Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 34, 135-165, 2013

Introduction to the Genomics Education Partnership and Collaborative Genomics Research in *Drosophila*

Julia A. Emerson[1], S. Catherine Silver Key[2], Consuelo J. Alvarez[3], Stephanie Mel[4], Gerard P. McNeil[5], Kenneth J. Saville[6], Wilson Leung[7], Christopher D. Shaffer[7] and Sarah C. R. Elgin[7]

[1]Amherst College, Department of Biology, P.O. Box 5000, Amherst MA 01002 USA
[2]North Carolina Central University, Department of Biology, 2246 MTSB, Durham NC 27701 USA
[3]Longwood University, Department of Biological and Environmental Sciences, 201 High St., Farmville VA 23909 USA
[4]University of California San Diego, Division of Biological Sciences, 9500 Gilman Drive 0355, La Jolla CA 92093 USA
[5]York College/CUNY, Department of Biology, 94-20 Guy R. Brewer Blvd., Jamaica NY 11451 USA [6]
Albion College, Biology Department, 611 E. Porter St., Albion MI 49224 USA
[7]Washington University, Department of Biology, Campus Box 1137, One Brookings Dr., St. Louis MO 3130 USA
(**jemerson@amherst.edu; ckey@NCCU.edu; alvarezcj@longwood.edu; smel@ucsd.edu; gmcneil@york.cuny.edu; ksaville@albion.edu; wleung@wustl.edu; shaffer@biology.wustl.edu; selgin@biology.wustl.edu**)

The GEP, a group of faculty from over 90 primarily undergraduate institutions, is using comparative genomics to engage students in research within regular courses. Using a versatile curriculum, GEP undergraduates undertake projects to improve draft genomic sequences and/or participate in the annotation of these improved sequences. An additional goal of the annotation curriculum is for students to gain a sophisticated understanding of eukaryotic gene structure. Students do so by carefully mapping the protein-coding regions of putative genes from recently-sequenced *Drosophila* species. This paper describes some ways in which the GEP's annotation tools have been adopted for use in undergraduate courses.

Keywords: Genomics Education Partnership (GEP), gene annotation, *Drosophila*, eukaryotic gene

Introduction

An effective method for teaching science is to engage students in doing science. The Genomics Education Partnership (GEP), a group of faculty from over 90 primarily undergraduate institutions, is using comparative genomics to engage students in research within regular academic-year laboratory courses (**http://gep.wustl.edu**). Using a versatile curriculum that has been adapted to many different class settings (as the core of a research-based laboratory course, as an independent research project, or as a smaller number of activities within the lab of a broader course), GEP undergraduates undertake projects to improve draft genomic sequences and/or participate in annotation of these improved sequences. GEP undergraduates have improved over 2.4 million bases of draft genomic sequence from several species of the fruit fly *Drosophila* and have produced hundreds of gene models using evidence-based manual annotation. An additional goal of the GEP is for students to gain a more sophisticated understanding of eukaryotic gene structure and mRNA processing than is often achieved in lectures. This paper provides examples of how GEP resources are being used at multiple colleges and universities, with a focus on gene annotation labs at Amherst College.

Amherst College is a private liberal arts college that enrolls approximately 1,800 undergraduate students. At Amherst, gene annotation has been integrated into the existing laboratory curriculum of the semester-long course, Biology 191 (*Molecules, Genes and Cells*). Biology 191 is one of two required introductory courses for the biology major at Amherst, and it is taken by many more additional students to fulfill the pre-medical requirements. Class enrollment is 90-120 students each year, and the course is predominately made up of first-semester sophomores, due to a chemistry pre-requisite. There are 3 hours of lecture and an optional (4th hour) question and answer session each week. Each student also attends one of the 3-hour lab sections each week, and lab enrollment is between 20 and 24 students per section. Students receive instruction and assistance in lab from one faculty instructor and two undergraduate teaching assistants. Using 1 hour of discussion and 9 hours of lab time over the last quarter of the semester, students receive training in DNA database searches and complete a gene annotation research project, carefully defining the protein-coding regions of a putative gene from a recently-sequenced *Drosophila* species (**http://flybase.org/static_pages/species/sequenced_species.html**). Lectures on gene structure and gene expression are completed just prior to the start of the annotation labs. Assessment of student learning gains in gene annotation is accomplished using the GEP's on-line pre- and post-course quizzes, and the students' overall experiences in the course are assessed using the GEP's on-line pre- and post-course surveys and an Amherst College course evaluation.

The following Student Outline begins with an overview of the GEP's objectives and then segues into a detailed introduction to the topic of gene annotation. The Student Outline is followed by *A Sample Annotation Problem*, which is a modified version of one of the GEP's annotation training tools. The S*ample Annotation Problem* is the second of two training activities that each Biology 191 student completes prior to starting his/her individual project. Figure 1 and the Implementation section below describe the organization of the gene annotation labs at Amherst College, while the Appendix contains shorter descriptions of how GEP materials have been used in courses at several other institutions.

Implementation

During the first week of gene annotation work, the lab coordinator uses PowerPoint slides and the GEP Web site to introduce all students in the course to the overall objectives of the annotation labs during Tuesday's '4th hour' (which precedes all the labs for the week). The pre-lab reading assignment for the week is the Student Outline (section 2), which includes revised parts of the GEP's *Annotation Instruction Sheet*. The instructor of each lab section provides a brief overview of the National Center for Biotechnology Information (NCBI) BLAST search engine to the students, who then work through the GEP's *Simple Introduction to NCBI BLAST* activity (**http://gep.wustl.edu/curriculum/course_materials_WU/annotation/tutorials_and_walk-throughs**).

Students work individually at desktop computers, although they are encouraged to talk with their neighbors/lab partner about what they are doing and learning. Most students complete the exercise during the lab period, and any students who do not finish are instructed to complete the activity as homework. All students are also expected to complete an 'open-book' quiz with 14 multiple choice questions drawn from the Student Outline and the *BLAST* activ-

