensus sequence for the Inr motif, where K (Keto) denotes either G or T and Y (pYrimidine) corresponds to either C or T.
6. Each box in the “Perfect Matches to Short Sequence (TCAKTY)” track corresponds to an instance of the motif. The sign “+” or “−” next to each bar denotes the orientation of the match while the number corresponds to the first base of the motif match.

**Q9.** Are there any perfect matches to the Inr consensus sequence in the region between 9,700-9,900? What are the coordinates and orientation of these matches?

**Q10.** Which base position(s) would you assign as the TSS of the *tra* gene based on the available evidence? Describe your reasoning.

**Q11.** Is there any ambiguity? In other words, do the three lines of evidence (RNA-Seq tracks, TSS as predicted by the modENCODE data, and the Inr consensus sequence location) point to exactly the same position as being the TSS? If they don’t, why might they differ? Could there be more than one TSS?

Let’s look at a different promoter region. Navigate to the Genome Browser Gateway page by clicking on the Genomes tab at the top of the page and select the *D. melanogaster* “Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)” assembly. Change the “Position/Search Term” field to “chr2R:18,867,350-18,867,430” and then click on the “GO” button.

7. Click on “hide all” and then enable the following tracks under “Mapping and Sequencing Tracks”:
   - Base Position: **full**
   - Short Match: **pack**

8. Search for TCAKTY, the Inr consensus sequence.
   - Click on the Short Match Link under “Mapping and Sequencing Tracks.”
   - Type “TCAKTY” in the “Short (2-30 base) sequence” field.
   - Click on the “Submit” button.

9. Under “Genes and Gene Predictions Tracks”:
   - FlyBase Genes: **pack**

10. Under “Expression and Regulation”:
    - TSS (Embryonic) (R5): **pack**
    - Click on a “refresh” button. Record the position(s) and orientation(s) of any matches to the Inr motif.
    - Repeat the search for TATAWAAR (the TATA Box motif).
Use the “Short Match” track to search for the Inr and TATA box motifs.

Q12. Are there any perfect matches to the Inr consensus sequence (Figure 11)? What are the coordinates and orientation of these matches? What about the TATA Box motif? Are these signals in good agreement with the beginning of the transcription unit?

Investigation 3: Map the 3’ end of the transcription unit

Introduction

After RNA polymerase II has started transcribing a gene (initiation), generally with the help of various transcription factors, it will proceed (in a process called elongation) all the way to the termination signal in order to produce a molecule of pre-mRNA. Let’s review what we know about the template, and then consider termination.

RNA polymerase II will use the template DNA to synthesize a primary transcript (pre-mRNA) by pairing purine bases with pyrimidine bases. Actually, the sequence of nucleotides that you observe on the tracks in the Browser corresponds to the “coding strand” of the DNA (complementary to the template strand); the coding strand is almost identical to that pre-mRNA, except that DNA has thymine versus RNA, which has uracil as the pyrimidine base that pairs with A.

Because DNA is antiparallel, if the coding strand that you see in the browser track runs 5’ to 3’, then the template strand runs in the 3’ to 5’ direction.

RNA polymerase binds to the promoter sequence on the template strand, constructing the transcribed mRNA in the 5’ to 3’ direction, just like the coding DNA strand that you see on the tracks. In fact, polymerases can only add nucleotides to the 3’ end (free –OH) of the growing RNA molecule.
Termination of mRNA transcription is different in eukaryotes than in prokaryotes. In eukaryotes, RNA polymerase II passes through one or more **AATAAA** sequences, which lie beyond the 3’ end of the coding region (i.e., thick black boxes in the FlyBase Genes track). The pre-mRNA molecule will thus carry the signal AAUAAA\(^2\). This AAUAAA signal is recognized by a special endonuclease that cuts at a site 11 to 30 nucleotides to its 3’ side. As you will learn in the mRNA processing module, a tail of polyriboadenylic acid, poly(A), is added by a special non-template-directed polymerase to the end of the transcript.

Pre-mRNA processing will be further studied in Modules 3 and 4.

1. We will try to identify the approximate end of the tra-RA transcript using the RNA-Seq data and will then search the DNA sequence for a termination signal (AATAAA). Return to the “July 2014 (Gene)” assembly, change the “enter position or search terms” field to “contig1:10,700-10,950” and then click “go” to navigate to the 3’ end of the tra gene. Examine the RNA-Seq read density in the “RNA-Seq Coverage” track (Figure 12).

![Figure 12](image.png)

**Figure 12**  Comparison of the transcription end site annotated by FlyBase (red arrow) versus changes in RNA-Seq read coverage in whole adult females (blue arrows).

**Q13.** At which base position do you see the RNA-Seq read coverage ending in the whole adult female sample? Zoom in close to the beginning of the pink area (no RNA-Seq coverage) in the RNA-Seq track.

**Q14.** What is the coordinate of the 3’ end of the tra-RA transcript according to the “FlyBase Gene” track? You will need to zoom in on the end of the FlyBase genes tracks.

You may observe that after decreasing, the amount of RNA-Seq reads in this region starts to increase again, continuing at a higher level to the end of the contig. This is because there is another gene downstream (to the right) very close to tra. We can ignore the region (starting at around position 10,900) where the RNA-Seq reads increase.

---

\(^2\) This signal is also referred to as the poly-A signal because a poly-A tail is added to the mRNA at its 3’ end. In the next module we will refer to it as a poly-A signal.
2. We will now look for a termination signal in this 3’ region of the tra gene. As we did when searching for the Inr consensus sequence, we can use the “Short Match” functionality to search for the AATAAA sequence.

3. Click on the “Short Match” link under the “Mapping and Sequence Tracks section”. Verify that the “Display mode” is set to pack and enter the sequence “AATAAA” into the “Short (2-30 base) sequence” field (Figure 13). Click on the “Submit” button.

![Figure 13](image_url) Use the “Short Match” track to search for the mRNA termination signal.

Q15. How many matches are there in the search region (contig1:10,700-10,950)?

Q16. How many of these matches are on the positive (+) strand of the DNA? Remember these sequences, like the Inr consensus sequence we discussed before, are strand specific and your gene is on the + strand.

Q17. Is the sequence(s) you found in the question above contained within the 3’ untranslated region of the transcript? Remember from Module 1 that the thick black boxes in the “FlyBase Genes” track represent coding (translated) regions while the thin black boxes represent non-coding (untranslated) regions.

Q18. Based on your analysis above, which position is the best choice for the termination signal? Describe your reasoning.

**Conclusion:**

In this lesson, you have seen how the primary transcript (the mRNA molecule) is produced from the template DNA by an RNA polymerase interpreting different signals on the DNA. We saw that DNA sequences upstream of the 5’ end (promoter) and near the 3’ end (terminator) are important parts of the transcription unit. The pre-mRNA molecule from the TSS site to the termination signals will undergo several modifications (processing) in addition to capping that you will learn about in the next few modules.
As discussed above, the reads produced by an RNA-Seq experiment are derived primarily from processed mRNA (not the pre-mRNA). Hence, we can explore several additional questions using the RNA-Seq Coverage track:

**Q19.** Do you see any correlation between the areas with high RNA-Seq read coverage (high peaks) and the different boxes in the tra-RA isoform? Zoom out 10X to get an overview. Remember that the thick boxes correspond to the coding regions, the thin boxes are the untranslated regions, and the lines with arrows are introns.

**Q20.** Where do you see regions in the RNA-Seq coverage data with no coverage at all?

**Q21.** If these regions with no RNA-Seq coverage occur within an initial transcript, what could have happened to these RNA sequences?
Module 3: Transcription Part II: What happens to the initial (pre-mRNA) transcript made by RNA pol II?

_Catherine Silver Key & Chiyedza Small_

**Objectives**

- Explain how the transcript generated by RNA polymerase II (the pre-mRNA) is processed to become mature mRNA, using the sequence signals identified in Module 2.
- Use the genome browser to analyze the relationships among:
  - pre-mRNA
  - 5’ capping
  - 3’ polyadenylation
  - splicing
  - mRNA

**Prerequisites**

- Understanding Eukaryotic Genes Modules: [https://thegep.org/ueg/](https://thegep.org/ueg/)
- Define pre-mRNA as the RNA that results from the process of transcription; this initial transcript includes exons and introns

**Class Instruction**

- Discuss the questions: What happens to the initial (pre-mRNA) transcript made by RNA pol II? Does it leave the nucleus ‘as is’ or do changes have to occur? (Hint: introns vs. exons) (Discuss with a partner then as a class).
- Mini-presentation illustrating that during pre-mRNA processing, three events occur:
  - 5’ capping,
  - 3’ polyadenylation
  - splicing out of introns
- Work through the genome browser investigation, with pauses to discuss the answers to the questions.
- Conclude with emphasis on main points:
  - Pre-mRNA is processed using 3 steps:
    - 5’ capping,
    - 3’ polyadenylation,
    - removal of introns through splicing (via spliceosome)
Introduction

In Module 2, you identified the transcription start site (TSS) for the A isoform of the *tra* gene (tra-RA). In this module, we will explore each of the three steps of pre-mRNA processing.

Setting up our Browser page (review):

1. Open a new web browser window and go to the UCSC Genome Brower Mirror site at [http://gander.wustl.edu/](http://gander.wustl.edu/). Follow the instructions given in module 1 to navigate to the contig1 project in the *D. melanogaster* "July 2014 (Gene)" assembly.
2. As you may remember from Module 1, contig1 is derived from chr3L in the *D. melanogaster* genome. This contig contains three different genes (CG32165, spd-2, and tra). Enter "contig1:9,500-11,000" into the "enter position or search terms" textbox and then click on the "go" button to navigate to the genomic region surrounding the tra gene.
3. Because the Genome Browser remembers your previous display settings, you should click on the "default tracks" button to reset the display to the default settings. Change the display mode for the "Base Position" track to "full" and verify that the "FlyBase Genes" track is set to "pack". Click on the "refresh" button.
4. Scroll down to the "RNA Seq Tracks" section and then click on the "RNA-Seq Coverage" link. Change the track display settings to the following, as we did in Module 2:
   5. Change the "Display mode" field to "full"
   6. Select the "Data view scaling" field to "use vertical viewing range setting"
   7. Change the "max" field under "Vertical viewing range" to 37
   8. Under the "List subtracks" section, select BOTH the "Adult Females" and the "Adult Males" (Select the check box next to subtrack to turn the subtrack on.)
   9. Click on the "Submit" button. Verify that the RNA-Seq Coverage track on the browser page is set to “full.”
Investigation: mRNA processing

The processing of pre-mRNA into mRNA involves three key steps (Figure 1):

- The addition of a 5’ cap
- The addition of a 3’ poly(A) tail
- The removal of introns through splicing

Removal of the introns during this process results in adjacent exons being brought together in the final mRNA message.

![Diagram of mRNA processing that converts a pre-mRNA to a processed mRNA.](image)

The first step in pre-mRNA processing occurs at the 5’ end of a messenger RNA. Recall that mRNA is synthesized in a 5’ to 3’ direction, so the 5’ end of the mRNA was synthesized first. Let’s examine the beginning of the tra gene. Type "contig1:9,825-9,870" into the "enter position or search terms " textbox and then click on the "go" button.

In Module 2, we identified the transcription start site (TSS) of the A isoform of tra at position 9,851.

To show the TSS’s that have been annotated by the modENCODE project, scroll down to the "Genes and Gene Prediction Tracks" and change the display mode for the "TSS Annotations" track to "pack", and click refresh. The modENCODE project looked for TSSs by using a chemical method to tag the special structure that occurs at 5’ ends of transcript, fishing out the RNA molecules that carried these tags, and mapping the sequence back to the genome, a method called “CAGE” (cap analysis of gene expression).
In addition, we will also display the "D. mel. cDNAs" track (also under the "Genes and Gene Prediction Tracks" section); change this to "pack". This track shows the alignment of *D. melanogaster* cDNAs (complementary DNAs, made by copying the mRNA) that have been sequenced by the Berkeley Drosophila Genome Project (BDGP). Click on the "refresh" button (Figure 2). These two tracks both provide an analysis based on the RNA population and mapping the positions of these sequences indicates where the transcript started.

Remember from Module 2 that we also found a match to TCAKTY, a common initiation signal just upstream, at 9,834 (display this using the “Short Match” track). All of these pieces of evidence argue for a TSS in this region.

![Figure 2](image)

**Figure 2** Change the display modes of the "D. mel. cDNAs" and "TSS Annotations" tracks to "pack".

The new Genome Browser image (Figure 3) shows the 5' end of the pre-mRNA transcript (i.e., the start of transcription) based on the CAGE experiment (modENCODE track) with the additional lines of support. On this end of the pre-mRNA, a modified guanine nucleotide (7mG) is added to the nucleotide at position 9,851, forming the 5' cap. Note that this additional nucleotide is **NOT** visible in the DNA track. It is added AFTER the transcript is made. This is the **first step** in pre-mRNA processing: capping.

![Figure 3](image)

**Figure 3** Addition of a 5' cap to the 5' end of the transcript.

**Q1.** What is the coordinate of the first nucleotide that is transcribed? In the DNA sequence, is it an A, C, T or G?
Q2. What are the coordinates for the start codon that codes for the first amino acid of the A isoform of the *tra* gene? (Assume reading frame +3.)

Q3. The region of the transcript from the 5’ cap to the nucleotide just upstream of the start codon is called the 5’ untranslated region (5’UTR) because it is part of the transcript that is not translated. How long (in ribonucleotides) is the 5’UTR?

The **second step** in pre-mRNA processing is **polyadenylation**.

10. To view the 3’ end of the tra-RA gene, change the "enter position or search terms" field to "contig1:10,633-11,000" and then click on the "go" button.

Polyadenylation means that **many** (poly) **adenine nucleotides** (ribonucleotides) are added to the 3’ end of the pre-mRNA **AFTER** transcription termination. This generates a poly-A tail (typically ~20 to ~250 As) that will be retained in the final mRNA, but it is not present in the "Base Position" track of the Genome Browser. This is because the poly-A tail does not exist in the DNA template but is simply added to the RNA by a special polymerase as a long run of adenine nucleotides.

Our previous analysis in Module 1 has shown that the last coding exon of tra-RA is in frame +2 and the stop codon is located at 10,754-10,756. We can use the Genome Browser to determine the end of the tra-RA transcript indicated by the cDNA track (in blue). (Note that this aligns with the cDNA although there is some discrepancy between the two as to the exact end of the transcript.)

Q4. How long (in base pairs) is this 3' untranslated region (3'UTR) as indicated by the cDNA track (in blue)?

Q5. **Zoom into the 3’ end of the FlyBase Gene**, near the termination site. What is the longest stretch of A nucleotides that you observe?