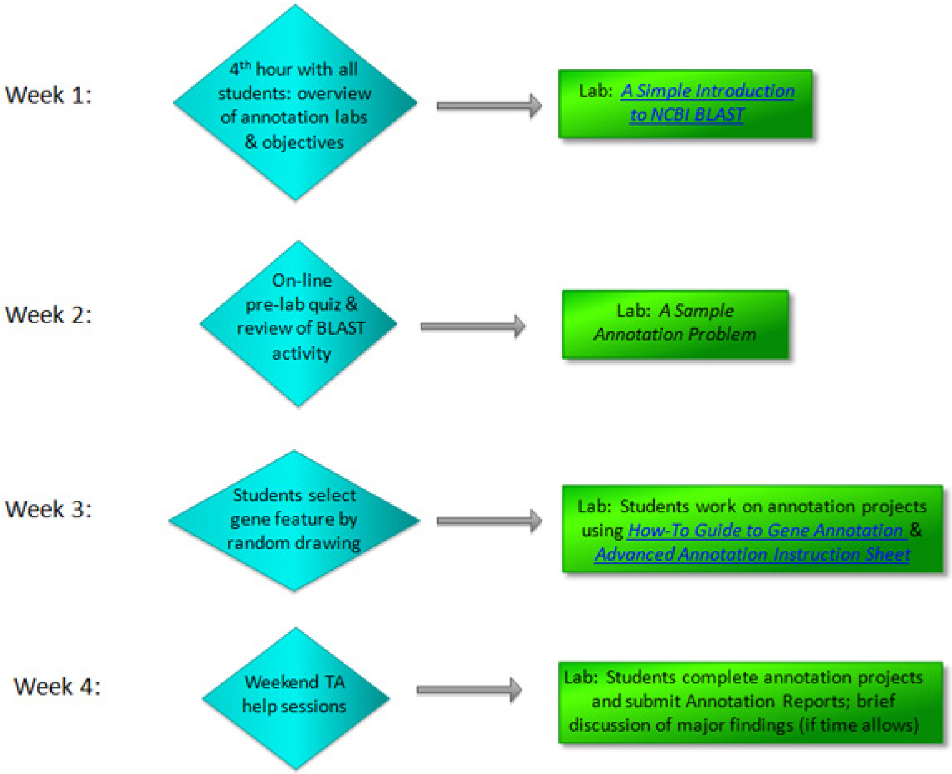

Week 1:
4th hour with all students: overview of annotation labs & objectives
Lab: A Simple Introduction to NCBI BLAST
Week 2:
On-line pre-lab quiz & review of BLAST activity
Lab: A Sample Annotation Problem
Week 3:
Students select gene feature by random drawing
Lab: Students work on annotation projects using How-To Guide to Gene Annotation & Advanced Annotation Instruction Sheet
Week 4:
Weekend TA help sessions
Lab: Students complete annotation projects and submit Annotation Reports; brief discussion of major findings (if time allows)

Figure 1. Organization of the gene annotation labs in the Biology 191 course at Amherst College

ity. The quiz and answer sheet are linked to the course Web site.

The quiz answer sheets are collected from the students as they walk into lab the next week and quickly graded by the two teaching assistants. The lab instructor then goes over questions that were answered incorrectly by large numbers of students, to make sure that all students understand the particular topics addressed by those questions.[1] Completion of the quiz (but not the actual grade) counts towards the final grade for each student's annotation project. Following the quiz discussion, students work individually on a revised and expanded version of the GEP's *A Simple Annotation Problem* (renamed *A Sample Annotation Problem*, see sections 3 and 4). Students are again encouraged to discuss the activity with their neighbors, and the lab instructor and teaching assistants circulate around the lab to answer additional questions. Midway through the lab period, all students receive additional instructions from the instructor for fine mapping of exon boundaries using the GEP's mirror of the UCSC Genome Browser. This includes how to examine + or – strand DNA sequences and how to evaluate reading frames and splice sites, all areas where students tend to get confused. Students then resume their work, and each student is required to show his or her completed *Sample Annotation Problem* worksheet to a teaching assistant before leaving lab 2. Any students who do not finish the *Sample Annotation Problem* and its Addendum are instructed to complete the activity as homework.[3]

At the beginning of the third week, students select a gene (via random draw) from one or more fosmids that contain genomic DNA from a particular species of *Drosophila*. The fosmids have been pre-screened by the lab coordinator to determine which predicted features are likely to be real protein-coding genes. The lab section's fosmid(s) are then projected on the large screen, and the lab instructor spends a few minutes going over information displayed in the GEP's mirror of the UCSC Genome Browser window and how the selected gene features were identified. Two or more students in each lab section are assigned to each gene, so that they can collaborate in completing the annotation. We also even out the workload somewhat between student teams by assigning multiple pairs of students to gene isoforms with ~10 or more exons.

After the gene drawing, students work on their selected isoforms using Amherst's *How-To Guide for Gene Annotation*. Students also receive copies of an *Advanced Annotation Instruction Sheet*, which addresses more challenging annotation problems and strategies. The *Advanced Annotation In-*

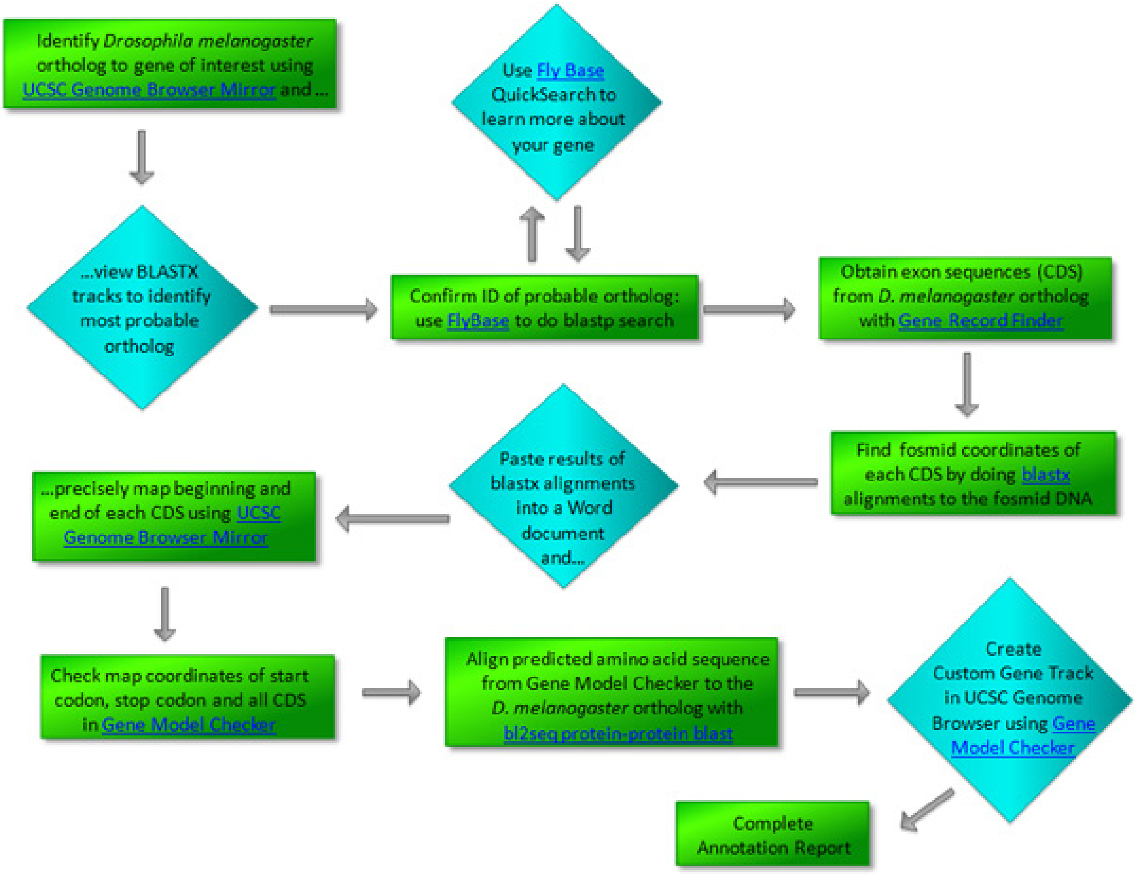

Identify *Drosophila melanogaster* ortholog to gene of interest using UCSC Genome Browser Mirror and ...
Use Fly Base QuickSearch to learn more about your gene
...view BLASTX tracks to identify most probable ortholog
Confirm ID of probable ortholog: use FlyBase to do blastp search
Obtain exon sequences (CDS) from *D. melanogaster* ortholog with Gene Record Finder
Find fosmid coordinates of each CDS by doing blastx alignments to the fosmid DNA
Paste results of blastx alignments into a Word document and...
...precisely map beginning and end of each CDS using UCSC Genome Browser Mirror
Check map coordinates of start codon, stop codon and all CDS in Gene Model Checker
Align predicted amino acid sequence from Gene Model Checker to the *D. melanogaster* ortholog with bl2seq protein-protein blast
Create Custom Gene Track in UCSC Genome Browser using Gene Model Checker
Complete Annotation Report

Figure 2. Summary of steps in the **How-To Guide for Gene Annotation**

struction Sheet is another modification of a GEP document and available as above. A diagram that summarizes the steps in the *How-To Guide*, which students follow to complete their annotation projects, is below (Fig. 2). Each teaching assistant holds a one-hour help session over the weekend after this third week, so that students can ask questions and get one-on-one assistance with their individual projects.

In the fourth and final week, students complete work on their gene models. Each student team submits two Word documents: (1) their 'working' Word file, which includes a screen shot of the Gene Record Finder (that displays the *D. melanogaster* coding sequence blocks for their particular isoform) and the results of the exon-by-exon BLAST searches; and (2) a completed Annotation Report for their isoform.[4] Any students who finish early are 'deputized' as TA's to help their fellow students. A course tradition is to have pizza and soft drinks during the final lab, and we use part of this time to have students briefly comment on anything of particular interest about their gene model to the other students. The lab coordinator then compiles the individual isoform reports into fosmid reports after the semester ends.[5]

Lessons Learned and Future Plans

The following comments refer to the numbers in the above section.

[1] In the future, we may turn this into an on-line quiz, which will be automatically graded and provide us with more advance notice of which concepts are confusing to the students.

[2] An answer key to the *Sample Annotation Problem* is posted on the course Web site at the end of the second week of annotation labs. Students can then compare their answers to the key, thereby reinforcing their understanding of annotation concepts and methodology. Students are also instructed to come to the next lab with any questions they still have about the activity after looking over the answer key.

[3] Thanksgiving Break occurs between the second and third weeks of the annotation labs, resulting in some regression in understanding and preparation. Thus, we may add a second pre-lab quiz prior to the third week of annotation, so students are a bit more prepared to begin their individual projects.

[4] As much as possible, we have all the genes on each claimed fosmid annotated by students in two different lab sections, so that we can compare the two annotation reports to one another and revise as needed.

[5] Since most individual projects and gene models are not carefully checked until they are graded, the final annotation reports cannot be compiled until after the labs are over (in case any revisions/additions to particular models need to be made). Thus, the lab coordinator is ultimately responsible for compiling and submitting the final fosmid annotation reports and data files to Washington University. We limit the number of fosmids that we claim for our course to three or four, to make sure this last step is not overly time-consuming (which it would be if each student in a class of over 100 students worked on a different gene/isoform). GEP researchers at Washington University and elsewhere use the submitted reports and data files in ongoing investigations of genome evolution in *Drosophila* (see **http://gep.wustl.edu/community/publications**).

Sample Student Evaluations of Gene Annotation Labs at Amherst College (Fall, 2010)

Which lab(s) were most engaging and effective in enhancing your understanding of biological concepts? Why?

"The genome sequencing for the last part of lab was very interesting, as it was an application of some of the things we had learned earlier in the semester about transcription and translation, while also adding to our biology skill set… it was very interesting and instructive in terms of genetics and evolution."

"These [gene annotation labs] were awesome. I was able to complete mine between lab time and about an hour outside, and I really learned a lot from it. Also, once I got all green checks on the Model Checker, I was so happy!"

Student Outline

Overview of the Genomics Education Partnership

The Genomics Education Partnership (GEP) is a national, collaborative, scientific investigation of a problem in genomics, involving wet-lab generation of a large data set (e.g., sequence improvement of genomic DNA) and computer analyses of the data (including **annotation** of genes, assessment of repeats, exploration of evolutionary questions, etc.). At present, the research problem entails generating finished sequence from the fourth (dot) chromosome (Fig. 3) of various species of *Drosophila*, annotating these sequences, and making comparisons among species and to control DNA sequences from the third chromosome to discern patterns of genome organization related to the control of gene expression.

The scientific interest is based on observations that the dot chromosome shows a mixture of heterochromatic and euchromatic properties. The chromosome stains intensely with fluorescent DNA-binding dyes, has a high density of repetitious sequences, is late replicating, and exhibits very low meiotic recombination, all properties of **heterochromatin**. At the same time, the distal 1.2-million base pair (Mb) region of the dot chromosome is amplified in **polytene chromosomes**, a property of **euchromatin**, and codes for ~80 genes, a gene density similar to that found in the euchromatic arms. (For a more detailed discussion of chromatin structure, see the Chromatin page of the modENCODE Educational Supplement at **http://modencode.sciencemag.org**.) Actively-transcribed regions of DNA are typically associated with euchromatic domains, while transcriptionally-silent regions of DNA are generally associated with heterochromatic domains. Thus, understanding chromosome organization and chromatin effects on dot chromosome gene expression requires careful analysis, not just of the genes present but also of the type and distribution of repetitive elements. While this amalgamation of heterochromatic and euchromatic properties of the dot chromosome is unusual in *Drosophila*, the composition of the dot chromosome (30% repetitious DNA) is actually similar to euchromatic regions in a mammalian genome. Hence, the dot chromosome is an ideal candidate for studying the effects of chromatin packaging on regulating gene expression, and the results from comparative genomics can provide many insights into these questions.

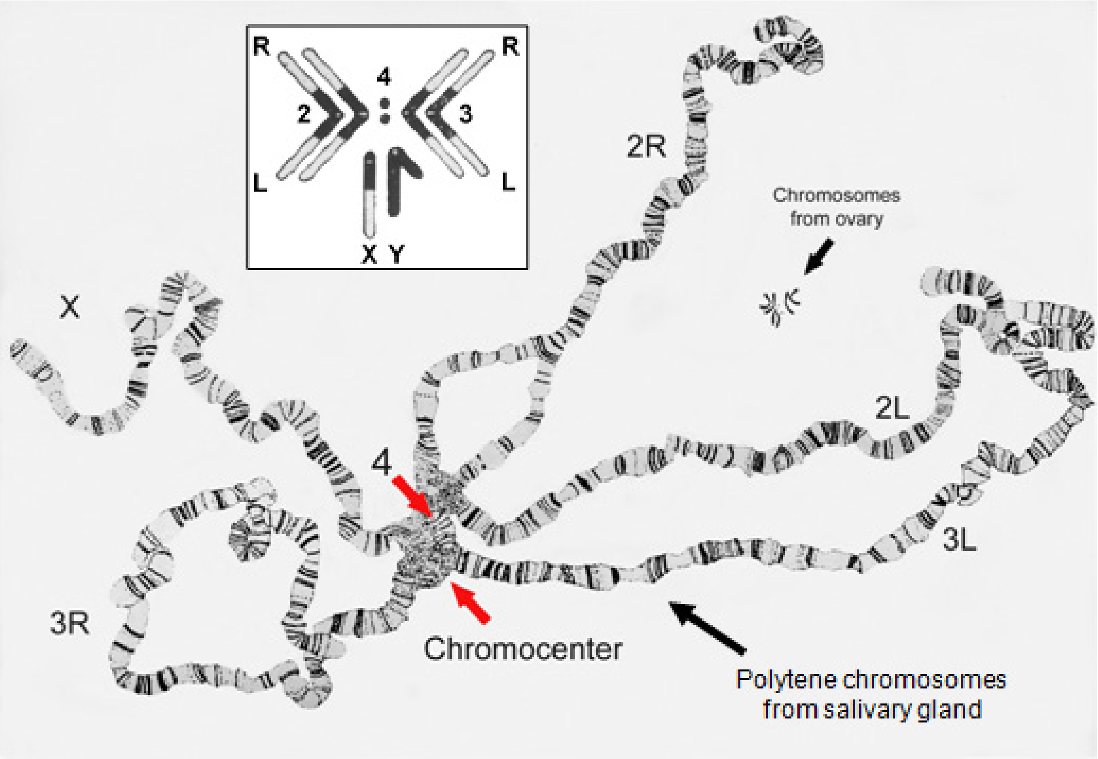


Figure 3. *Drosophila melanogaster* chromosomes (from Painter, 1934; used with permission from Oxford University Press)

Sequence improvement is usually done in the context of an entire semester course or independent research project. However, some publically-available DNA assemblies are of sufficient quality that no additional improvement is needed. In addition, fosmid DNA that is successfully improved by GEP students is then deposited into the GEP's Project Management System for future annotation. Thus, all members of the Genomics Education Partnership have the opportunity to contribute to the completion of an annotation research project, even if they do not do the sequence improvement themselves.