Q6. **Do your findings support the conclusion that the poly(A) sequence observed in the mature mRNA transcript is not in the template DNA?**

11. **Perform a "Short Match" search for the poly-A signal (AATAAA)** using the protocol you learned in Module 2. This search should place the poly-A signal at 10,818-10,823 (Figure 4). As mentioned in Module 2, the transcript is cleaved 11 to 30 nucleotides downstream of the poly(A) signal sequence, and then 150-200 adenines are added to the pre-mRNA. The nucleotides between the stop codon and the end of the poly-A tail comprise the 3' UTR.
Figure 4  Previous analysis placed the stop codon for the A isoform of tra in frame +2 (blue arrow) and the poly-A signal at 10,818-10,823 (red arrow).

We can see the polyadenylation sequence that is associated with the processed mRNA by examining the cDNA (BT028774) that has been aligned to this region.

12. Click on the "BT028774" feature under the "cDNAs from D. melanogaster" track and then click on the "View details of parts of alignment within browser window" link (Figure 5).
Figure 5  Examine the alignment of D. melanogaster cDNA BT028774 against contig1.

The next figure shows the actual alignment between the *D. melanogaster* cDNA BT028774 and the genomic sequence in contig1 (Figure 6). Nucleotides that are identical between the two sequences are shown in blue capital letters while nucleotides that differ are shown as black lowercase letters. The light blue bases denote the start and the end of the gap in the alignment. The side-by-side alignment shows the pairwise alignment between the cDNA (top) and the contig1 sequence (bottom) within the viewing region (i.e., contig1:10,633-11,000).
Q7. Scroll up to the 'cDNA BT028774' area. After which coordinate (number in the cDNA) do you see the polyadenylation track (in lower case black letters)?

Q8. How many ‘A’ ribonucleotides have been added to the tra mRNA (represented in the cDNA)?

Q9. Locate the AATAAA termination signal in the cDNA sequence. How many nucleotides 3’ of the final ‘A’ in the signal sequence does the poly(A) run start? (This number should be between 11-30 nucleotides.)

The final step in pre-mRNA processing is splicing out of introns and merging adjacent exons into one continuous open reading frame so that the mRNA is ready for translation into a protein.
13. Change the "enter position or search terms" field to "contig1:9,870-10,170" and then click on the "go" button to navigate to the first intron of the tra-RA transcript (Figure 7).

![Image of genomic region surrounding the first intron of tra-RA]

**Figure 7** The genomic region surrounding the first intron (red arrows) of tra-RA.

14. Zoom in to the region near the end of the first exon of tra-RA.

**Q10.** Which two nucleotides are found just after the end of the first exon of tra-RA? ________

Repeat this determination, identifying the two nucleotides at the start of intron 2 of tra-RA. ________

These two nucleotides are a signal for **splicing** to occur at the 5’ end of an intron; these represent the first two bases of the intron, often called the donor site (or 5’ splice site).

**Q11.** At which base does exon 1 end? __________

15. Zoom out and then zoom in to the region near the beginning of the second exon of tra-RA.

**Q12.** Which two nucleotides are found right before the start of tra-RA exon 2? ______

16. Zoom out and then zoom in to the region near the beginning of the third exon of tra-RA.

**Q13.** Which two nucleotides are found right before the start of tra-RA exon 3? ______

These two nucleotides are the signal for **splicing** out of the 3’ end of the intron, often called the acceptor site (or 3’ splice site). These represent the last two bases of the intron.

**Q14.** At which base does exon 2 of tra-RA begin? What is its coordinate? __________

**Conclusions**

In this module, we learned about the three key steps that are involved in converting the pre-mRNA into a mature mRNA:
The addition of a 5' cap
The addition of a 3' poly(A) tail
The removal of introns through splicing
  - Note that introns are removed during this process and adjacent exons are brought together in the mRNA message

After mRNA processing, the mature mRNA (tra-RA) can now exit the nucleus so that it can be translated into a protein (tra-PA) by the cytoplasmic ribosomes.
Module 4: Splicing – Removal of introns from messenger RNA by splicing

Meg Laakso

Objectives
- Identify splice donor and acceptor sites that are best supported by RNA-Seq and TopHat splice junction predictions.
- Utilize the canonical splice donor and splice acceptor sequences to identify intron-exon boundaries.

Prerequisites
- Understanding Eukaryotic Genes Modules 1–3: [https://thegep.org/ueg/](https://thegep.org/ueg/)
- Define pre-mRNA as the RNA that results from the process of transcription; this initial transcript includes exons and introns

Class Instruction
- Review pre-mRNA processing using appropriate figures from the textbook or Module 3
- Investigation 1: Students familiarize themselves with RNA-Seq data
  - Review consensus sequences for splice donor and splice acceptor sites
- Investigation 2: Students find splice donor and acceptor for intron 1
  - Review RNA splicing of intron 1 using tra-RA as the example
- Investigation 3: Students find remaining splice donor and splice acceptor for intron 2
  - Discuss length of pre-mRNA vs. length of spliced mRNA
  - Identify isoforms with different TSSs or alternative splicing patterns

Associated Videos
- RNA-Seq and TopHat Video: [https://youtu.be/XD-egRcHYL4](https://youtu.be/XD-egRcHYL4)
- Genes and Isoforms Video: [https://youtu.be/ce6nVSiiv7I](https://youtu.be/ce6nVSiiv7I)

Table of Contents

Investigation 1: Examining RNA-Seq data ................................................................. 2
Investigation 2: Identifying splice sites ................................................................. 6
Investigation 3: Identify the 5' splice donor and 3' splice acceptor sites for intron 2 .......... 10
Homework Part 1: Identifying splice sites for intron 1 ............................................. 12
Homework Part 2: Identifying the 5' splice donor and 3' splice acceptor sites for intron 2 ........ 13
Investigation 1: Examining RNA-Seq data

We will continue to focus on isoform A of transformer (referred to as tra-RA). Here we will focus on data from experiments that assess the RNA population in cells. This data can be used to help us identify exons and introns for the gene under study.

All RNAs in the cell are collectively known as the “transcriptome,” as almost all RNA is produced by transcription from a DNA template. (In some cases, RNA is made from an RNA template.) The transcriptome includes messenger RNAs, ribosomal RNAs, transfer RNAs, and other RNAs that have specialized functions in the cell.

RNA can be harvested from cells or a whole organism like Drosophila and converted to DNA, then sequenced to produce RNA-Seq (RNA Sequencing) data. First, extracted mRNA that has been fully spliced is copied back to DNA with the enzyme called reverse transcriptase. Short fragments of the copied or complementary DNA are sequenced, and then these segments are mapped back to the genome. By analyzing the mapping data, it is possible to know which and how many messenger RNAs have been synthesized.

This is a powerful technique that allows us to see when and where different genes are expressed. This kind of information can help researchers and clinicians know which genes are expressed in different types of cancer, for example. Here and in the next modules we are going to use RNA-Seq data to explore how the transformer (tra) gene is expressed in male vs. female Drosophila.

RNA-Seq data can indicate where transcription occurs, to the exact nucleotide. The number of RNA-Seq fragments that map to a given site also tells us how many copies of the RNA are present in the sample. Remember from Module 3 that the initial RNA transcripts are quickly processed to remove the introns. Hence in total RNA from a cell, sequences from exons will be much more abundant than sequences from introns. We will use RNA-Seq data to help us find the exon-intron boundaries for tra-RA, the isoform of the tra gene that is expressed in female fruit flies.

1. Open a new web browser window and go to the GEP UCSC Genome Brower Mirror site at https://gander.wustl.edu/.
2. Select the “Genome Browser” link from the left side panel. (Detailed instructions for navigating to the Genome Browser Gateway page are given in Module 1.)

Reminder: use the dropdown menus on the Genome Browser to select the following:

1. Select “D. melanogaster” under the “REPRESENTED SPECIES” field.
2. Select “July 2014 (Gene)” in the “D. melanogaster Assembly” field.
3. Enter “contig1” in the “Position/Search Term” field.
4. Then click “GO” and the screen below will appear (Figure 1). As you will remember, this section of DNA is 11,000 base pairs long and is part of the left arm of chromosome 3, which is about
28,100,000 bp long. If your Browser window is showing other evidence tracks, reset by clicking on “default tracks.”

![UCSC Genome Browser on D. melanogaster July 2014 (Gene) Assembly (dm3gene)](image)

**Figure 1** A screenshot of “contig1” project.

5. Let’s start by setting the evidence tracks we want to see. Click on the “hide all” button. Then open only the tracks that will provide data for this investigation. Note that you will not be able to see the DNA sequence until you “zoom in.”

6. Change the RNA-Seq coverage to “full,” then click “refresh.”

You will see blue and red histograms representing RNA-Seq data generated using RNA samples from adult females and adult males, respectively. This allows us to infer the pattern of RNA synthesis, and to look for similarities and differences between the sexes. Both similarities and differences are apparent!

7. Zoom in to view only the *tra* gene by entering **contig1:9,800-10,900** in the “enter position or search terms” text box (Figure 2).

8. We are now looking at the region of chromosome 3 where the *tra* gene is located. Compare the blue and red histograms for Adult Females and Adult Males. Note that there are numbers on the y-axis that show how many RNA reads (sequences from transcripts) map to that position.
Figure 2  Histograms representing RNA-Seq data in transformer (tra).

Q1. List two ways in which the histograms are similar.

a. _______________________________________________________________

b. _______________________________________________________________

Q2. List one way that the histograms differ (other than color).

a. _______________________________________________________________

To recap: The blue histogram represents the sequenced messenger RNA from female fruit flies and represents the form of the tra gene referred to as isoform A (tra-RA). The red histogram represents the RNA-Seq data from male fruit flies. The mRNA in males is different from that in females, and this second isoform of the tra gene is called isoform B (tra-RB).

Q3. Recall that our other evidence (see Module 3) indicates that tra-RA (blue, female-specific) extends from 9,851 to 10,846. How many gaps do you see in the histograms in this interval? __________

Each gap corresponds to an intron — there is very little RNA-Seq signal for that part of the gene because the intronic RNA was spliced out before the mRNA was collected and sequenced. [*Hint: Notice that in the first intron, the RNA-Seq histogram for tra-RA (blue, female-specific) has a long stretch of very low bars. (The blue bars are much shorter than the bars for the expressed exons, and much shorter than the corresponding region in the red histogram of RNAs from male fruit flies.) This
region is part of the intron in female flies, and it reflects the alternative splicing event that creates the female-specific isoform A. However, since a small fraction of the RNA-Seq reads is derived from pre-mRNAs, we can find intronic sequences in the RNA-Seq data. **We can use a subset of the RNA-Seq reads that span adjacent exons to verify that this region is part of the intron in the mRNAs produced by female flies (e.g., exon junction predictions derived from spliced RNA-Seq reads, discussed next in Module 4, Investigation 2).**

**Q4.** How many exons do you see? ______________

Remember that the exon is the “expressed” part of the gene and there will be either a sharp, or broad, peak in the RNA-Seq histogram.

**Q5.** Do females or males make more transformer mRNA, or do they express it at about the same level? __________________

Now that you've examined the evidence yourself, let's go back and review.

- **Gaps (introns)** – Figure 3 is a screenshot of the genome browser with gaps circled. Note that there are 2 gaps in females, and 2 gaps in males. The first gap in females looks a little strange because it doesn’t have clean boundaries, suggesting a mixed population of processed transcripts.
- **Exons** – Brackets have been drawn underneath the RNA-Seq data for Adult Female flies, corresponding to the three exons that are expressed.
- **Isoforms** – Note that isoform A and isoform B of the *tra* gene are the result of alternative splicing. For other genes, isoforms may have different transcription start sites. Isoforms of a gene always have different mRNA sequences, but they may have the same protein sequence. To learn more about isoforms and genes, watch the [Genes and Isoforms video](#).
Q6. Using the information you’ve gathered so far, make a diagram of the tra-RA (female-specific) isoform with 3 exons and 2 introns. Represent exons as rectangular boxes and introns as lines connecting the boxes. Number each exon and intron (start from the left with “exon 1”).

Q7. Where is the promoter in relation to the exons and introns? _____________________________.
On your diagram, mark the putative Transcription Start Site in your diagram with a bent arrow pointing in the direction of transcription.

Investigation 2: Identifying splice sites

We will again focus on tra-RA to identify splice sites, that is, the exact nucleotides where splicing occurs to remove introns from the pre-mRNA.

Background:

Two software programs, called TopHat and Bowtie, use the RNA-Seq data to graphically represent the exon junctions. The resulting graphic on the genome browser coincidentally looks something like a little bowtie (two small boxes connected by a thin line) (Figure 4). The boxes represent the sequenced mRNA (the exons), and the line represents a gap (the intron). The exon junction can be inferred when
the first part of a sequenced fragment from the RNA population matches (for example) DNA positions 50-100 and the second part of the same fragment matches DNA positions 200-250; the RNA from positions 101-199 must have been taken out of the middle!

Short sequences are present at the beginning and end of each intron that allow the spliceosome—the molecular machinery that cuts out introns—to precisely remove the intron, leaving only the exon sequences in the mature mRNA. The first two nucleotides of the intron are the splice donor site and almost always the nucleotides “GT”. The last two nucleotides of the intron are the splice acceptor site and almost always the nucleotides “AG”. (Recall your observations in Module 3.) For more information on RNA-Seq and the search for splice junctions, watch the RNA-Seq and TopHat video.

![Figure 4](image.png)

**Figure 4** A diagram of intron-exon junctions.

9. Using the same Genome Browser page, reset the Browser by clicking on “**hide all**.” Then open the tracks that will provide the information we want for Investigation 2. *(Note: If you stopped after Module 4: Investigation 1, then you may want to return to page 2 for a reminder of how to get to the Genome Browser page.)*

10. Change base position to “**dense.**”

Note that you will not be able to see the DNA sequence until you “zoom in.”

11. Change RNA-Seq coverage to “**full,**” then click “**refresh.**”

You will again see blue and red histograms representing the RNA-Seq data (indicating the amount of mRNA synthesized) in females and males, respectively. We will focus on the blue histogram (Adult Females) again.

12. As we did in Module 3, let’s customize the RNA-Seq track by setting the “Data view scaling” field to “**use vertical viewing range setting**” and the “max” field under “Vertical Viewing range” to 37. Remember that you gain access to these settings by clicking on the “RNA-Seq Coverage” link under the RNA-Seq Tracks green bar.

13. Change Exon Junctions to “**full,**” then click “**refresh.**”

The rectangular boxes joined by a thin black line will help you identify the exon-intron boundaries. Our graphical output viewer will look like the screenshot below (Figure 5).
Figure 5  View of the tra RNA-Seq Data with the splice donor site in intron 1 circled.

14. Zoom in to the area that is circled — click and drag the cursor just above the numbers, or use zoom buttons.
15. Set the screen so that you can see about 15-20 nucleotides, as shown in the example below (Figure 6).