What is Gene Annotation?

The first complete genome sequences of a bacterium (*Haemophilus influenza*) and a eukaryote (*Saccharomyces cerevisiae*) were published in consecutive years in the mid-1990's (Fleischman et al., 1995; Goffeau et al., 1996). Since that time, hundreds of organisms have had their genomes sequenced. However, while continually-updated technologies are rapidly increasing the number of sequenced genomes, ascribing functions to various parts of the genome is nowhere near as automated. In the process of **annotation**, decisions are made as to which particular DNA sequences within a genome contain biological information. This involves identifying features in the DNA sequences and, in many cases, determining which of these **features** are likely to be **genes** (Walter and Wilkerson, 2006).

The expression of protein-coding genes in cells employs mRNA intermediates. Gene annotation in complex eukaryotes is complicated by the fact that the coding regions of most protein-coding genes are interrupted, consisting of coding **exons** and non-coding **introns**. Introns are often much longer than the exons of many genes, and they must be spliced out of the initial RNA transcript. Thus, gene annotation includes careful mapping of all the exon-intron boundaries, to create a gene model that results in the translation of a full-length polypeptide chain. Gene annotation is facilitated by computer programs, which scan the newly-sequenced DNA for **intron splice sites** and **open reading frames**; the programs may also search for **consensus sequences** of known gene regulatory regions (such as **promoters** and **enhancers**), which dictate when and where a gene is transcribed into RNA (Fig. 4). Human annotators combine these (often contradictory) computational predictions with sequence alignment and expression data (RNA-seq) to create the best-supported gene model (Walter and Wilkerson, 2006).

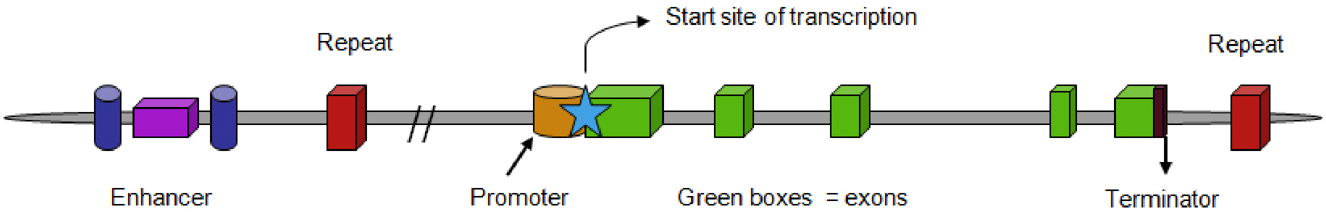


Figure 4. Diagram of a single, hypothetical eukaryotic gene. Some of the rectangles code for amino acid sequences of a protein (the exons), others contain regulatory information (promoters and enhancers), and some areas are transcribed but not translated (introns between the exons).

GEP members use a comparative genomics approach to gene annotation, comparing new genomic DNA sequences (and mRNA sequences, if available) from various *Drosophila* species to known reference DNA and mRNA sequences from *Drosophila melanogaster*. A discussion of the various rules and criteria that are used to generate gene models for the orthologous sequences follows an overview of the organization of the gene annotation labs.

Overview of Gene Annotation Labs at Amherst College

In these labs, students work in pairs to annotate one specific feature of a 40-60 kb **fosmid** of DNA from a species of *Drosophila*. The GEP annotation strategy makes extensive use of a Washington University mirror of the University of California at Santa Cruz (UCSC) Genome Browser (including custom 'evidence tracks' for GEP projects), NCBI BLAST and FlyBase, all publically-accessible Internet sites. Links to these Web sites are on the first page of *A Sample Annotation Problem* (section 3).

You will first complete an on-line activity entitled *A Simple Introduction to NCBI BLAST*, which introduces you to the National Center for Biotechnology Information (NCBI) BLAST search engine and the GEP's own Gene Record Finder Web site. In this activity, you will analyze a DNA sequence from *Drosophila yakuba*, a close relative of the widely-used model organism, *Drosophila melanogaster*.

Genomic sequences from species such as *Drosophila mojavensis* or *Drosophila grimshawi*, which are more distantly related to *Drosophila melanogaster*, are generally more difficult to annotate. However, you will next work with a highly-conserved gene from *Drosophila grimshawi* by working through *A Sample Annotation Problem*, filling in answers to questions in the worksheet as you go along. This exercise will reinforce the concept of interrupted genes in eukaryotes, which consist of coding **exons** and non-coding **introns**, and guide you through the basic annotation procedure of mapping the exon/intron boundaries. You will also be introduced to the Gene Model Checker in this second activity, so you know how to assess your own gene models.

The next scheduled lab(s) will be devoted to work on your individual annotation project, with help from the teaching assistants and the lab instructor. We will first provide an overview of the preliminary analyses of the claimed *Drosophila* fosmids from the GEP, in which the number of **features** (putative genes) in the fosmid(s), the number of **isoforms** for each feature, and the relative complexity of the different features was pre-determined using various gene predictors. Fosmids from specific species of *Drosophila* are selected for annotation based on these and additional criteria. You and your lab partner will then (by random draw!) select one or two features and/or isoforms from the pre-selected fosmids to annotate for your individual project.

Once you have passed (*to your instructor's satisfaction!*) all parts of the Gene Model Checker and the final BLAST alignment, you will then complete an Annotation Report for your mapped feature. Students who are annotating different isoforms of the same gene should meet to create one group gene report for that feature. The Annotation Reports should be completed by the end of the final lab. If time allows, each pair of students (or group if the feature has multiple isoforms) will give a few-minute overview of your feature to the rest of the lab section (with PowerPoint slides if you wish), talking a little bit about the gene and what it encodes (if it is known), the challenges you faced with the annotation, (e.g., any problems or ambiguous spots that you encountered) and anything of interest that you learned about the gene by annotating it.

Conceptual Guide to Gene Annotation

The following guide is intended to help you think about the various types of evidence you should consider as you attempt to annotate genes in *D. mojavensis*. The same considerations will also apply to species that are closer to *Drosophila melanogaster*, such as *Drosophila erecta*. However, in many of these species the level of conservation will be high enough that some of the other forms of evidence will rarely need to be considered. Your job as a gene annotator is to learn as much as you can within the annotation labs time frame and apply these skills and knowledge to come up with your best gene model.

The basic idea when attempting to create a gene model for any feature is to determine a series of base pair coordinates that describes the structure of the gene. In cases where evidence of expression (**expressed sequence tags** [ESTs], mRNA sequences, or RNA-seq data) is available, one may be able to identify the coordinates of the full-length transcript including the **5' and 3' untranslated regions**; otherwise one must focus on the protein-coding domains. The coordinates assigned in these cases would describe the base position of the beginning and end of each piece of coding sequence that together make up most of the **exons** in the final (mature) mRNA.