Figure 6  TopHat data (in black) for adult males and adult females indicating an exon junction.

The blue histogram stops at the end of exon 1. The last three nucleotides of exon 1 are “G-A-G.”

Q8. What is the coordinate of the last nucleotide of tra-RA exon 1? ____________________

Q9. What are the first two nucleotides of tra-RA intron 1? ____________________
This is called the 5' splice site or “splice donor site.”

16. Zoom out so that you can see all of the tra gene. Let's use TopHat to help us find the 3' end of the intron. Examine the Exon Junctions track. The first intron-exon junction predicted by TopHat (black) seems to align with the red histogram data from males; the second junction aligns better with the blue histogram data from females (Figure 7).

![Figure 7](image)

**Figure 7** TopHat and RNA-Seq data for males and females.

17. Let’s examine the 3' end of intron 1 more closely. Change the “enter position or search terms” field to contig1:10,140 and then click “go.” Zoom out 10x and then 3x (Figure 8).

![Figure 8](image)

**Figure 8** Graphical viewer centered on the junction between intron 1 and exon 2 of the tra-RA (female specific) isoform.

**Q10.** What are the last two nucleotides of tra-RA intron 1? ________________________________

This is called the 3' splice site or 'splice acceptor site'.

**Q11.** What is the coordinate of the first nucleotide of tra-RA exon 2? ________________
Investigation 3: Identify the 5’ splice donor and 3’ splice acceptor sites for intron 2

We can use the same approach to map the exon-intron boundaries of the second intron.

Review steps #2–5 in Investigation 2. Then repeat the process to answer the following questions about the 5’ splice donor and 3’ splice acceptor sites for the tra-RA (female specific) intron 2.

18. Zoom into the sequence surrounding the 5’ splice donor for the tra-RA (female specific) intron 2.

Q12. What are the last three nucleotides of tra-RA exon 2? __________________________

Q13. What is the coordinate of the last nucleotide of tra-RA exon 2? __________________________

Q14. What are the first two nucleotides of tra-RA intron 2? __________________________

19. Zoom out as needed so that you can see all of intron 2. Use TopHat to find the 3’ end of intron 2.

20. Click and drag so that the end of intron 2 is centered in the viewer. Then zoom in so that you can see the nucleotide sequence.

Q15. What are the last two nucleotides of tra-RA intron 2? __________________________

Q16. What is the coordinate of the first nucleotide of tra-RA exon 3? __________________________

Q17. Using the information you’ve gathered so far, make a graphical picture of the tra-RA (female specific) isoform with 3 exons and 2 introns. Number each exon and intron at the corresponding DNA coordinates. Add the coordinates for first and last nucleotide of the exons that you have found so far. Add the sequences of the splice donor and splice acceptor sites at the appropriate locations.

Q18. Where do you think the promoter is located in relation to your gene model? What evidence do you have to support your idea, using the evidence tracks we have displayed (Base Position, RNA-Seq Coverage, Exon Junctions)?

Q19. Bonus question! Support your hypothesis by gathering additional data. Recall our explorations in Modules 2 and 3. You might want to open the tracks “D. mel. cDNAs,” and “TSS Annotations,” both in full. What type of evidence is shown by each of these tracks (refer to Module 2)? Finally, to see tra-RA as currently annotated in FlyBase, open the “tra Isoform” track on full, or to see both isoforms open the “FlyBase Genes” track in pack. Do these results support your model? Do any ambiguities remain?
Module 4 Homework: Determining splice sites for the \textit{spd-2} gene

\textit{Meg Laakso}

21. Open a new web browser window and go to the GEP UCSC Genome Brower Mirror site at https://gander.wustl.edu/.

22. Follow the instructions given in Module 1 to navigate to the contig1 project in the \textit{D. melanogaster “July 2014 (Gene)” assembly}.

\textbf{Reminder:} use the drop-down menus on the Genome Browser to select the following:

- Under “REPRESENTED SPECIES” field = “D. melanogaster”
- Under “D. melanogaster Assembly” field = “July 2014(Gene)”
- In the “Position/Search Term” field = “contig1”

Then click “GO” and the screen below will appear (Figure 9). As you will remember, this section of DNA is 11,000 base pairs long and is part of the left arm of chromosome 3, which is about 28,100,000 bp long. If your Browser window is showing other evidence tracks, reset by clicking on “default tracks.”

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{contig1.png}
\caption{A screenshot of “contig1” project.}
\end{figure}
23. Let's start by setting the evidence tracks we want to see. Click on the “hide all” button. Then open only the tracks that will provide data for this investigation.

- **Base position “dense”** - Note that you will not be able to see the DNA sequence until you zoom in. If the base position is changed to “full”, you can see the amino acid tracks also.
- **RNA-Seq coverage “full”** - You will see blue and red histograms representing RNA-Seq data generated using RNA samples from adult females and adult males, respectively.
- **Exon Junctions “full”** - Viewing the exon junctions will help you to find the splice donor and splice acceptor sites.
- **D. mel. cDNAs “full”** - This track was used extensively in previous modules and is useful for confirming the sequence of the mature mRNA. Remember that a cDNA is a DNA copy of an mRNA.

Zoom in to view only the spd-2 gene by entering “contig1:5,750-9,800” in the “enter position or search terms” text box.

We are now looking at the region of chromosome 3 where the spd-2 gene is located.

**Q1.** How many exons does spd-2 have? __________

**Q2.** How many introns does spd-2 have? __________

**Homework Part 1: Identifying splice sites for intron 1**

You remember from class that short sequences are present at the beginning and end of each intron that allow the spliceosome to precisely remove each intron, leaving only the exon sequences in the mature mRNA. The first two nucleotides of the intron are “GT”, and the last two nucleotides are “AG” (Figure 10).

![Figure 10](image)

**Figure 10** A bowtie diagram.

24. Zoom in to the end of exon 1. Set the screen so that you can see about 15-20 nucleotides.

**Q3.** What is the coordinate of the last nucleotide in exon 1? ________________

**Q4.** What are the first two nucleotides of intron 1? ________________
25. Zoom out so that you can see all of intron 1 and use TopHat to help find the end of the intron. Examine the Exon Junctions track.
26. Then zoom in so that you can see the nucleotide sequence.

Q5. What are the last two nucleotides of intron 1? ________________________

Q6. What is the coordinate of the first nucleotide in exon 2? ________________

Homework Part 2: Identifying the 5' splice donor and 3' splice acceptor sites for intron 2

Let's use the same approach to map the exon-intron boundaries for intron 2.

27. Review the steps you used in Part 1, then repeat the process to answer the following questions about the 5' splice donor and 3' splice acceptor sites for intron 2.
28. Zoom in to the sequence surrounding the 5' splice donor for intron 2.

Q7. What is the coordinate of the last nucleotide in exon 2? ________________

Q8. What are the first two nucleotides of intron 2? ________________________

29. Zoom out as needed so that you can see all of intron 2. Use TopHat to find the end of intron 2.
30. Click and drag so that the end of intron 2 is centered in the viewer. Then zoom in so that you can see the nucleotide sequence.

Q9. What are the last two nucleotides of intron 2? ________________________

Q10. What is the coordinate of the first nucleotide in exon 3? ________________

Q11. Using the information you've gathered so far, make a graphical picture of the spd-2 gene with 3 exons and 2 introns. Number each exon and intron. Add the coordinates for first and last nucleotide of the exons that you have found so far. Add the sequences of the splice donor and splice acceptor sites at the appropriate locations. Finally, add a bent arrow for the transcription start site.
Module 5: Translation

Carina Endres Howell & Leocadia Paliulis

Objectives

- Determine the codons for specific amino acids and identify reading frames by looking at the Base Position track in the genome browser
- Assemble exons to maintain the open reading frame (ORF) for a given gene
- Define the phases of the splice donor and acceptor sites and describe how they impact the maintenance of the ORF
- Identify start and stop codons of an assembled ORF

Prerequisites

- “Understanding Eukaryotic Genes” Modules 1–4: https://thegep.org/ueg/
- Overview of the ribosome, tRNAs, and associated proteins involved in translation (Initiation Factors, Elongation Factors and Release Factors)
- Overview of the DNA codon table

Class Instruction

- Review the process of translation: Overview of the ribosome, tRNAs, and associated proteins involved in translation (Initiation Factors, Elongation Factors and Release Factors)
- Review the DNA codon table
- Work through the activities using the Genome Browser, with pauses to discuss the answers to the questions.
- Conclude with emphasis on main points:
  - mRNAs are translated into amino acids using triplet codons.
  - Identification of ORFs.
  - The ORF must be maintained across splice sites to generate a working mRNA.
  - The assembled ORF begins with a start codon and ends with a stop codon.

Associated Videos

- Splicing and Phase Video: https://youtu.be/9VWjyu3PUJ0

Table of Contents

Investigation 1: Examining Open Reading Frames (ORFs) in the tra gene ......................................................... 2
Investigation 2 ......................................................................................................................................................... 5
Investigation 1: Examining Open Reading Frames (ORFs) in the *tra* gene

Introduction: Review of reading frames.

In this exploration, we will continue to focus on the *transformer* gene (referred to as *tra*-RA or just *tra*), and will learn about how the *tra* mRNA is translated into a string of amino acids.

Given that DNA is double-stranded, and that the genetic code is based on triplets (3 consecutive bases), there are six possible reading frames. One can determine a reading frame by dividing the sequence of nucleotides in DNA or RNA into a set of consecutive, non-overlapping triplets. There are three possible reading frames (read 5'→3') in the forward direction on the top strand of DNA, and three (read 5'→3') in the reverse direction on the complementary bottom strand of the same DNA molecule. Hence, there are six possible reading frames for each gene (see illustration in Module 1).

Once it is determined in which direction a particular gene is transcribed (for review see Modules 2 and 3 on transcription), there remain three choices for the reading frame. To determine which of these reading frames is used during translation, evidence such as the presence of an initiation codon and the absence of stop codons is used. As you learned in Module 1, the initiation codon is ATG in the coding DNA strand (AUG in the mRNA) and specifies the amino acid methionine. Additional triplets code for the other 19 amino acids, and three triplets are stop codons, causing termination of translation. These stop codons are TAA, TAG and TGA in DNA, or UAA, UAG and UGA when found in mRNA. An Open Reading Frame is a string of consecutive codons that is uninterrupted by stop codons. Every mRNA contains one ORF that is translated by the ribosome from start codon to stop codon.

Let’s investigate reading frames for the *tra* gene.

1. Go to the GEP UCSC Genome Browser Mirror site at [https://gander.wustl.edu/](https://gander.wustl.edu/) and follow the instructions given in Module 1 to open contig1 of *Drosophila melanogaster* using the July 2014 (Gene) assembly.
2. The screen below will appear (Figure 1). As you will remember, this section of DNA is 11,000 base pairs long and is a small part of the left arm of chromosome 3, which is about 28,100,000 bp long.
3. Zoom in to view only the first exon of the tra-RA gene by entering **contig1:9,840-9,900** in the “enter position or search terms” text box and hitting the go button.

4. Open only the tracks that will provide information for this investigation. Set the Base Position track to **full**, and click **refresh**. Now we can see the amino acid tracks as well, giving the results from conceptual translation.

5. There are three possible reading frames for this transcript in the forward direction, here indicated by the numbers 1, 2 and 3 (red arrow) (Figure 2).

6. Three reading frames are possible in the forward direction, as one could start translating with the first base, the second base or the third base. In this case, starting translation with the first base shown in Figure 2 (i.e., position 9,840 in contig1) leads us to the reading frame 3. Starting with the second base leads us to the reading frame 1, and starting with the third base lead us to the reading frame 2. Starting with the fourth base is equivalent to starting with the first base, only missing the first codon. The DNA “top strand” is read from left to right (indicated by the arrow in the browser, right under the word “contig1” that looks like this \( \rightarrow \)). If you click on that arrow, the three reading frames in the reverse direction will appear, as the “bottom
“strand” is read from right to left. As you learned in Module 1, genes have directionality. The *tra* gene is read from the “top strand” (left to right).

7. Notice that the third reading frame has a green M codon (methionine) at the location where the thick black rectangle indicates the first coding exon or CDS (Coding DNA Sequence) of tra-RA mRNA. Remember that the codon for methionine (ATG in DNA) is the start signal, the first codon used in translation. This gives us our first piece of evidence that reading frame 3 is the one used in translation of the first CDS of the *tra* gene. For simplicity let’s call this first CDS “CDS1” to distinguish it from other CDSs in the *tra* gene. Note that there is a stretch of RNA transcript upstream (to the left) of the ATG; this is the 5’UTR (5’ untranslated region), found at the 5’ end of all eukaryotic mRNAs.

8. Next carefully examine reading frame 2. Notice that in this reading frame there is no codon for methionine (no start codon) in the region that maps to the first exon. This gives us evidence that reading frame 2 is probably not being used during translation of CDS1 of the *tra* gene.

9. Finally, look at reading frame 1. Notice that there is a stop codon at the beginning of that reading frame (indicated by a red box with an asterisk in it). This evidence indicates that reading frame 1 probably is not being used during translation of CDS1 of the *tra* gene.

10. Let’s move on to looking at the reading frames for exon 2 of tra-RA. Zoom in to view only the second exon of the *tra* gene by jumping to “contig1:10,120-10,570” using the “enter position or search terms” text box. Remember that both the RNA-Seq data and the cDNA data have been used to map the positions of exons.

**Q1.** First examine reading frame 1. Are there any stop codons in the reported exon? ______ If there are early stop codons, do you think this is the reading frame used during translation? ________________

**Q2.** Examine reading frame 2. Are there any stop codons in this reading frame within the exon? ______

**Q3.** Examine reading frame 3. Are there any stop codons in the reported exon? ______

**Q4.** Using the evidence above, which reading frame maintains an Open Reading Frame (ORF) across exon 2 of tra-RA? ________________ Is this the same reading frame as that used for exon 1? ______

11. Finally, take a look at exon three (contig1:10,600-10,850). We anticipate that since this is the last CDS there will be one or more stop codons, the first of which will mark the site of translation termination. The 3’ UTR (3’ untranslated region) extends downstream from this point to the site of poly(A) addition where the last exon ends (see Module 3). Here all three reading frames have an ORF followed by one or more stop codons in the exon.

How can we figure out which reading frame is correct? We will investigate this in the next section by looking specifically at the splice junction.
Investigation 2

We can combine what we know about reading frames with what we know about splicing to learn exactly how tra-RA is put together. We’ll note where the start codon, splice sites, and stop codon are so we can construct a gene model. Then, in Module 6, we’ll use these same types of information to solve some mysteries about tra-RB.

1. Using the same Genome Browser page, reset the Browser by clicking on “hide all.” Then open the tracks that will provide the information we want for Investigation 2.

   - **Base Position “full”**
     - Note that you will not be able to see the DNA sequence or amino acid tracks until you zoom in.

   - **FlyBase Genes “pack”**

   - **RNA-Seq Coverage “full”**
     - You will see blue and red histograms representing the RNA-Seq data (indicating the amount of mRNA synthesized) in females and males, respectively. We will focus on the blue histogram (Adult Females) again.

   - **Exon Junctions “full”**
     - These rectangular boxes joined by a thin black line will help us identify the exon-intron boundaries.

2. As we did in Module 3, let’s customize the RNA-Seq track by setting the “Data view scaling” field to “use vertical viewing range setting” and the “max” “Vertical viewing range” to 37. (Remember that you gain access to these settings by clicking on the “RNA-Seq Coverage” label under the “RNA Seq Tracks” green bar found in the bottom section of the page.) Under the “List sub-tracks” section, unselect the “Adult Males” track.

3. Let’s find the start codon for tra-RA. Zoom in on where the FlyBase Genes track shows that the translation starts (where the tracked black box gets thicker for the tra-RA isoform) as seen in Figure 3.
Q5. Give the coordinates for the entire start codon for tra-RA (start codon coordinates should be three consecutive numbers, for example: nucleotides 212-214). _____________________

Q6. Which reading frame should we follow along to see the predicted amino acid sequence of tra-RA?  
_________________

Q7. Zoom out to see the entire exon. Are there any stop codons in this reading frame in the first exon?  
_________________

4. Now zoom in and find the last base of the first exon for tra-RA using your RNA-Seq data and Exon Junctions data (Figure 4).

Q8. Give the coordinate for the very last base of the first exon.  ___________________

So that we can follow the polypeptide through until we identify the stop codon, we need to figure out which reading frame we should follow in the second exon. This is not as easy as you might think, because in eukaryotes, different exons of the same gene may utilize different reading frames. You saw
an example of this previously in Module 1. Sometimes we can infer the correct reading frame given the pattern of start and stop codons within the region of the exon, identified by RNA-Seq data. But that sort of information does not always give a definitive answer — there may be more than one possible reading frame for a given exon. To figure out which reading frame is being translated at exon 2, we need to check the end of the first exon to see how many bases of the last codon are present before the 5’ splice site consensus sequence. To do this, look closely at reading frame 3, just before the splice site (Figure 5).

![Figure 5](image.png)

**Figure 5**  Region at the end of Exon 1 of the *tra* gene.

Note that the splice site (red box) cuts off the last codon of the first exon (blue box) after just one base. Therefore, we would say this exon has a “**phase 1** end because there is a partial codon at the end of the exon that is 1 base long.** (If there were a fully completed codon before the splice site, it would be in **phase 0**, and if there were two bases before the splice site, it would be in **phase 2**.)

For this exon with a phase 1 end, we will need two more bases from the next exon to complete the codon. Knowing this we can identify the reading frame that will be used in the second exon. Navigate to the 3’ splice site of intron 1 (i.e., the location where the first intron ends and the second exon begins; Figure 6). To review splicing and the concept of phase, watch the [Splicing and Phase video](#).

**Q9.** Based on the evidence you see in the browser, give the coordinate for the first base of the second exon of tra-RA. ______

**Q10.** Do you observe an appropriate splice acceptor site just upstream within the intron? ______

**Q11.** Knowing that exon 1 ends with a partial codon of 1 base, what reading frame is being used in the second exon? ____
Now we will be using reading frame 2, because, after the splice site, there are two bases left in the codon. These two bases plus the one base left from the first exon make a complete codon.

5. Next, zoom out and look at reading frame 2 for all of exon 2 of tra-RA. You can see that there are no stop codons in this reading frame, which lends support to our conclusion that this is the proper reading frame.

6. Now, let’s do the same for the 5’ splice site of intron 2 for tra-RA. Zoom in on that splice site (Figure 7).

Q12. Give the coordinate of the base prior to the 5’ splice site of intron 2. ________________

Q13. How many bases are left in the codon before the splice site (i.e., is this phase 0, phase 1, or phase 2)? ______________________________

7. Navigate to the start of the final exon (Figure 8).
**Q14.** Locate the 3′ splice site of Intron 2. Give the coordinate of the first base in exon 3 for tra-RA.
____________________

**Q15.** Which reading frame is being translated in the final exon? ________________

8. Now locate the first stop codon in the translated reading frame. Stop codons are shown as red boxes with asterisks (red arrows) as shown in Figure 9.

**Gene model for tra-RA:**
- Coordinates for start of translation: __________________
- Coordinate for last base of exon 1: __________________
- Coordinate for first base of exon 2: __________________
- Coordinate for last base of exon 2: __________________
- Coordinate for first base of exon 3: __________________
- Stop codon coordinates: _____________________________
Take the coordinate information above to draw a map of tra-RA using rectangles to represent exons and connecting lines to represent introns. Label the ends of the exons with the appropriate coordinates and indicate the transcription start site for the tra-RA initial transcript. Below this map, provide a map of the processed mRNA after intron removal. Below this map, indicate the regions that are translated into a protein. Give precise coordinates. Color coding may be helpful.

In Module 6, we will compare this model of tra-RA with a model of tra-RB.

**Q18.** To cement your knowledge of gene structure, you could construct a similar map of the *spd-2* gene. How many exons does this gene have? ____ How many introns? ____ How many isoforms? ____ Use the same approach to determine the coordinates for the exons, and the coordinates for the coding region (another name for the region that is translated).
Module 6: Alternative Splicing

Leocadia Paliulis

Objectives

• Demonstrate how alternative splicing of a gene can lead to different mRNAs.
• Show how alternative splicing can lead to the production of different polypeptides and result in drastic changes in phenotype.

Prerequisites

• Understanding Eukaryotic Genes Modules 1–5: https://thegep.org/ueg/

Class Instruction

• Introduce tra-RB
• Discuss differences between tra-RB and tra-RA. Reinforce concept of isoform.
• Investigation 1: How can there be different mRNAs encoded in the same gene?
• Investigation 2: Examine the tra polypeptides by looking at the three possible reading frames. Review concept of reading frame and introduce phase if not previously introduced. Students will construct a gene model for tra-RB, using sequence information, and RNA-Seq data as evidence.
• Discussion of gene models/wrap-up

Associated Videos

• Genes and Isoforms Video: https://youtu.be/ce6nVSiiV7I
• Splicing and Phase Video: https://youtu.be/9VWjyu3PUJ0
• RNA-Seq and TopHat Video: https://youtu.be/XD-egRcHYL4

Table of Contents

Investigation 1: ................................................................................................................................. 1

Investigation 2. What are the consequences of alternative splicing on the polypeptides produced from each isoform? ........................................................................................................ 6

Investigation 1:

In this investigation, we will focus on tra-RB, the second isoform of the tra gene, and will explore how multiple different mRNAs and polypeptides can be encoded by the same gene. The story of tra-RB is an exciting story of sex, alternative splicing, and poison exons!
1. To begin, open a web browser.
2. Go to the GEP UCSC Genome Browser Mirror site at https://gander.wustl.edu/ and follow the instructions given in Module 1 to open contig1 of *Drosophila melanogaster*, using the July 2014 (Gene) assembly rather than the Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) assembly. Once you are on the Genome Browser page, set “Base Position” (under the “Mapping and Sequencing Tracks” bar) to “full” so that you will be able see the three possible reading frames (remember that you will not see individual bases or amino acids until you zoom in, though). Also set “FlyBase Genes” (under the “Genes and Gene Predictions” bar) to “full”. Don’t forget to click on one of the “refresh” buttons to see your changes.
3. Enter the following coordinates into the “enter position or search terms” text box: contig1:9,700-11,000 and hit the “go” button to get a good view of the tra gene (Figure 1).

![UCSC Genome Browser on D. melanogaster July 2014 (Gene) Assembly (dm3gene)](image)

Figure 1  Center the browser on the tra gene.

4. Let’s consider what we know about tra-RA and learn more about tra-RB.

Q1. Given that exons are shown by the black boxes, and introns are shown by thin lines with arrowheads in the FlyBase Genes track, what does this tell us about the first intron of tra-RB compared to that of tra-RA?

____________________________________________________________________________

5. Now let’s look at the patterns of transcription. Scroll down to “RNA Seq Tracks,” click on the “RNA-Seq Coverage” link. Change the track display settings as we did in Module 2 (“Display mode” field on full; “Data view scaling” field set to “use vertical viewing range setting”; “max” field under “Vertical viewing range” to “37”), but this time check both “Adult Females” and “Adult Males” under “List subtracks.” Hit the “Submit” button (Figure 2). Back on the browser main page, Under “RNA seq tracks” → “Exon Junctions” use the dropdown menu to select full, then hit refresh. To review the use of RNA Seq data, watch the RNA-Seq and TopHat video.
Now we can see the RNA-Seq data for males (red) and females (blue). Recall that peaks in RNA-Seq Read Coverage tracks usually correspond to the regions of the genome that are being transcribed. These two samples generally show similar RNA-Seq read coverage along the entire span of the *tra* gene. However, the adult female sample shows substantially lower RNA-Seq read coverage at around 9,971-10,145 (red box in Figure 3). We can also see the RNA-Seq Exon Junctions track, which shows the location of splice sites supported by the RNA-Seq data (as you saw in Module 4). Recall that the black
boxes in the FlyBase Genes track are exons and the thin lines with arrowheads show the locations of the introns. Notice that the diagrams for the first and second RNA-Seq Exon Junctions tracks have the same 5’ splice site but different 3’ splice sites. Let’s see what we can find out about these splice sites. First, we need to establish the reading frame for the first exon. Zoom in on the 5’ end of the transcript around position 9850-9860.

**Q2.** Given what you know about the initiation of translation, which of the 3 possible reading frames is used for both the tra-RA and tra-RB products? ____________________

Now zoom in on the location of the 5’ splice site at the end of the first exon in both tra-RA and tra-RB (Figure 4). We will also be thinking about the concept of phase here. To review splicing and phase, watch the [Splicing and Phase video](#).

**Q3.** Give the coordinate for the last base of the first exon for tra-RA_______

**Q4.** Give the coordinate for the last base of the first exon for tra-RB_______

**Q5.** What is the consensus sequence for the 5’ splice site (donor site)? ________________

**Q6.** What are the coordinates for the 5’ splice site in tra-RA? ________________

**Q7.** What are the coordinates for the 5’ splice site in tra-RB? ________________

**Q8.** What is the phase at this splice site? ___________

6. Now zoom out and zoom in on the start of the second exon in tra-RB, just after the 3’ splice site. We can identify the second exon by the RNA-Seq data, in particular using the RNA-Seq Exon Junctions data (Figure 5).
Q9. What are the coordinates for the first base of the second exon in tra-RB? __________
Q10. What is the consensus sequence for the 3’ splice site? ________________
Q11. What are the coordinates for the 3’ splice site in intron 1 of tra-RB? ____________
Q12. What phase do we anticipate? ______
Q13. Given this, what is the reading frame for tra-RB exon2? _____
Q14. Does this make sense, given the location of stop codons? ____________

7. Now zoom out and zoom in on the 3’ splice site for tra-RA. This can be identified from the RNA-Seq data, particularly the RNA-Seq Exon Junctions (Figure 6).

Q15. What are the coordinates for the first base of the second exon in tra-RA? __________
Q16. What is the consensus sequence for the 3’ splice site? ________________
Q17. What are the coordinates for that sequence in intron 1 of tra-RA? ______________
Q18. Given the phase at the donor site, what phase are we looking for here? ________
Q19. Given this, what is the reading frame for tra-RA exon 2? ________
Q20. Does this make sense, given the location of stop codons? 

The 3’ acceptor site for the second intron in tra-RA is found inside the second exon of tra-RB. This intron is *alternatively spliced*. Alternative splicing is one way eukaryotes produce different proteins from the same coding regions of DNA. Here the alternative decision is made in a sex-specific manner; male fruit flies have targeted the spliceosome to use the first 3’ acceptor site identified by the RNA-Seq Exon Junction data, while female fruit flies have targeted the spliceosome to use the second 3’ acceptor site identified. This change in splicing has profound effects – in fact, it drives the programming of male and female characteristics in the developing fly. To review alternative splicing, watch the [Genes and Isoforms video](#).

8. Reset your browser by entering “contig1:9,700-11,000” into the “enter position or search terms” text box and hit the “go” button. Let’s analyze the consequences of this alternative splicing on production of a protein product.

Q21. Look back at Module 5, Q17. Using the coordinates you’ve written in your answer to that question, calculate the number of amino acids in the protein translated from tra-RA. Do this by calculating the size of each exon, added the sizes of all exons together, and dividing by 3 (the number of bases in a codon).

Now look at the tra-RB isoform:

Q22. Write down the coordinates for exon 1.

Q23. Looking at exon 2 of tra-RB, is the entire second exon translated or is there a stop codon within the exon?

Q24. Write down the coordinates for the translated portion of exon 2.

Q25. How many amino acids does the protein translated from the tra-RB isoform have?

Q26. Is it likely that the protein translated from tra-RB could play the same functional role played by the protein translated from tra-RA?

The Tra protein has an important function in female *Drosophila* and is itself a splicing factor that regulates splicing. Careful annotation of genes, as we have done here, can provide many insights into biological control mechanisms.

**Investigation 2. What are the consequences of alternative splicing on the polypeptides produced from each isoform?**

Now that we know that *tra* is alternatively spliced to make two isoforms, tra-RA and tra-RB, and that males express one isoform while females express the other, let’s try to figure out how alternative splicing affects the polypeptides produced from translating these mRNAs. To do this, we need to produce a gene model for tra-RB and compare it to the gene model for tra-RA that you constructed in Module 5, showing where the start codons and stop codons appear in each isoform.
Use what you learned in Module 5 to construct a gene model for tra-RB. Locate the start codon, splice sites, and the stop codon. Construct the gene model below.

Q27. Coordinate for start of translation: ______________________
Q28. Coordinate for last base of exon 1: ______________________
Q29. Coordinate for first base of exon 2: ______________________
Q30. Coordinate for last base of exon 2: ______________________
Q31. Coordinate for first base of exon 3: ______________________
Q32. Stop codon coordinates: ______________________________

Points for discussion:

How does the polypeptide translated from the tra-RB isoform differ from the polypeptide translated from the tra-RA isoform? What are the consequences of these differences on protein function?

Discuss how the bigger mRNA leads to creation of a smaller polypeptide!!

Consider how alternative splicing could allow many different proteins to be encoded by the same gene.