Your gene model must be consistent with what is known about the basic biology of transcription, mRNA splicing, and translation. For example, since it is known that RNA polymerase does not hop back and forth between the two strands of a double-stranded DNA molecule, a gene model typically does not include sequences from both strands (see McManus et al., 2010 for exceptions). It must start on one strand, continue down the length of that strand and end on the same strand. At the minimum, your gene model must include the base position of the start codon for translation, the position of the beginning and ending of each coding exon and the position of the stop codon of translation. For species sufficiently close to *Drosophila melanogaster* (e.g., *Drosophila erecta*), it might also be possible to identify the 5' and 3' untranslated regions of the mature mRNA by sequence similarity to *D. melanogaster* cDNA sequences. It may also be possible, as stated above, to identify these regions if there are expression data available for your feature. Two activities that describe how to use expression data in gene annotation *(Exercise #2: Using mRNA* and *EST Evidence in Annotation* and *Browser-Based Annotation and RNA-Seq Data)* are posted on the GEP Web site at the following link: **http://gep.wustl.edu/curriculum/course_materials_WU/annotation/annotation_exercises**.

Your first step in annotation will be to collect and consider all the evidence you can gather about the sequence you are annotating. Once you have gathered the available evidence, each piece of evidence should be weighed against all the other evidence and used to make your gene model. The goal is to make the best gene model you can that integrates all the evidence in a way that maximizes the use of high-quality evidence, avoids internal conflicts and only uses low-quality evidence when no higher-quality evidence can be found.

The types of evidence fall into three basic categories: expression, conservation and computation. **Expression** evidence is derived from sequencing RNA that has been isolated from the organism of choice. The sequences are typically mapped back to the genome to provide evidence of transcription. **Conservation** defines those types of evidence that rely on the assumption that the new species being annotated had a common ancestor with *Drosophila melanogaster*. When there is no expression evidence, conservation will be your most important evidence in constructing a gene model. Based on the principle of Occam's razor, which declares that the best explanation for anything is the one with the fewest assumptions, the best gene model in a new species is usually the model that assumes the fewest differences between it and the gene model in *Drosophila melanogaster* (i.e. the one that has the most similarity).

The third general type of evidence is **computational**. Many computer programs have been created that attempt to recognize various features in DNA sequence. Several of these programs have already been run on the sequences you will be annotating, and the results are available for viewing on various genome browsers. These programs are designed to identify evidence for a wide variety of biological features including genes, repeats, or various other features (e.g., intron/exon boundaries). Each of these programs has been optimized for its given purpose and as such, provides at least a hint as to a possible biological function of any given sequence. Without expression and conservation evidence, computational analysis is usually the only evidence you can fall back on to create your gene models.

Finally, if expression, conservation, and computational evidence fail to provide enough evidence for a given gene model, a few simple rules can be used to assist in creating a gene model. These rules are based on philosophical consideration of how best to "get things wrong" and are discussed in the Summary and outline that follows. Additional documents that describe techniques for tackling annotation projects with a range of difficulties are available on the GEP web site at **http://gep.wustl.edu/curriculum/course_materials_WU/annotation/specific_issues.**

Basic Biology

Before we consider types of annotation evidence in more detail, we will review a few details of basic molecular biology, to guide you in your generation of a gene model. While it is impossible in a short tutorial to cover all the relevant basic biology (you should already know about transcription and translation), there are a few specific details that should be discussed.

Introns

Unlike bacteria, many genes in eukaryotes have introns (Fig. 5). These sequences are removed from the primary RNA transcript based on sequences found within the intron. In cells, splicesomes recognize RNA consensus sequences to remove the introns and join the exons together, creating an mRNA with a single open reading frame. In gene annotation, it is likewise critical to find the precise beginning and end of each exon, so that the resulting gene model also encodes a full-length polypeptide. The sequence at the beginning (5' end) of the intron is called the **donor site**, while the 3' end of the intron is called the **acceptor site**. *Most eukaryotic introns begin with the nucleotide sequence 'GT' and end with the sequence 'AG' (the GT-AG rule) in the genomic DNA*. However, the larger **consensus sequences** that define intron donor and acceptor sites have a lot of tolerance for mismatches and can evolve quite quickly.

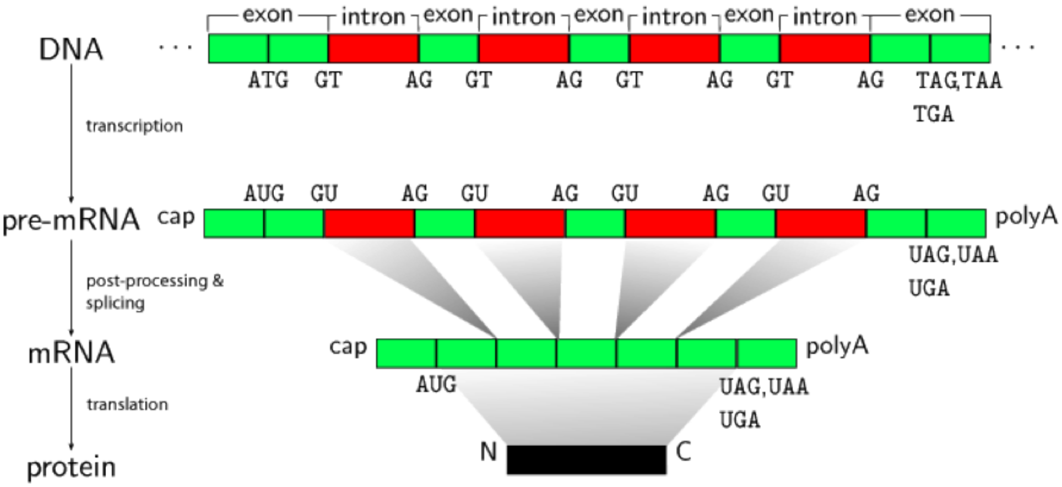


Figure 5. Eukaryotic gene structure. The donor (GT) splice site and the acceptor (AG) splice site are the first two and last two nucleotides, respectively, of the DNA sequence of each intron (in red above) (Ben-Hur et al., 2008). (Figure used with permission from S. Sonnenburg, G. Rätsch & B. Schölkopf)

For the purposes of your analysis, you can identify putative intron/exon boundaries in three ways. First, you can assume that any *de novo* gene prediction program will predict donor and acceptor sites consistent with the basic biology of splice site sequence composition, thus any gene model generated by a *de novo* gene prediction site can be used to help you pick splice sites. Second, a computer program designed to find and score putative donor and acceptor sites has been run on all GEP fosmid sequences. The results of this analysis are displayed in the GEP mirror of the UCSC Genome Browser for your sequence. To simplify the results, the potential sites predicted using this program have been split into high, medium, and low quality; sites of any quality can be used as part of your gene model. Finally and probably the least reliable way to find donor and acceptor sites is to look at the sequence by eye and scan for the base sequences known to be used by the splicing mechanism. While searching by eye is the lowest quality of evidence for the prediction of an intron/exon boundary, it is often the only evidence for a given splice site. In this case, any GT before a stop codon can be considered as a potential intron donor site, while any AG is a possible acceptor site. In all cases, conservation of exon length between the reference *Drosophila melanogaster* sequence and the sequence being annotated should be a guiding factor in choosing between possible intron splice sites (see **Conservation** section below).

In rare cases (in the range of about 1 in 100), a "non-canonical" donor site with the sequence GC (instead of the canonical GT) has been found in *Drosophila melanogaster*. Non-canonical donor sites will never be used in gene models generated by most *de novo* gene predictors; however, evidence collected so far in annotation of *Drosophila virilis* suggest that these GC donor sites are also used in a few genes in this species. Of the non-canonical sites found in *Drosophila virilis*, about 50% of the time the same non-canonical site is used in *Drosophila melanogaster*.

Finally, it should be noted that examination of a large number of introns in *Drosophila melanogaster* establishes the minimum size of an intron as 43 bases (Guo et al., 1993; Talerico and Berget, 1994). It is reasonable to assume that this limit also exists in the other *Drosophila* species. *Thus, any gene model that predicts the presence of an intron smaller than 42 bases is highly unusual and would require expression data (e.g., EST, RNA-seq) to support the hypothesis*.

mRNA structure

Once the start codon, the stop codon, and all the intron/exon boundaries have been identified, you will use the GEP's on-line Gene Model Checker to create the final coding sequence of the mRNA as well as the predicted amino acid sequence of the encoded protein. Remember that in eukaryotes, each processed, mature mRNA contains a single open reading frame that extends from the start codon through all the internal exons and ends with a stop codon. Your gene model should likewise produce a putative message that contains a single long open reading frame with no internal stop codons. *If the Gene Model Checker report shows that your gene model has internal stop codons, you should double check and adjust your intron/exon boundaries until no internal stop codons are found. Although stop codon read-through has been observed in* Drosophila*, such cases are rare, and if suggested, need to be accompanied by mRNA and protein evidence.*

Expression

Evidence of expression from cDNA/EST/RNA-Seq tracks can provide strong evidence for the locations of intron/exon boundaries. However, there are important factors to consider when evaluating expression data. First, some genes are expressed at very low levels or only for very short periods of time. As such, the lack of expression data for any gene should not be considered strong evidence in favor of discounting the presence of a gene. No matter how much expression data are collected, there will always be real genes that lack expression evidence. Another issue to be aware of is the high level of noise found in RNA-seq data. For example, we have observed a non-trivial amount of variation in the apparent splice sites in RNA-seq data that have been mapped across introns (i.e., TopHat track). In situations like this where there is a lot of variation in the evidence for the exact position of a splice donor/acceptor site, we will follow a "majority rules" approach.

Conservation

Conservation can take many forms and all of the following should be considered when generating a gene model. They are presented here in order of importance with the most important first:

Conservation of primary amino acid sequence

This is certainly one of the most important forms of evidence to guide you in construction of your gene models. It is reasonable to assume that for almost any gene found in the various *Drosophila* species, the encoded protein is serving a very similar if not identical function in the different species (i.e. the proteins are serving as functional **orthologs**). As such, one would expect the amino acid sequences to be very similar. Your use of programs that search for similarity to identify regions of similar amino acid sequence will be the foundation upon which you will build your gene models. Conservation of this type is found using computer programs like BLAST and Clustal. When conservation between the two species is very high, the identification of intron/exon boundaries is easy since the boundary will be very close to the end of the alignment. As the extent of amino acid conservation decreases, the identification of intron/exon boundaries will need to rely on other evidence discussed below. See the *Advanced Annotation Instruction* sheet for more on Clustal analysis and what to do if the default BLAST search fails.

Conservation of gene structure

The creation or removal of an intron in orthologous genes is a very rare event, even over evolutionary time scales. This means that the best gene model for your *Drosophila* isoform will almost always have the same basic structure (i.e., the same number of introns and exons) as the gene model in *Drosophila melanogaster*. This rule however is not absolute; sometimes the only gene model that fits most of the evidence has a new or missing intron, or fuses two introns into one, so if you can find no way to construct a gene model that maintains the same number of exons, go with a gene model that keeps the total number of exons as close to *Drosophila melanogaster* as possible.

Conservation of exon length

In a surprisingly-large number of cases, we find exons that have a very similar length even when there is no detectable conservation of the encoded amino acids found near the intron/exon boundary. This has happened enough that we can come to consider more carefully any putative donor or acceptor sites that conserve exon length. For example, consider the following alignment in which BLASTX was used to find similarity between a piece of *Drosophila virilis* genomic DNA sequence and the sequence of a 45 amino acid long exon from *Drosophila melanogaster*.

```
Exon sequence:

   1  CGSVVPSADYAYSPAYTQYGGTYGSYSYGTSSGLIYNPAS
  41  GPITT

BLAST alignment:
  Query: 14253    CGSVVPSADYAYSPAYTQYGGTYGSYSYGTSSGLI    14357
                  C SVVP +DYAY+PAYTQYGG YGSY YGT SGLI
  Sbjct:     1    CSSVVPGSDYAYNPAYTQYGGAYGSYGYGTGSGLI    35
```

In this case, we can see that the alignment starts out well (amino acid 1 of the exon is aligning to some sequence in the genome) but ends before the end of the exon, we are missing the last 10 amino acids (remember that BLAST only gives a **local** alignment; that is, it does not report sequences that do not have significant similarity). In cases like this, we would concentrate our search for donor sites around base 14387 (30 bases or 10 codons down from the end of the alignment). While any donor downstream of 14357 (the end of the above alignment) would be potential candidate, donor sites found near 14387 would be strong candidates for use in the final gene model, especially if they have a high score on the donor site detection algorithm (see following).

Computational evidence

While conservation is in most cases the best evidence for constructing your gene model, you will not always have sufficient similarity to construct a viable gene model. There are also cases in which conservation will give support for several different gene models with no way to pick among the consistent models. In these cases, computational evidence is your next best source. The best approach is to rely on conservation as much as possible and adjust your models based on the computational evidence. There are two main sources for information you will want to consider as you try and determine the best gene model, splice site prediction programs and *ab initio* gene finders.

Splice site prediction program

Several of the information tracks available to you on the UCSC Genome Browser show the results of the splice site prediction program GeneSplicer. The output of this program tags potential splice donor and acceptor sites and gives them a score of between -10 and +10. In order to simplify the output, we have classified those sites with scores above 7 as high quality, scores between 0 and 7 as medium quality and scores between -10 and 0 as low quality. In general, this information can be used to help you pick donor/acceptor sites when there is no conservation. For the purposes of the GEP project, you should always pick a donor/acceptor site that maintains the open reading frame and maximizes conserved amino acids; however when there is little or no conservation or there are two or more possible donor/acceptor sites very close together, sites tagged by GeneSplicer are better candidates than sites that are not tagged, and in general, the higher the score the better.

Ab initio *gene prediction algorithms*

The creation and optimization of *ab initio* gene finders is an active field of study and, as such, many different programs are available to create gene prediction sets. Many of these have been run on the section of DNA that you will be working on. The results of these analyses are available on the Genome Browser for your section of DNA. While each program has its strengths and weaknesses, for the purposes of gene model creation (selection of intron/exon boundaries) they should be considered of equal quality. The most common usage of the information created here is a majority rules/vote system. Failing any evidence from basic biology, conservation or other algorithms, the splice site that was picked by the most different programs would be picked as the donor/acceptor site.

Last and certainly least

It is certainly possible that you may run across situations where you will have ambiguous evidence and must choose between a small number of consistent choices with no evidence to help you decide (this is often the case when using the conservation of exon length rule). In these cases, when all else fails, the policy is to go with the choice that creates the largest protein. The reason for this is that it is better to add a few extra amino acids to a protein than to have a few amino acids missing. This is because if the amino acids are missing there is no way to find them in a BLAST search, but BLAST is fairly tolerant of having a few extra amino acids tucked inside an alignment. Thus it is best to err by including extra amino acids rather than missing amino acids.

It is also possible that you will run across situations where there may be only very weak evidence for one gene model over another, yet the weaker model gives a longer protein. To balance these decisions, the GEP has set a policy for the use of the computational donor/acceptor sites when picking your gene model. In general, when picking among a group of consistent intron/exon boundaries, choose the longest exon that has a boundary no more than one step (low, medium, high) worse than a boundary that creates a shorter exon. In other words, when two choices differ by two steps, go with the higher valued boundary (longer unlabelled vs. shorter medium scoring, pick the medium candidate; longer low scoring vs. shorter high scoring, pick the highscoring candidate). Alternatively, when two choices differ only by one step (unlabelled vs. low, low vs. medium, and medium vs. high) pick the boundary that gives the longer protein.