Based on the gene structure of the two isoforms of tra shown in the "FlyBase Genes" track, provide a hypothesis that could explain this difference in RNA-Seq read coverage between the adult males sample and adult females sample.
Annotation of a *Drosophila* Gene

Wilson Leung

**Prerequisites**
- Lecture: *Annotation of Drosophila*
- Lecture: RNA-Seq Primer
- BLAST Walkthrough: *An Introduction to NCBI BLAST*

**Resources**
- FlyBase: [https://flybase.org](https://flybase.org)
- The GEP UCSC Genome Browser, Gene Record Finder, and the Gene Model Checker are available under the “Resources & Tools” section of the F Element project page ([https://thegep.org/felement/](https://thegep.org/felement/))

**Introduction**
The overall GEP gene annotation strategy is discussed in the “*Annotation of Drosophila*” presentation. This walkthrough illustrates how you can apply the GEP annotation strategy to construct a gene model on the *contig10* project from the *Drosophila biarmipes* Muller F element [Aug. 2013 (GEP/Dot) assembly]. In this walkthrough, we will discuss the strategies for identifying the putative ortholog using BLAST, determining the gene structure using the Gene Record Finder and FlyBase, mapping the individual coding exons using the “Align two sequences” functionality in BLAST, and verifying the final gene model using the Gene Model Checker. Once we have verified the gene model, we will complete the “Gene Report Form” section of the “F Element Project Annotation Report” in preparation for project submission.

**Examine the project region**
Open a web browser and navigate to the GEP UCSC Genome Browser ([https://gander.wustl.edu](https://gander.wustl.edu)). Click on the “Genome Browser” link in the left sidebar to access the Genome Browser Gateway page. Enter “*D. biarmipes*” into the “Enter species or common name” field. Select “Aug. 2013 (GEP/Dot)” under the “*D. biarmipes Assembly*” field, and enter “*contig10*” under the “Position/Search Term” field. Click on the “Go” button (Figure 1).

Because the Genome Browser remembers your previous track display settings, we will hide all the evidence tracks and then turn on the subset of evidence tracks that are required for this walkthrough. Scroll down to the list of buttons below the Genome Browser image and then click on the “hide all” button to hide all the evidence tracks (Figure 2).
Use the drop-down boxes in the track configuration sections to change the display options for the following evidence tracks and then click on the “refresh” button (Figure 3):

- Under “Mapping and Sequencing Tracks”
  - Base Position: **full**

- Under “Genes and Gene Prediction Tracks”
  - D. mel Proteins: **pack**
  - Genscan Genes: **pack**
  - N-SCAN: **pack**

For this walkthrough, we will only evaluate the results from two gene predictors (i.e. Genscan and N-SCAN). For your own annotation projects, you should evaluate the predictions from all the gene predictors available through the GEP UCSC Genome Browser.
Interpreting the D. mel Proteins track
The “D. mel Proteins” track shows the results of the BLASTX alignments of D. melanogaster annotated proteins against the D. biarmipes contig10 genomic sequence. This track shows that contig10 contains three regions with sequence similarity to D. melanogaster genes (i.e. CG33978, CG31997, and Arl4). Each rectangle in the “D. mel Proteins” track corresponds to a region of sequence similarity between the D. melanogaster protein (subject) and the translated genomic sequence of the D. biarmipes contig (query). A line connects multiple BLASTX matches to the same D. melanogaster protein and the direction of the arrows denotes the orientation of the match. Because the “D. mel Proteins” track simply demarcates regions with significant sequence similarity, BLASTX could have merged multiple exons into a single alignment block, missed a weakly conserved exon, or break a large exon into multiple alignment blocks. Consequently, you should not use this track to infer the gene structure (i.e. number or placement of exons) of the putative ortholog.

Each protein-coding gene annotated by FlyBase has an annotation symbol that begins with the prefix “CG” (i.e. Computed Gene). FlyBase also assigns a different gene symbol to genes that have been characterized experimentally. (The gene symbol for a gene that has not been characterized experimentally is the same as its annotation symbol.) For example, the gene symbol for ADP ribosylation factor-like 4 is Arl4 and its annotation symbol is CG2219. All the GEP annotation resources and tools refer to D. melanogaster genes using their FlyBase gene symbols.

The features on the “D. mel Proteins” track are shown in different colors. To better understand the color scheme, click on the “D. mel Proteins” link in the track configuration section (Figure 4).

Figure 4  Click on the “D. mel Proteins” link in the track configuration section to learn more about this evidence track.

We can use the controls at the top section of the “D. mel Proteins Track Settings” page to change the display mode and filter the BLASTX matches by score (Figure 5). The “Description” section mentions that the color of each feature corresponds to its bit score (i.e. the statistical significance of the match). It also includes a table that shows the range of bit scores for each color (warmer color denotes a more significant match). For example, features with bit scores greater than 500 are in red and features that have bit scores between 200 and 500 are in brown. The “Methods” section shows the list of custom BLASTX parameters that were used to produce this evidence track.

Click on the “Submit” button to return to the Genome Browser view of contig10. Based on the color scheme of the “D. mel Proteins” track, we know that the red matches to the three isoforms of G33978 and the A and B isoforms of Arl4 are more statistically significant than the brown matches to the A and B isoforms of CG31997 and the C isoform of Arl4.

Interpreting the gene prediction tracks
Examination of the Genscan and N-SCAN gene prediction tracks show that the two gene predictors disagree on the features that are found in the first 30kb of contig10 (Figure 6). Genscan predicted a single gene that spans the first 30kb of the contig while N-SCAN predicted a smaller gene that spans from 25–30kb.

The “D. mel Proteins” track indicates that the first 10kb of contig10 has significant similarity to CG33978 while the region between 25–30kb has significant sequence similarity to CG31997. Hence the “D. mel Proteins” track is in concordance with the N-SCAN prediction contig10.001.1 (blue arrow in Figure 6) and it suggests that Genscan might have merged two adjacent genes into a
single feature (contig10.1, red arrow in Figure 6). Consequently, our subsequent analysis will be based on the N-SCAN prediction contig10.001.1 instead of the Genscan prediction.

This walkthrough will only investigate the feature at 25–30kb. For some GEP annotation projects, you would need to investigate all three regions with BLASTX alignments and gene predictions.

**Identify the ortholog**
The first step in our investigation of the N-SCAN gene prediction contig10.001.1 is to identify the putative *D. melanogaster* ortholog. While the “D. mel Proteins” track provide us with an overview of the interesting features within contig10, **we should not rely on this track to assign the ortholog.** The “D. mel Proteins” track shows the regions of the contig with significant sequence similarity to *D. melanogaster* proteins but it does not necessarily mean that the contig region contains the best match to the *D. melanogaster* protein. In addition, multiple genes could appear at the same region of the contig (e.g., because they contain the same conserved domains). Because the rest of the gene annotation is predicated on the identification of the correct ortholog, we should always perform a BLASTP search of the feature of interests against the collection of *D. melanogaster* annotated proteins in order to identify the putative *D. melanogaster* ortholog.

We will search the predicted protein sequence against the collection of annotated proteins sequences in *D. melanogaster* using the FlyBase BLAST service. Click on the N-SCAN prediction contig10.001.1 on the Genome Browser and then click on the “Predicted Protein” link to retrieve the protein sequence (Figure 7). Select the protein sequence and copy it onto the clipboard.

![Figure 7](image.png)  
**Figure 7** Click on the “Predicted Protein” link to retrieve the protein sequence for the N-SCAN prediction contig10.001.1.
Open a new tab and then navigate to FlyBase (https://flybase.org). Click on the “BLAST” button on the upper left of the main page and then paste the predicted protein sequence into the “Sequence” field. Change the “Database” field to “Annotated proteins (AA)” and verify that the “Program” field is set to “blastp: AA → AA”. Under the “Species” section, verify that the checkbox next to “Drosophila melanogaster” is selected (Figure 8). Click on the “BLAST” button to run the BLASTP search.

Figure 8 Perform a BLASTP search against the collection of D. melanogaster annotated proteins.

Examination of the BLAST Hit Summary table shows that there are two matches (to the B and A isoforms of CG31997) that are much more statistically significant (i.e. have lower E-values) than the rest of the matches (E-values = 3.28622e-66 versus E-values > 1, Figure 9).

Figure 9 The first two BLASTP matches to D. melanogaster proteins have much lower E-values than the rest of the matches.

Scroll down to the alignment section so that we can examine the BLASTP alignment to the B isoform of CG31997 (i.e. CG31997-PB). The first line of the alignment output contains the metadata that are associated with the D. melanogaster protein. For example, the “loc” field denotes the genomic location of this protein in the D. melanogaster assembly (Figure 10). The value before the colon corresponds to the chromosome where the protein is found (i.e. chromosome 4, also known as the Muller F element). Because most genes remain on the same Muller element across
the different *Drosophila* species, the fact that CG31997 is located on the Muller F element in *D. melanogaster* lends further credence to the hypothesis that the contig10 region surrounding the NSCAN gene prediction contig10.001.1 contains the putative ortholog of CG31997.

![Figure 10](image)

Because the number of chromosomes changes across the different *Drosophila* species, Hermann Müller developed a nomenclature (A–F) to describe the chromosome arms that are found in most *Drosophila* species (*Muller, 1940*). For example, the Muller F element is the 4th chromosome in *D. melanogaster* but the 6th chromosome in *D. mojavensis*. See the Synteny Table page at FlyBase for additional information ([https://flybase.org/maps/synteny](https://flybase.org/maps/synteny)).

Examination of the BLASTP alignment for CG31997-PB shows that the end of the predicted protein has high levels of sequence similarity with this *D. melanogaster* protein. However, the beginning of the alignment shows lower sequence similarity with the *D. melanogaster* protein and the first three amino acids of the NSCAN prediction are missing from the alignment (Figure 11). Hence we need to investigate this region further when we construct the *D. biarmipes* gene model.

![Figure 11](image)

Collectively, the BLASTP search results indicate that the region surrounding the NSCAN contig10.001.1 prediction in *D. biarmipes* contig10 likely contains a putative ortholog of the *D. melanogaster* gene CG31997. However, because of the high error rates associated with gene predictions, we need to perform additional analysis on the entire genomic region surrounding the gene prediction in order to construct the gene model.
Determine the gene structure
Before we can construct the orthologous gene model, we need to ascertain the gene structure (e.g., the number of isoforms and exons) of the *D. melanogaster* CG31997 gene using the Gene Record Finder. Open a new tab and navigate to the F Element project page on the GEP web site (https://thegep.org/felement/). Click on the “Gene Record Finder” link under the “Resources & Tools” section. Enter “CG31997” into the text box and then click on the “Find Record” button (Figure 12).

Gene symbols in *Drosophila* are case-sensitive (for example, *Dl* and *dl* correspond to two different genes in *D. melanogaster*). By convention, the gene symbol begins with a lowercase letter if the mutant phenotype is first characterized by a recessive allele. The gene symbol begins with an uppercase letter for the wild-type phenotype or if the mutant phenotype is first characterized by a dominant allele. See sections 1.2.2 and 1.2.3 of the genetic nomenclature page at FlyBase for additional details (https://wiki.flybase.org/wiki/FlyBase:Nomenclature#1.2.2).

The Gene Record Finder shows that CG31997 has two isoforms (A and B) in *D. melanogaster*. The CDS usage map (under the “Polypeptide Details” tab) shows that both isoforms have the same set of coding exons (i.e. 1_10739_0, 2_10739_2, and 3_10739_1). (The coding exons are ordered from 5’ to 3’ from left to right in the CDS usage map.) Hence the differences between these two isoforms are limited to the untranslated regions (UTRs).

To visualize the differences between these two isoforms using FlyBase GBrowse, click on the “View in GBrowse” link under the “Graphical Viewer” column in the “Gene Details” section (Figure 13). (A graphical overview of the two isoforms is also available under the “mRNA Details” section.)
Figure 13  The **Gene Record Finder** shows that **CG31997** has two isoforms (A and B) in *D. melanogaster*. Click on the “View in GBrowse” link to view the gene models using FlyBase GBrowse.

Consistent with the **Gene Record Finder** record, the GBrowse “Transcript” track shows that the coding exons (orange boxes) are the same in the A and B isoforms of CG31997. The difference between the two isoforms is found in the UTRs (grey boxes). Specifically, the 3’ UTR of the B isoform overlaps with but is longer than the 3’ UTR of the A isoform (Figure 14).

In *Drosophila*, the suffix following the gene name corresponds to the name of the isoform. The “R” suffix corresponds to a RNA record whereas the “P” suffix corresponds to the protein record. This means that CG31997-RB is the mRNA record and CG31997-PB is the protein record for the B isoform of CG31997.

Based on parsimony (i.e. minimizing the number of changes compared to *D. melanogaster*), we expect to find both the A and B isoforms of CG31997 in our *D. biarmipes* contig10 sequence. For
this walkthrough, we will only focus on the annotation of the coding exons. Consequently, we only need to determine the coordinates of the three coding exons (CDS) for one of the isoforms (e.g., isoform B) because the set of coding exons for both the A and B isoforms are the same.

**Determine the approximate location of the coding exons**

The next step in our analysis is to determine the exact coordinates of each coding exon of CG31997-PB in our contig10 project. Because the BLAST algorithm does not take the positions of potential splice sites into account when it generates the alignment, BLAST often extends the alignment beyond the coding exon boundary and into the intron. To ameliorate this issue, the GEP annotation protocol recommends mapping each coding exon separately to determine their approximate locations and then further refine the exon boundaries by searching for compatible splice donor and acceptor sites by visual inspection using the GEP UCSC Genome Browser.

In addition to comparing a query sequence against a collection of subject sequences in a database (e.g., nr, RefSeq), the NCBI BLAST web service also allows us to compare two or more sequences against each other (using the program bl2seq). In order to map the amino acid sequences of each CDS against the contig10 sequence, we must translate the entire contig10 sequence in all six reading frames (i.e. three reading frames in the plus and minus strands, respectively) and then compare each conceptual translation against the CDS sequence. This means that we can use either TBLASTN or BLASTX to perform this search (depending on whether we treat the CDS sequence as the query or the subject sequence, respectively). In this walkthrough, we will perform BLASTX searches using the contig10 genomic sequence as the query and each CDS sequence as the subject.

To setup this BLASTX search, open a new tab and navigate to the NCBI BLAST web site (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Click on the “blastx” image under the “Web BLAST” section (Figure 15) and then select the “Align two or more sequences” checkbox.

![Figure 15](https://example.com/image.png)  
Figure 15  Click on the “blastx” image under the “Web BLAST” section to access the NCBI BLASTX service.

In order to perform the BLASTX search, we need to obtain the genomic sequence for contig10. For this walkthrough, the sequence file is available in the exercise package (contig10.fasta). Click on the “Browse...” or the “Choose File” button in the “Enter Query Sequence” section and select the contig10.fasta file. Alternatively, you can obtain the genomic sequence using the “DNA” link.
(under the “View” menu) on the GEP UCSC Genome Browser (Figure 16), and then copy and paste
the sequence into the “Enter Query Sequence” textbox.

For the GEP annotation projects, the contig sequence file is available in the “src”
folder of the annotation package.

![UCSC Genome Browser](image)

Figure 16 Using the “DNA” link (under the “View” menu) of the GEP UCSC Genome Browser to retrieve the contig10
genomic sequence.

For the subject sequence, we can use the Gene Record Finder to retrieve the amino acid sequence for
each CDS. Instead of searching each CDS from 5’ to 3’, we recommend searching for sequence
similarity using the larger CDS’s first to anchor the gene model. (Many genes in D. melanogaster
have small initial CDS.) The size of each CDS is listed in the “Size (aa)” column of the CDS
sequence table. In this case, the CDS 3_10739_1 is the largest CDS (with 62aa) among the three
CDS’s in CG31997.

Select the “Gene Record Finder” tab in your web browser and click on the row with the FlyBase ID
“3_10739_1” in the CDS sequence table, select the sequence in the “Sequence viewer” panel
(including the header which begins with a > sign) and copy it onto the clipboard (Figure 17).
Select the NCBI BLAST web browser tab and paste the sequence for CDS 3_10739_1 into the “Enter Subject Sequence” text box (Figure 18).

The default NCBI BLAST parameters are optimized for searching the query sequence against a large collection of sequences in a database. When we are using BLAST to compare only two sequences against each other, we need to change some of these alignment parameters because the default parameters could potentially mask the conserved regions of the coding exon.
Click on the “Algorithm parameters” link to expand this section. Change the “Compositional adjustments” field to “No adjustment” and uncheck the “Low complexity regions” filter under the “Filters and Masking” section. To reduce the number of spurious matches, we will also change the “Expect threshold” to “1e-2” under the “General Parameters” section. Because NCBI BLAST uses different word sizes for database versus bl2seq searches, we should also verify that the “Word size” parameter is set to “3” (Figure 19). Click on the “BLAST” button to run the BLASTX search.

![Figure 19](image)

**Figure 19** Customize the BLASTX search parameters in the “Algorithm parameters” section for bl2seq searches.

Because the E-value of a BLAST hit depends on the length of the alignment, you may need to increase the “Expect threshold” in order to detect sequence similarity to short CDS's. (Shorter alignments have higher E-values because they are more likely to occur by chance.)

By default, NCBI BLASTP, BLASTX, and TBLASTN use a word size of 6 for database searches and a word size of 3 for bl2seq searches. The larger word size improves the performance of BLAST but it might miss matches to short protein sequences with weak sequence similarity. Hence we should verify that the word size parameter (under the “Algorithm parameters” section) is set to 3 when we use BLAST to compare a CDS sequence against the contig sequence.

The BLASTX results show only a single match (with E-value 6e-39) to the CDS 3_10739_1 (Figure 20). Click on the “Alignments” tab to view the corresponding BLASTX alignment. The “Subject” coordinates show that the alignment covers all 62aa of the CDS. The “Query” coordinates correspond to the region within contig10 (i.e. 27,286–27,471) that shows sequence similarity to CDS 3_10739_1 when it is translated in the first reading frame in the positive strand (i.e. frame +1). Hence we can place CDS 3_10739_1 at 27,286–27,471 on contig10.
We can apply the same procedure to place the other two CDS's on contig10. The BLASTX search of the next largest CDS (1_10739_0) shows only a partial alignment (at 25,685–25,834 in frame +2) with an E-value of $7e^{-12}$ (Figure 21). The first two amino acids of the CDS is missing from the alignment and the beginning of the CDS shows lower sequence similarity to contig10 than the end of the CDS. However, this is the only match with an E-value that is less than $1e^{-2}$ within contig10.

The BLASTX search of CDS 2_10739_2 against contig10 placed this CDS at 27,081–27,194 in frame +3. However, the last amino acid of this CDS is missing from the alignment (Figure 22).
The results of the BLASTX exon-by-exon searches are summarized in the table below:

<table>
<thead>
<tr>
<th>FlyBase ID</th>
<th>CDS Size</th>
<th>Query Range</th>
<th>Query Frame</th>
<th>Subject Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_10739_0</td>
<td>46</td>
<td>25685-25834</td>
<td>+2</td>
<td>3-46</td>
</tr>
<tr>
<td>2_10739_2</td>
<td>39</td>
<td>27081-27194</td>
<td>+3</td>
<td>1-38</td>
</tr>
<tr>
<td>3_10739_1</td>
<td>62</td>
<td>27286-27471</td>
<td>+1</td>
<td>1-62</td>
</tr>
</tbody>
</table>

Examination of the query ranges for the BLASTX alignments of the three CDS of CG31997 shows that they are collinear: all the CDS's are placed on the positive strand and the query ranges for the CDS's are in ascending order. Consequently, despite the amino acids that are missing from the BLASTX alignments, the exon-by-exon search results support the hypothesis that the putative ortholog of CG31997 is located at 25–28kb of the D. biarmipes contig10.

Using RNA-Seq data to verify the placement of the initial coding exon
While the best BLASTX alignment placed the initial CDS 1_10739_0 at 25,685-25,834 in contig10, the start codon of this initial CDS is missing from the alignment. In addition, the alignment to the beginning of the CDS shows much weaker sequence similarity to contig10 than the end of the CDS (Figure 21). Consequently, we would have more confidence in the annotation of this CDS if it were supported by expression data.

As part of the modENCODE project, the Baylor College of Medicine Human Genome Sequencing Center (BCM–HGSC) has produced RNA-Seq data for D. biarmipes using the adult males, adult females, and mixed embryos samples. These RNA-Seq reads (100–125bp in length) are derived primarily from processed mRNA (i.e. after the introns have been removed). Hence genomic regions with RNA-Seq read coverage usually correspond to transcribed exons (i.e. include both the translated and untranslated regions).

See the “RNA-Seq Primer” for an overview on the different types of RNA-Seq data available on the GEP UCSC Genome Browser.

To visualize the RNA-Seq data for D. biarmipes, go back to the web browser tab with the GEP UCSC Genome Browser, change the display mode for the “RNA-Seq Alignment Summary...” track to “show” and then click on the “refresh” button. The three new evidence tracks that appear on the Genome Browser correspond to the three samples where RNA-Seq data are available (i.e. mixed embryos, adult females, and adult males). The height of the histograms within each track corresponds to the number of RNA-Seq reads that have been mapped to each position of the D. biarmipes contig10 sequence. Hence the RNA-Seq summary track shows that CG31997 has the highest level of expression among the three features in contig10 (Figure 23).
Examine the RNA-Seq read coverage using the RNA-Seq Alignment Summary track.

To learn more about the display conventions of the RNA-Seq Alignment Summary tracks, click on the “RNA-Seq Alignment Summary” link under the “RNA Seq Tracks” section and then click on the “Mixed Embryos RNA-Seq Summary” link. Under the “List subtracks” section, we find that the green, red, orange, and blue colors correspond to the read depth of each nucleotide (A, T, G, and C, respectively) while the grey color corresponds to the number of reads that have high mapping quality at each genomic position (Figure 24). We can use the controls on the track settings page to alter the display settings of the Alignment Summary Tracks.
Regions with high read depth but low mapping quality (bright bases in the Alignment Summary track) are often caused by differences between the RNA-Seq reads and the contig sequence. These discrepancies could be caused by errors in the contig sequence. See the “Sequence Updater User Guide” for additional details on how to identify and report errors in the contig sequence.
To ascertain if the BLASTX alignment for CDS 1_10739_0 at 25,685–25,834 is supported by the RNA-Seq data and to determine the location of the start codon, enter “contig10:25685-25834” into the “enter position or search terms” text box and then click on the “go” button. Zoom out 3x so that we can examine the region surrounding the BLASTX alignment block.

The RNA-Seq Alignment Summary tracks for all three samples show high RNA-Seq read depth within and upstream of the BLASTX alignment block, consistent with the hypothesis that this region is being transcribed in *D. biarmipes*. The BLASTX alignment for CDS 1_10739_0 begins at 25,685 in frame +2. Examination of the “Base Position” track in the Genome Browser shows that there is only a single start codon (green rectangle) upstream of 25,685 (at 25,673±25,675) in frame +2 before the first stop codon (red rectangle) at 25,640–25,642 (Figure 25).

![BLASTX Alignment to D. melanogaster Proteins](image)

Figure 25  There is only one start codon (green rectangle) in frame +2 upstream of the start of the BLASTX alignment at 25,685 before the first stop codon.

Using this start codon at 25,673–25,675 in our gene model will increase the size of this CDS compared to the orthologous CDS in *D. melanogaster*. However, this is the only start codon available that would allow us to preserve both isoforms of CG31997 in *D. biarmipes*. The expansion of this CDS 1_10739_0 relative to the *D. melanogaster* is also supported by the available RNA-Seq data and the N-SCAN gene prediction. Consistent with the *D. melanogaster* gene model (Figure 14), the RNA-Seq coverage upstream of the start codon likely corresponds to the 5' UTR.

See the “Browser-Based Annotation and RNA-Seq Data” exercise for a more detailed discussion of the challenges associated with interpreting RNA-Seq data.

While regions with RNA-Seq read coverage is a strong indicator of transcription, we should treat the lack of RNA-Seq coverage as a negative result. For example, a transcript might only be expressed at low levels or at a tissue or developmental time point that have not been sampled by RNA-Seq. Hence we cannot make any inferences based on the lack of RNA-Seq coverage.
Identifying splice sites

Using the combination of the exon-by-exon BLASTX alignments and the RNA-Seq data, we can define the span of the coding region for the CG31997-PB ortholog in D. biarmipes contig10 (i.e. 25,673–27,471). The next step of our annotation is to identify the exon boundaries for the three CDS's in our gene model. For eukaryotic genes with multiple coding exons, introns in the pre-messenger RNA (pre-mRNA) are usually removed by the spliceosome prior to the translation of the mature mRNA into a protein product. (In some cases, introns in the pre-mRNA can also be excised by self-splicing introns that form a ribozyme.) The 5’ end of the intron (i.e. splice donor site) usually has the sequence **GT** (GU in the pre-mRNA) while the 3’ end of the intron (i.e. splice acceptor site) usually has the sequence **AG** (Figure 26).

![Figure 26](image)

In D. melanogaster, approximately 99% of the introns have a **GT** splice donor site and 1% of the introns have a **GC** non-canonical splice donor site. Almost all of the introns have an **AG** splice acceptor site. U12-type introns have an AT splice donor site and an AC splice acceptor site but they are rare in D. melanogaster (found in less than 1% of all unique introns). The GEP comparative annotation protocol posits that all introns have a GT splice donor site and an AG splice acceptor site unless the D. melanogaster gene model uses a non-canonical splice site or the non-canonical splice site is supported by RNA-Seq data.

Determine the phases of the donor and acceptor splice sites

During splicing, introns (which usually begins with a GT and ends with an AG) are removed from the pre-mRNA so that adjacent exons are placed next to each other. This means that the ends of an exon do not necessarily correspond to the ends of the complete codon. The number of nucleotides between the last complete codon and the splice donor site is known as the **phase** of the splice donor site. Similarly, the number of nucleotides between the splice acceptor site and the first complete codon is known as the phase of splice acceptor site. Because the phases of the splice sites depend on the placement of the complete codon, the phases of the donor and acceptor sites are predicated on the reading frame of each CDS.
In addition, in order to maintain the open reading frame across adjacent CDS's, the phases of the donor and acceptor sites of adjacent CDS's must be compatible with each other. Specifically, the sum of the donor and acceptor phases of adjacent CDS's must either be 0 (i.e. no additional codon) or 3 (i.e. a complete codon). The use of incompatible splice donor and acceptor sites will introduce a frame shift into the translation of the CDS following the splice acceptor site.

Because the BLASTX alignment for the initial CDS (1_10739_0) of CG31997-PB terminates at 25,834 and the alignment includes the last amino acid of this CDS (Figure 21), we expect to find the splice donor site for this CDS at around position 25,834 of contig10.

To examine this region more closely in the Genome Browser, enter “contig10:25834” into the “enter position or search terms” text box and then click on the “go” button. Zoom out 10x and then zoom out another 3x to examine the 30bp surrounding this position. The GT splice donor site closest to 25,834 is located at 25,836-25,837. This splice donor site is in phase 2 relative to frame +1, in phase 1 relative to frame +2, and in phase 0 relative to frame +3 (Figure 27).

Figure 27  The phase of the splice donor site depends on the reading frame.
However, because the reading frame of this CDS is dictated by the BLASTX alignment (i.e. frame +2, Figure 21), the splice donor site at 25,836–25,837 is in phase 1. This means that the splice acceptor site of the adjacent CDS must be in phase 2 in order to maintain the open reading frame.

The GEP annotation strategy prefers the gene model that minimizes the number of changes compared to D. melanogaster (i.e. the most parsimonious gene model). Because the total change in the size of the coding region depends on the positions of both the splice donor and acceptor sites, we should identify the locations of alternate splice donor site candidates in the region surrounding the BLASTX alignments. We can then select the pair of compatible splice donor and acceptor sites that is best supported by the available expression data and computational predictions while also minimizing the change in the total size of the coding region compared to D. melanogaster.

In principle, we can search for potential splice donor site candidates up to the end of the first in-frame stop codon (i.e. 25,843). For CDS 1_10739_0, the closest phase 0 splice donor site relative to frame +2 is located at 25,826–25,827 while the closest splice donor site in phase 2 is located at 25840–25841 (Figure 28).

In order to ascertain which of the three potential splice donor sites is the best candidate, we need to determine the phase of the splice acceptor site for the next CDS. Because the BLASTX alignment shows that the first amino acid of the next CDS (2_10739_2) aligns to contig10 at 27,081-27,083 in frame +3 (Figure 22), we will examine this region more closely using the Genome Browser to determine the phase of the best acceptor site.

Enter “contig10:27081” into the “enter position or search terms” text box and then click on the “go” button. Zoom out 10x and then zoom out another 3x to examine the 30bp region surrounding this position. There is only a single splice acceptor site (at 27,077–27,078) within this region and this splice acceptor site is in phase 2 relative to frame +3 (Figure 29). Gene predictions from Genscan and N-SCAN as well as the RNA-Seq read coverage all support this potential splice acceptor site.
This phase 2 splice acceptor site is compatible with the phase 1 donor site at 25,836–25,837 that we have identified earlier for CDS 1_10739_0. The extra nucleotides (i.e. G + AT) near the splice sites will form an additional amino acid (D, Figure 30). Collectively, our analysis suggests that the CDS 1_10739_0 ends at 25,835 with a phase 1 splice donor site and CDS 2_10739_2 begins at 27,079 with a phase 2 splice acceptor site.
Verifying splice junctions using *TopHat* predictions

In addition to sequence similarity to the *D. melanogaster* CDS’s that have been detected by *BLASTX*, this splice junction is supported by Genscan, N-SCAN as well as the coverage of RNA-Seq reads in the mixed embryos, adult females, and adult males samples. We can gather additional evidence to support this splice junction using a RNA-Seq analysis tool called *TopHat*.