Summary

The following is a list of the important rules for annotation based on the above discussion. All models should follow these rules as much as possible. The rules are listed in the order of importance: the best model will follow a rule higher on the list at the expense of a lower rule. Most models will not follow all rules; it is your job as an annotator to create the best model in spite of this. For example, you may need to decide between a model that follows many less important rules but breaks a single more important rule and a model that follows a more important rule over the less important rules. This balancing act is where human ability far exceeds computers. Rules are ranked into four classes:

Inviolate rules – rules for which counter examples are almost never seen. Clear and convincing wet bench experimental evidence would be required to convince scientists that this rule should not apply to your model. Since no wet bench work is being done for these projects, these rules should never be broken. However, you can make a note of your suspicions in the annotation report.

Important rules – rules for which exceptions are only rarely seen. You may choose to make a model that does not follow this rule but you must note in your annotation report that this rule was not followed and document why you have decided not to follow this rule.

Basic rules – these are rules or observations that are seen more often than not but are also not followed in a significant number of models. You should make models that follow these rules if you can but be careful not to ignore more important rules just to follow these rules. You do not need to document that you did not follow rules of this type.

Tie-breaking rules – rules to help make models when all the more important rules do not help. You may wish to note the use of these rules in your annotation report to help those reviewing your annotations understand why you picked the model you did.

Refer frequently to the one page-summary on the next page while you annotate and create your gene model.

Rules for Gene Annotation

Inviolate rules (In Basic Biology):

1. CDS of gene must begin with ATG and end with a stop codon; no internal, in-frame stop codons.
2. Exons are found in order along the source DNA.
3. The last two bases of an intron sequence must be AG.
4. Intron sequences should be at least 42 nt.

Important rules:

In Basic Biology:

5. In most cases, an intron sequence should begin with GT. GC use is rare but can be selected if use of GT breaks inviolate or other important rules.
6. Use data in RNA-Seq tracks (both mapped reads and TopHat splice-site junctions) if available.

In Conservation:

7. Conserved amino acids identified by single exon BLAST shown in high quality alignments (i.e. high % identity and properly placed) should be included in exons.
8. The number of exons between informant and new species should be conserved.
9. The organization of exons to generate the various *D. melanogaster* isoforms should be conserved. Note: some genes have alternate splice sites for one or more exons and a gene model must be created for each unique isoform.

Basic rules:

In Conservation:

10. Identification of conservation should be done in the following order, (based on speed, sensitivity, ease of use, and specificity).
    a. Protein-DNA BLASTx or tBLASTn with increasing expect thresholds
    b. DNA-DNA CLUSTALW alignment
    c. DNA-DNA BLASTn using very large E-score cutoff values (e.g. 1e10)
11. Failing identification of exons by expression or conservation alignments as above, the highly-conserved regions identified in the “Comparative Genomics” (multiz) tracks should be checked.
12. Attempt to conserve exon length even if the specific amino acids are not conserved.
13. If it is difficult to identify exons based on the above, repeat the whole process using the gene models generated by the GEP in a close neighbor species as your assigned species instead of the model found in *Drosophila melanogaster*.

In Computation:

14. Exons that cannot be found by any type of conservation may be identified using predictions from *ab intio* gene finders

Tie-Breaking rules (In Basic Biology):

15. Longer exons are better; include more amino acids in the exons between the start and stop codons.

Glossary of Terms

***ab initio* gene finding**: process whereby genomic DNA sequence alone is systematically searched for certain tell-tale signs of protein-coding genes (as opposed to experimental evidence in the form of an identified mRNA or protein molecule encoded by that DNA sequence.)

Chromatin: the DNA-protein complex found in eukaryotic chromosomes

***Euchromatin**: chromatin that is diffuse and non-staining during interphase; may be transcribed*

Heterochromatin: chromatin that retains its tight packaging during interphase; often not transcribed

Coding DNA Sequences (CDS): the subset of exon sequences that are translated into protein

cDNA (complementary DNA): a DNA copy of an mRNA molecule, manufactured using the enzyme **reverse transcriptase**

Consensus sequence: the most common nucleotide (or amino acid) at a particular position after multiple, related sequences are aligned and similar functional sequence motifs are found. Example: transcription factors that recognize particular patterns in the promoters of the genes they regulate

Exons: gene sequences that are transcribed into RNA and are present in the mature (spliced) mRNA molecule

Expressed Sequence Tag (EST): partial, single (i.e., one shot) sequence read of a cDNA molecule. The sequence is of relatively-low quality, usually 500 to 800 nucleotides in length.

Feature: any region of defined structure/sequence in a genomic fragment of DNA. Features would include genes, pseudogenes and repetitive elements. Most people are interested in identifying the protein-encoding genes.

Fosmid: a hybrid plasmid cloning vector, which has been manufactured to accept DNA inserts of ~40,000 base pairs (bp) [normal plasmids are able to carry only 1-20 kb]; usually propagated in *E. coli*, which can each only contain one fosmid

Homologous genes: genes that have similar sequences because they are evolutionarily related; there are two different types of homology

Orthologs: related genes in different species, which are derived from the same gene in a common ancestral species

***Paralogs**:* related genes within a species, which have arisen by a duplication event

Introns: gene sequences that are transcribed into RNA but are removed during splicing

Isoform: any of several different forms of the same protein. Isoforms may be produced from different alleles of the same gene, from the same gene by alternative splicing, or may come from closely-related genes.

Open Reading Frame (ORF): a segment of the genome that potentially codes for a polypeptide chain (or part of a polypeptide chain); the part of an mRNA molecule that potentially codes for a polypeptide. ORFs are located between the start codon (ATG; AUG in the mRNA) and a stop codon (TAA, TAG, TGA; UAA, UAG, UGA in the mRNA) for translation. One must sort through six possible reading frames for any putative protein-encoding genomic DNA fragment: three on one strand of DNA heading in the 3’ direction (each shifted by one base with respect to the previous reading frame) and three on the other strand of DNA heading in the opposite direction (Fig. 6).

Polytene chromosomes: giant chromosomes that form when multiple rounds of DNA replication occur and the sister chromatids remain attached to and aligned with each other

5’ Untranslated Region (5’ UTR): mRNA sequence between the 5’ cap and the start codon for translation

3’ Untranslated Region (3’ UTR): mRNA sequence between the stop codon for translation and the poly A tail