Because the RNA-Seq reads are derived primarily from processed mRNAs (where the introns have been removed), the subset of RNA-Seq reads that span multiple exons (i.e. spliced RNA-Seq reads) can provide us with additional evidence for a splice junction. When we map a spliced RNA-Seq read against the genome, part of the spliced read will map to one exon while the rest of the read will map to another exon. The region between these two alignment blocks would correspond to the intron. Splice site prediction tools such as *TopHat* and regtools can recognize this distinct mapping pattern of spliced RNA-Seq reads in order to infer the possible locations of the splice junction (Figure 31).

![Figure 31](image)

*Figure 31*  *TopHat* uses the spliced RNA-Seq reads to infer splice junctions.

The width of the boxes in the “RNA-Seq *TopHat*” track corresponds to the extent of the spliced RNA-Seq reads that support the splice junction (grey dotted lines in Figure 31). The most important part of a *TopHat* prediction in the “RNA-Seq *TopHat*” track is the line that connects the two boxes because it corresponds to the intron inferred by *TopHat*. 

23
To view the TopHat splice junction predictions for the first intron of CG31997, scroll down to the “RNA Seq Tracks” section and change the display mode for the “RNA-Seq TopHat” track to “pack” (Figure 32, top). Enter “contig10:25673-27194” into the “enter position or search terms” text box and then click “go” so that we can visualize the genomic region between the first and second coding exons.

![RNA Seq Tracks](image)

Figure 32  Multiple TopHat splice junction predictions for the first intron of CG31997-PB.

Examination of the features in the “TopHat junctions for Mixed Embryos” shows that the TopHat splice junction prediction JUNC00000027 is consistent with our annotated splice acceptor site for CDS 2_10739_2 at 27,077±27,078. However, TopHat also predicted two additional splice acceptor sites (i.e. JUNC00000025 and JUNC00000026) further upstream (red arrow in Figure 32, bottom). Hence we need to examine these splice junctions to ascertain if they are valid splice acceptor sites.
Enter “contig10:27,020-27,090” into the “enter position or search terms” text box and then click on the “go” button. Because the BLASTX CDS alignment for CDS 2_10739_2 is in frame +3, we can reject the splice acceptor candidate suggested by JUNC00000025 at 27,029-27,030 because it is located upstream of two in-frame stop codons (at 27,042–27,044 and 27,051–27,053, respectively) in frame +3 (Figure 33). The splice junction JUNC00000024 has the same splice acceptor site as JUNC00000025 but it is connected to a splice donor site at 25,484–25,485 (upstream of CDS 1_10739_0). We can apply the same rationale (i.e. presence of in frame stop codons) to reject this candidate as a potential splice junction for CDS 1_10739_0 and 2_10739_2.

While the splice acceptor site suggested by the TopHat junction JUNC00000026 at 27,057–27,058 is located before the two stop codons, this splice acceptor site is in phase 1 relative to frame +3, which means it is incompatible with the phase 1 donor site for the previous CDS 1_10739_0 (Figure 34).

Investigate the additional TopHat splice junction predictions

One possible explanation for these additional splice junctions predicted by TopHat is that there could be additional novel isoforms of CG31997 in D. biarmipes where the region upstream of CDS 2_10739_2 and the exons located further upstream are part of the 5’ untranslated region (5’ UTR). Specifically, CDS 2_10739_2 contains a methionine at 27,090–27,092 that could correspond to the start codon for these novel isoforms. The region between the splice acceptor sites predicted by TopHat and this start codon would be part of the 5’ UTR. Similarly, CDS 1_10739_0 would be another part of the 5’ UTR for these novel isoforms.

However, because these proposed isoforms do not exist in D. melanogaster, we need to have strong RNA-Seq evidence in D. biarmipes that supports the additional splice junctions. The score of each
TopHat junction prediction corresponds to the number of spliced RNA-Seq reads that supports the prediction. Hence we can use this score to assess the level of confidence that are associated with each TopHat prediction.

Click on the TopHat junction “JUNC00000027” in the “TopHat Junctions for Mixed Embryos (Eggs)” track. We find that this junction has a score of 1187, indicating that this splice junction is supported by 1,187 spliced RNA-Seq reads (Figure 35, left). Go back to the previous page and then click on the TopHat junction “JUNC00000025”. We find that this junction only has a score of 2 (Figure 35, middle). Hence the junction JUNC00000025 is only supported by two spliced RNA-Seq reads. Similarly, we find that the splice junction JUNC00000026 has a score of 1 (Figure 35, right), which means that it is only supported by a single spliced RNA-Seq read. Consequently, we have high levels of confidence in the JUNC00000027 TopHat prediction and low confidence in the JUNC00000025 and JUNC00000026 splice junction predictions.

We can configure the settings for the “RNA-Seq TopHat” track to filter out TopHat splice junction predictions that are only supported by a small number of spliced RNA-Seq reads. Scroll down to the “RNA Seq Tracks” section and click on the “RNA-Seq TopHat” link. Change the “Show only items with score at or above” to “10” and then click on the “Submit” button. This will filter out all the TopHat junctions that are supported by nine or fewer RNA-Seq reads.

Enter “contig10:25,673-27,194” into the “enter position or search terms” text box and then click on the “go” button so that we can examine the genomic region surrounding the splice junction between CDS 1_10739_0 and 2_10739_2. After applying the score filter, we find that there is only one splice junction prediction in each of the three samples (Figure 36).

Note that there is a trade-off between sensitivity and specificity when evaluating RNA-Seq data. RNA-Seq reads could be placed incorrectly in the assembly (e.g., because of base calling errors, repetitive sequences in the genome). Although rare, splicing errors also occur. Hence we are generally skeptical of splice junctions that are only supported by a small number of RNA-Seq reads. However, rare transcripts or genes expressed only at low levels would also show low RNA-Seq read coverage. Consequently, we cannot a priori determine the appropriate cutoff scores for the TopHat splice junction predictions.
The annotators at FlyBase also encounter the same trade-off with the RNA-Seq data when they create gene models in *D. melanogaster*. For example, the FlyBase gene report for CG31997 [https://flybase.org/reports/FBgn0051997.html](https://flybase.org/reports/FBgn0051997.html) includes a note in the “Comments on Gene Model” section (under “Gene Model and Products”) indicating that the *D. melanogaster* gene models do not account for all of the RNA-Seq junctions that are only supported by a small number of spliced RNA-Seq reads (Figure 37).

The GEP annotation strategy is based on parsimony (i.e. minimizing the number of changes compared to *D. melanogaster*). Our initial hypothesis is that the same set of isoforms in *D. melanogaster* also exists in the *D. biarmipes* ortholog. We will only postulate novel isoforms based on multiple lines of evidence: high RNA-Seq coverage, strong TopHat predictions, sequence conservation with other closely related *Drosophila* species (e.g., *D. takahashii*), and computational gene predictions.
Collectively, our analysis of the *TopHat* splice junctions support the placement of the splice donor site at 25,836–25,837 for CDS 1_10739_0 and the splice acceptor site at 27,077–27,078 for CDS 2_10739_2 in contig10. The *TopHat* results indicate that there might be additional splice acceptor sites for CDS 2_10739_2. However, because these splice junctions are only supported by a small number of spliced RNA-Seq reads, there is insufficient evidence to propose novel isoforms of CG31997 in *D. biarmipes*.

**Annotating the remaining splice sites**

We can apply the strategy described above to annotate the splice donor site for CDS 2_10739_2 and the splice acceptor site for CDS 3_10739_1. The *BLASTX* alignment for CDS 2_10739_2 spans from 27081-27194 in frame +3 but the last amino acid of this CDS is missing from the alignment (Figure 22). Hence we would expect to find the splice donor site at around 27,197 (i.e. 27194+3). Enter “contig10:27,197” into the “enter position or search terms” text box and then click on the “go” button. Zoom out 10x and then zoom out another 3x.

![Phase 2 splice donor site at the end of CDS 2_10739_2.](image)

Examination of this region in the Genome Browser shows that there is only a single splice donor site (at 27,200-27,201) prior to the first in frame stop codon in frame +3. This phase 2 splice donor site is supported by the RNA-Seq Alignment Summary tracks, the RNA-Seq *TopHat* tracks as well as the NSCAN gene prediction (Figure 38). This also means that the splice acceptor site for CDS 3_10739_1 must be in phase 1.

Because the *BLASTX* alignment for CDS 3_10739_1 begins at 27,286 in frame +1 (Figure 20), we will search for a phase 1 splice acceptor site near this position. Enter “contig10:27,286” into the “enter position or search terms” text box and then click on the “go” button. Zoom out 10x and then zoom out another 3x. The acceptor site at 27,283–27,284 is in phase 1 relative to frame +1 and it is supported by the *D. biarmipes* RNA-Seq data and the NSCAN gene prediction (Figure 39).
Based on the available evidence, we will annotate the end of the CDS 2_10739_2 at 27,199 and the start of the CDS 3_10739_1 at 27,285.

**Verifying the gene model using the Gene Model Checker**

Our analysis of the exon-by-exon BLASTX alignments and the evidence tracks on the GEP UCSC Genome Browser allow us to precisely define the start and end positions of each of the three coding exons of CG31997-PB. To verify that our proposed gene model satisfies the basic biological constraints (e.g., begins with a start codon, has compatible splice sites, ends with a stop codon), we will check our gene model coordinates using the Gene Model Checker.

Open a new tab and navigate to the F Element project page on the GEP web site (https://thegep.org/felement/). Click on the “Gene Model Checker” link under the “Resources & Tools” section (Figure 40).
The first part of the configuration will define the analysis region. Select “D. biarmipes” under the “Species Name” field, then select “Aug. 2013 (GEP/Dot)” under the “Genome Assembly” field. Enter “contig10” into the “Scaffold Name” field.

Enter “CG31997-PB” under the “Ortholog in D. melanogaster” field. Under the “Coding Exon Coordinates” field, enter a comma-delimited list of coordinates for the three coding exons: “25673-25835, 27079-27199, 27285-27468”.

Note that the coordinates for the “Coding Exon Coordinates” field **do not include the stop codon**. We will enter the stop codon coordinates separately in the “Stop Codon Coordinates” field.

Because CG31997-PB is on the positive strand relative to contig10, we should verify that “Plus” is selected under the “Orientation of Gene Relative to Query Sequence” field. The Gene Model Checker will automatically infer the stop codon coordinates (i.e. “27469-27471”) when you select the “Stop Codon Coordinates” field. (Figure 41). Click on the “Verify Gene Model” button to run the Gene Model Checker.

Once the analysis is complete, the right panel contains the results of the Gene Model Checker analysis. The “Checklist” tab enumerates the list of criteria that have been checked by Gene Model Checker (Figure 42). For example, the Gene Model Checker verifies that our proposed gene model begins with a start codon and ends with a stop codon. It also verifies that the splice junctions
contain the canonical splice donor and acceptor sites. Some of the items on the checklist have been skipped because they do not apply to a complete gene (e.g., Acceptor for CDS 1).

Figure 42 The checklist produced by the Gene Model Checker for our D. biarmipes CG31997-PB gene model.

The Gene Model Checker checklist is designed to highlight unusual features in the gene model. Warnings and failures reported by the Gene Model Checker do not necessarily mean that the proposed gene model is incorrect. However, the annotator should provide additional evidence that justifies the unusual annotation (e.g., non-canonical splice donor site, stop codon read through).

In addition to verifying the basic gene structure, the Gene Model Checker also compares the proposed gene model against the putative D. melanogaster ortholog using a protein alignment and a dot plot. Click on the "Dot Plot" tab to examine the dot plot between the D. melanogaster protein (x-axis) and the protein sequence for the submitted model in D. biarmipes (y-axis). The alternating color boxes correspond to the different coding exons in the two sequences. Dots in the dot plot correspond to regions of similarity between the D. melanogaster protein and the submitted D. biarmipes gene model.

If the submitted sequence is identical to the D. melanogaster ortholog, then the dot plot will show a straight diagonal line with a slope of 1. Changes in the size of the submitted model compared to the D. melanogaster ortholog will alter the slope of this line. In this case, the dot plot shows that the last two CDS's of CG31997-PB in D. melanogaster and D. biarmipes have similar lengths but the initial CDS (1_10739_0) of CG31997-PB in D. biarmipes is substantially longer than the orthologous CDS in D. melanogaster. Furthermore, the dot plot also did not detect any sequence similarity between the beginning of the submitted model and the beginning of the D. melanogaster ortholog (Figure 43).
The dot plot alignment shows that the main differences between the *D. melanogaster* protein CG31997-PB (x-axis) and the submitted *D. biarmipes* gene model (y-axis) are located within the first CDS.

To further investigate the discrepancies in the dot plot, we will examine the protein alignment between the two sequences (Figure 44). Click on the “View protein alignment” link above the dot plot. The alignment shows the comparison of the *D. melanogaster* protein (top) against the conceptual translation for the submitted *D. biarmipes* gene model (bottom). Similar to the dot plot, the alternating colors correspond to the different coding exons.

The gap in the dot plot between the second and third CDS's of *D. melanogaster* CG31997-PB and the *D. biarmipes* gene model is caused by amino acids near the splice site boundary that are similar but not identical (red box).
The protein alignment shows that the last two CDS’s have high levels of sequence similarity between the *D. melanogaster* ortholog and the *D. biarmipes* gene model. The symbols in the match line denote the level of similarity (* indicates conserved amino acids, : denotes amino acids with highly similar chemical properties). Hence the gap in the dot plot between the second and third CDS can be attributed to similar but not identical amino acids near the splice site boundary.

Consistent with the results of our exon-by-exon BLASTX analysis, the protein alignment shows that the end of the first CDS is very highly conserved while the beginning of the CDS alignment shows multiple gaps (-) in the *D. melanogaster* protein sequence compared to the submitted *D. biarmipes* model (Figure 45). These gaps in the first CDS explain why the *D. biarmipes* gene model is 8 amino acids longer than the *D. melanogaster* ortholog.

### Figure 45
The three gaps in the protein alignment of *D. melanogaster* CG31997-PB (top) against the *D. biarmipes* model (bottom) accounts for the 8 extra amino acids in the *D. biarmipes* model compared to *D. melanogaster* (i.e. 156 versus 148aa).

You should provide an explanation for any large gaps or changes in slope in the dot plot in the F Element Project Annotation Report Form. In particular, a large vertical or horizontal gap at the beginning or the end of an exon in the dot plot often indicates the presence of alternate splice sites that would minimize the change in the size of the CDS compared to *D. melanogaster*. You should include a detailed explanation in the F Element Project Annotation Report in order to support the truncation or expansion of the CDS compared to the *D. melanogaster* model (e.g., location of the compatible splice donor or acceptor sites, RNA-Seq TopHat splice junctions).

Our previous analysis has shown that there are no alternate start codons available for CDS 1_10739_0 and that the expansion of this CDS compared to the *D. melanogaster* ortholog is strongly supported by the *D. biarmipes* RNA-Seq data (Figure 25). To verify our previous observations, we can view the submitted gene model within the context of the other evidence tracks in the GEP UCSC Genome Browser.

Select the “Checklist” tab and then click on the magnifying glass icon next to the “Check for Start Codon” criteria. A new window will appear with our submitted gene model shown in the red.
“Custom Gene Model” track. Zoom out 10x so that we can examine the entire reading frame for CDS 1_10739_0 in frame +2 (Figure 46). The Genome Browser view shows that our gene model begins at the only available start codon in frame +2 before the first in-frame stop codon. This proposed start codon location is consistent with the N-SCAN gene prediction and it is also supported by the RNA-Seq data.

![Gene Model Checker](image)

**Figure 46**  Click on the magnifying glass icon in the Gene Model Checker checklist to view the submitted gene model in the context of the other evidence tracks on the GEP UCSC Genome Browser.

The “Gene Model Checker User Guide” contains a more comprehensive overview of the Gene Model Checker. The user guide also includes a walkthrough on how to use the Gene Model Checker to identify and diagnose problems in the proposed gene model.

**Download the files required for project submission**

In addition to the F Element Project Annotation Report Form, you must prepare three additional data files in order to submitting a project to the GEP: a General Feature Format (GFF) file, a transcript sequence (fasta) file, and a peptide sequence (pep) file. The Gene Model Checker automatically creates these three files when you verify a gene model. We can download these files by selecting the “Downloads” tab (Figure 47) and then right-click (control click on macOS) on each of the links and select “Save Links As...” or “Download Linked File As ...” to save each file onto your computer.
The GFF, transcript, peptide sequence files for all the genes and isoforms in your project should be combined into a single file prior to project submission. You can use the “Annotation Files Merger” (available through the F Element project page on the GEP web site) to create the combined GFF, peptide, and transcript sequence files for your entire project. See the Annotation Files Merger User Guide for additional details.

Conclusion
This walkthrough illustrates many of the key steps of the GEP annotation strategy. Using the “D. mel Proteins” and the gene predictions tracks on the GEP UCSC Genome Browser, we identified three features of interest within the contig10 project from the D. biarmipes Muller F element. To further investigate one of these features within contig10, we performed a BLASTP search to compare the N-SCAN prediction contig10.001.1 against a database of D. melanogaster annotated proteins at FlyBase. This BLASTP search indicates that the D. biarmipes genomic region surrounding this N-SCAN prediction likely contains an ortholog of the D. melanogaster gene CG31997. Using the Gene Record Finder, we determined the overall gene structure (e.g., number of isoforms and unique CDS’s) of CG31997 in D. melanogaster.

We then compared each of the unique CDS’s of CG31997 against the contig10 sequence using NCBI BLASTX to determine the approximate placement of each CDS. We further refined the placement of the CDS’s using the RNA-Seq evidence tracks on the GEP UCSC Genome Browser. We used the “RNA-Seq Alignment Summary” track to verify the placement of the start codon and the “RNA-Seq TopHat” tracks to identify the splice donor and acceptor sites.

Once we have determined the coordinates of all the CDS’s, we verified our proposed gene model using the Gene Model Checker. The Gene Model Checker checklist confirms that our proposed gene model satisfies the basic biological constraints of a eukaryotic gene. We also examined the dot plot and the protein alignment between our proposed gene model and the D. melanogaster ortholog to verify that the differences between the two sequences are genuine.

The final step is to document the results of our analysis in the F Element Project Annotation Report Form. We have included a sample F Element Project Annotation Report for CG31997 in the package for this walkthrough (Sample_GEP_Annotation_Report.docx).
Some of the genes in *D. biarmipes* are more challenging to annotate than the example described in this walkthrough. The “Annotation Instruction Sheet” contains additional strategies on how to identify small or weakly conserved coding exons. The “Annotation Strategy Guide” illustrates how the concepts described in the “Annotation Instruction Sheet” can be applied to more challenging annotation cases.
F Element Project: Annotation Report

Faculty instructor(s): Sarah C.R. Elgin
College/university: Washington University in St. Louis
Course number: Bio 4342
Course name: Research Explorations in Genomics

Project Details

Project name: contig10
Project species: D. biarmipes
Date of submission: 08/15/2021
Size of project in base pairs: 43,013
Number of genes in project: 3

Does this report cover all of the genes or is it a partial report? Partial report
If this is a partial report, please indicate the region of the project covered by this report:
From base 25,000 to base 28,000

Note: For each gene described in this annotation report, you should also prepare the corresponding GFF, transcript and peptide sequence files as part of your submission.
Complete the following Gene Report Form for each gene in your project. Copy and paste the sections below to create as many copies as needed within this report. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms.

**Gene Report Form**

Gene name (e.g., *D. ananassae eyeless*): *D. biarmipes CG31997*
Gene symbol (e.g., *dana_ey*): *dbia.CG31997*

Approximate location in project (from 5’ end to 3’ end): **25673-27471**
Number of isoforms in *D. melanogaster*: 2
Number of isoforms in this project: 2

Complete the following table, including all of the isoforms in this project:

<table>
<thead>
<tr>
<th>Name(s) of unique isoform(s) based on coding sequence</th>
<th>List of isoforms with identical coding sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG31997-PB</td>
<td>CG31997-PA</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species: **NA**

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (*e.g.*, changes in canonical splice sites, gene structure, etc.):

**NA**

**Note:** For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, irrespective of whether their coding sequence is identical to that of another isoform.
Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

All of the coordinates reported in this section should be relative to the coordinates of the original project sequence.

Location(s) in the project sequence with consensus errors:
NA

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

2. Generate a VCF file which describes the changes to the consensus sequence

Use the Sequencer Updater to create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. Paste a screenshot with the list of sequence changes into the box below:
Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

Gene-isoform symbol (e.g., dana_ey-PA): dbia.CG31997-PB

Names of any additional isoforms with identical coding sequences: dbia.CG31997-PA

Is the 5’ end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 5’ end: _____________

Is the 3’ end of this isoform missing from the end of the project? No
If so, how many putative exons are missing from the 3’ end: _____________

(Define “putative exons” based on the exons present in the *D. melanogaster* ortholog)

1. Gene Model Checker checklist

Enter the coordinates of your final gene model for this isoform into the *Gene Model Checker* and paste a screenshot of the checklist results into the box below:

**Note:** For projects with consensus sequence errors, report the exon coordinates relative to the original project sequence. Include the VCF file you have generated above when you submit the gene model to the *Gene Model Checker*. The *Gene Model Checker* will use this VCF file to automatically revise the submitted exon coordinates.
2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the “Checklist” tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:

1. A sequence alignment track (e.g., D. mel Proteins)
2. At least one gene prediction track (e.g., Genscan)
3. At least one RNA-Seq track (e.g., RNA-Seq Coverage)
4. A comparative genomics track (e.g., D. mel. Net Alignment, Conservation)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:

![Gene Model Screenshot]

Low-frequency RNA-Seq exon junctions not annotated:
The evidence from the RNA-Seq TopHat evidence tracks and Multiz alignments suggest that there might be additional isoforms because of alternative splicing at the 5’ end of this gene (red arrows in the screenshot above). However, because most of the TopHat junctions are supported by less than 10 reads, there is insufficient evidence to postulate the presence of multiple novel isoforms in D. biarmipes compared to D. melanogaster.
**Extra CDS predicted by the SNAP gene predictor:**

*SNAP* predicted a CDS at 26,502-26,584 (blue arrow in the screenshot above) between the first and second CDS's of *CG31997*. The RNA-Seq Alignment Summary track shows that the region surrounding this region has low (<20 reads) RNA-Seq read coverage and the region is adjacent to a hAT DNA transposon fragment (see screenshot below).

NCBI *BLASTX* search of the genomic region surrounding the *SNAP* CDS prediction (contig10:26400-26700) against the *nr* database did not detect any significant (E-value < 1e-5) sequence similarity to known proteins in the *nr* database (see screenshot below).
A NCBI BLASTN search of this region against the nt database detected 19 significant matches to predicted mRNAs in *Drosophila subpulchrella* and *Drosophila suzukii* (see screenshot below). Both *Drosophila* species are members of the suzukii subgroup.

The E-values for the *D. subpulchrella* matches range from $9\times10^{-10}$ to $3\times10^{-9}$, and they correspond to four different predicted genes (LOC119559709, LOC119559300, LOC119559298, and LOC119547467). The E-values for the *D. suzukii* matches range from $1\times10^{-0.8}$ to $5\times10^{-0.6}$, and they correspond to five different predicted genes (LOC108011950, LOC108013970, LOC118879467, LOC118878470, LOC108014610). All of these matches are RefSeq predictions that have not been confirmed experimentally. There are no significant matches to RefSeq records that are supported by experimental evidence and no significant matches to mRNAs in other species outside of the suzukii subgroup.

Collectively, while we could not reject the possibility that this region of contig10 contains an untranslated region of a nearby gene, there is insufficient evidence to postulate a novel
isoform of CG31997 in D. biarmipes compared to D. melanogaster. Given the proximity of this feature to the hAT DNA transposon and the multiple matches to predicted transcripts in D. subpulchrella and D. suzukii, an alternative explanation is that the feature is part of a transposon that is found in D. biarmipes, D. subpulchrella, and D. suzukii. Hence we have omitted this predicted CDS in our annotation of the CG31997 ortholog in D. biarmipes.

3. Alignment between the submitted model and the D. melanogaster ortholog

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative D. melanogaster ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the “View protein alignment” link under the “Dot Plot” tab) or you can generate a new alignment using the “Align two or more sequences” feature at the NCBI BLAST web site. Paste a screenshot of the protein alignment into the box below:

![Alignment of Dmel.CG31997-PB vs. Dbia3.CG31997-PB](image-url)
4. Dot plot between the submitted model and the *D. melanogaster* ortholog

Paste a screenshot of the dot plot (generated by the *Gene Model Checker*) of your submitted model against the putative *D. melanogaster* ortholog into the box below. Provide an explanation for any anomalies on the dot plot (e.g., large gaps, regions with no sequence similarity, indications of significant insertions or deletions).

**Note:** Large **vertical and horizontal gaps** near exon boundaries in the dot plot often indicate that an incorrect splice site might have been picked. Please re-examine these regions and provide a justification as to why you have selected this particular set of donor and acceptor sites.
The dot plot shows that the last two CDS's of CG31997-PB are highly conserved between the proposed *D. biarmipes* gene model and the *D. melanogaster* ortholog. Examination of the protein alignment at the end of the second and third CDS's indicate that the amino acids have similar chemical properties even though they are not identical. In addition, the lengths of these two CDS's are the same between *D. biarmipes* and *D. melanogaster*.

The dot plot shows that the beginning of the first CDS of CG31997-PB is only weakly conserved between *D. biarmipes* and *D. melanogaster*. In addition, the dot plot shows that the first CDS of the *D. biarmipes* gene model is longer than the orthologous CDS in *D. melanogaster*. The protein alignment shows that there are 8 additional amino acids within the first CDS in the proposed *D. biarmipes* gene model compared to *D. melanogaster*.

Examination of this region in the *GEP UCSC Genome Browser* shows that there is only one methionine in frame +2 that could serve as the start codon for CG31997-PB (see screenshot below). The expansion of this CDS is consistent with the *BLASTX* alignment, the *N-SCAN* gene prediction, and the available RNA-Seq data. Consequently, our annotation has expanded the size of this CDS (1_10739_0) in order to retain this isoform in *D. biarmipes*. 
Select gene prediction

Identify *D. melanogaster* ortholog (workflow)

Review *D. melanogaster* gene structure

Select isoform to annotate

Map each coding exon (CDS) to contig

Determine CDS coordinates (workflow)

Verify isoform model

Click on gene prediction and the “Predicted Protein” link

FlyBase blastp search of predicted protein against the *D. melanogaster* “Annotated Proteins” database

Determine number of isoforms and unique coding exons (CDS):
  - *Gene Record Finder* tables
  - Genome Browsers (e.g., *JBrowse*)
  - Obtain *D. melanogaster* CDS sequence from *Gene Record Finder*
  - Use NCBI *blastx* ➔ Align two or more sequences to compare contig against CDS
    - Record start and end positions, coverage, and reading frame
  - Identify the best supported donor (GT/GC) and acceptor sites (AG)
  - Minimize changes compared to *D. melanogaster*

Provide explanations for:
  - Errors and warnings in checklist
  - Large gaps in dot plot
  - Gaps near splice sites

---

**UCSC Genome Browser**  **Gene Record Finder**  **Gene Model Checker**

**FlyBase**  **NCBI BLAST (bl2seq)**
**Identify *D. melanogaster* ortholog**

FlyBase `blastp` search of predicted protein against the *D. melanogaster* "Annotated Proteins" database

- Has at least one significant match (E-value < 1e-5)?
  - Yes
  - No

  - Matches multiple *D. melanogaster* genes?
    - Yes
      - Assign gene with the **lowest E-value** as the putative ortholog
    - No

  - Multiple matches with E-values similar to the best match?
    - Yes
      - Assign gene with the **highest sequence identity** as the putative ortholog
    - No

  - Multiple matches with sequence identity similar to the best match?
    - Yes
      - Assign gene with the **highest alignment coverage** as the putative ortholog
    - No

  - Multiple matches with alignment coverage similar to the best match?
    - Yes
      - Assign putative ortholog based on **synteny** with *D. melanogaster*
    - No

Consult instructor

NCBI `blastp` search of predicted protein against the `nr protein` database

- Has at least one significant match (E-value < 1e-5)?
  - Yes
  - No

- Potential novel gene
  - Yes
  - No

- Probable misprediction
  - Yes
  - No

Consult instructor
Determining CDS Coordinates

**query** - CDS for *D. melanogaster* exon
**subject** - nucleotide (nt) sequence of interest (turn off low complexity filter and use no compositional adjustment)

For **highly conserved alignments**: look for splice sites nearest the exon boundaries
For **less conserved alignments**: attempt to extend the exon boundaries to include additional sequence from the Open Reading Frame (ORF), while still maintaining correct splice phases; conserve exon size

1. If available, check RNA-Seq evidence to prove canonical or alternative splice site exists (e.g., TopHat or regtools junctions)
2. Gene prediction tracks in the *GEP UCSC Genome Browser* to support alignment to selected site or demonstrate alternative solution
3. **BLASTX** track to verify level and length of conservation
4. Check closely-related species for choice of splice site (**BLAT** analysis): sequencing error?
5. Perform **Clustal Omega** searches to test level of conservation between species

---

**flybase.org** to check the intron sequence of interest:
Is the non-canonical splice site present for this gene in *D. melanogaster*?