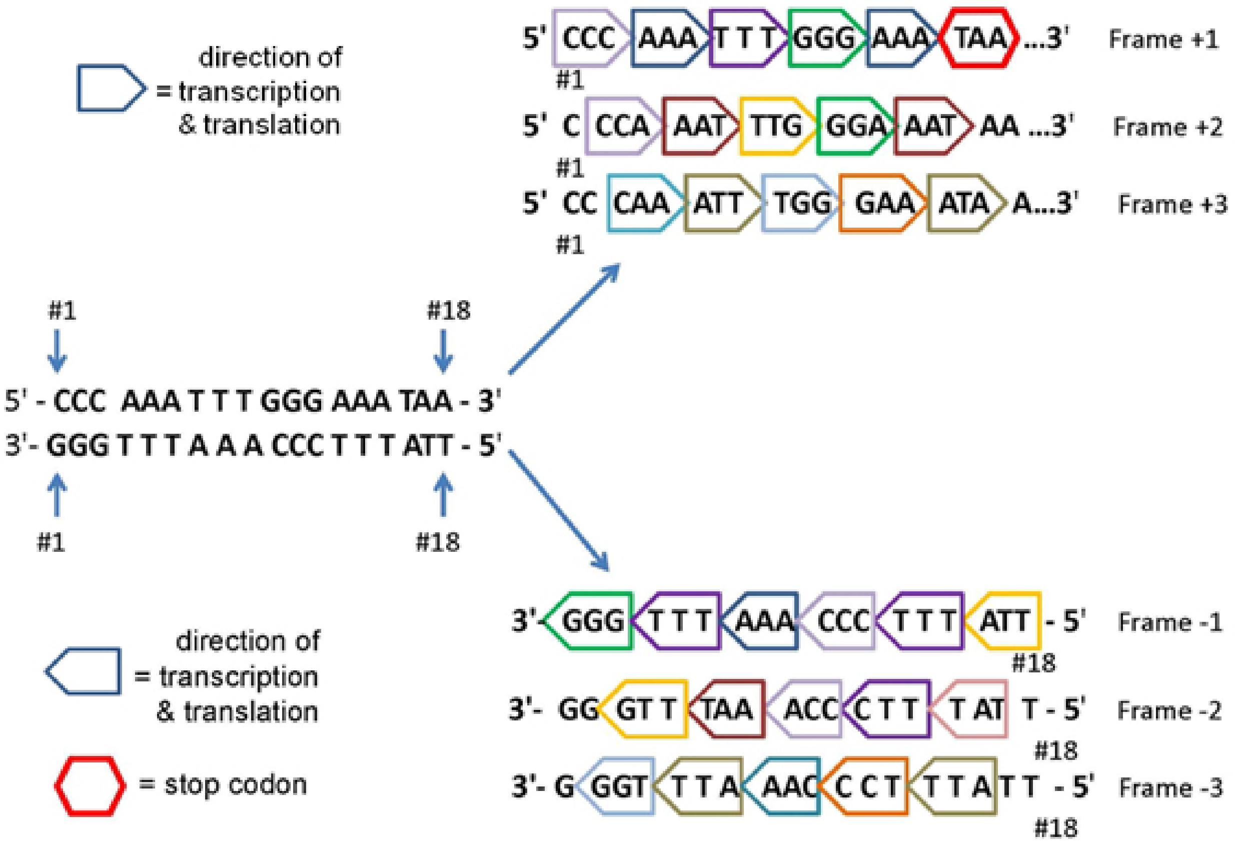

direction of
= transcription
& translation
5' CCC AAA T T T GGG AAA TAA ...3' Frame +1
#1
5' C CCA AAT TTG GGA AAT AA ...3' Frame +2
#1
5' CC CAA ATT TGG GAA ATA A...3' Frame +3
#1
#1
#18
5' - CCC AAA T T T GGG AAA TAA - 3'
3'- GGG T T T A A A CCC T T T ATT - 5'
#1
#18
3' GGG T T T AAA CCC T T T ATT - 5' Frame -1
#18
3'- GG GT T TAA ACC C T T T AT T - 5' Frame -2
#18
3'- G GGT T T A AAC C C T T T A T T - 5' Frame -3
#18
direction of
= transcription
& translation
= stop codon

Figure 6. There are six possible reading frames for any region of DNA. The colored block arrows mark the codons in each of the six reading frames. The different colors of the block arrows represent different amino acids encoded by the various codons. Note that the amino acid sequence for each of the reading frames is different. Note also the stop codon in reading frame +1.

A Sample Annotation Problem

Prerequisites

1. BLAST Exercise: *A Simple Introduction to NCBI BLAST*, available at

   http://gep.wustl.edu/curriculum/course_materials_WU/annotation/tutorials_and_walkthroughs

2. Familiarity with concepts in the Student Outline

Resources

1. The BLAST Web server, available at **http://blast.ncbi.nlm.nih.gov/Blast.cgi**
2. FlyBase, available at **http://flybase.org/**
3. The GEP UCSC Genome Browser Mirror, available at **http://gep.wustl.edu/** under

   ‘Projects’ -> ‘Annotation Resources’

4. The Gene Record Finder, available at **http://gep.wustl.edu/** under ‘Projects’ ->‘Annotation Resources’.

   The Gene Record Finder relies on the D. melanogaster gene annotations published by FlyBase. Because FlyBase releases updates to the D. melanogaster gene annotations approximately every 8 weeks, the screenshots and accession numbers described in this document might differ from the data available in more recent FlyBase releases. Thus, the GEP has setup a separate version of the Gene Record Finder that corresponds to the FlyBase release used to develop this document. This version of the Gene Record Finder is available at: **http://gander.wustl.edu/~wilson/dmelgenerecord_able/**.

Important Note

Mozilla Firefox is the Internet browser of choice in the teaching labs at Amherst College. However, other Internet browsers work equally well with all of the above databases and on-line search engines. It works best to open each of the above Web sites in a separate window (NOT tabs), so that you can have them all visible at one time on your monitor. You will also need an open Word document file, into which you will paste your BLAST alignments and other information.

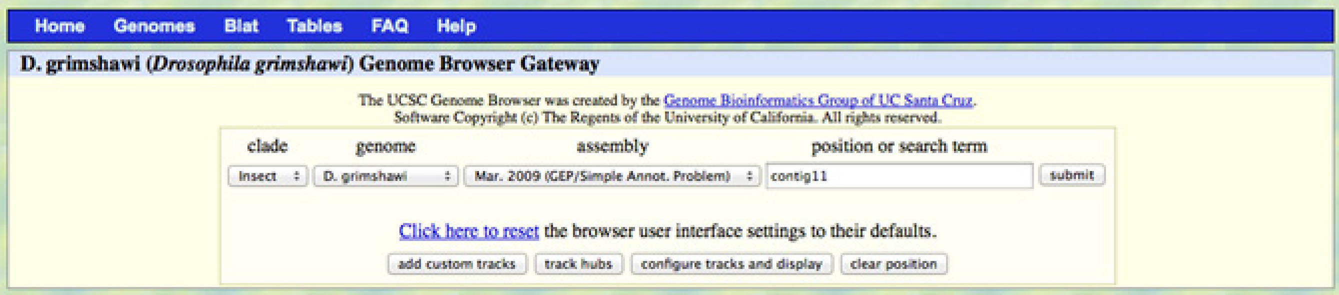
Introduction

This worksheet will guide you through a series of basic steps that have been found to work well for annotation of species closely related to *Drosophila melanogaster*. It provides a technique that can also be the foundation of annotation in other more divergent species, but in those cases other special techniques will probably be needed. The given example uses a gene from the fourth (dot) chromosome of *Drosophila grimshawi*. Although the *D. grimshawi* and *D. melanogaster* lineages diverged over 30 million years ago, the gene that you will examine in this activity is highly conserved between the two species. Thus, annotation of this particular *D. grimshawi* gene is relatively straightforward. Students should consult the *Advanced Annotation Instruction Sheet* for additional strategies in tackling the annotation of genes with lower levels of similarity.

While this worksheet will do some click-by-click guidance, some familiarity with the NCBI Blast pages and the UCSC Genome Browser Mirror is assumed. Those users completely unfamiliar with these sites may wish to familiarize themselves with the use of these pages before attempting to use them in this annotation exercise. There are online tutorials and user guides available on the use of the Genome Browser at **http://genome.ucsc.edu/training.html**. BLAST training is available online at the NCBI web page and documentation for GEP programs is available on the GEP Web site (under Help -> Documentations -> Web Framework, then click on or scroll to the 'Guide to Other GEP Tools' section and choose the appropriate guide). *Note: As you work through this exercise on a computer, the screenshots you get may not exactly match the images shown in the figures since Web sites are frequently updated.*

Identifying the Ortholog

The first step in annotating a potential feature is to identify the *D. melanogaster* ortholog. We can start by examining the UCSC Genome Browser on the GEP website. Go to **http://gep.wustl.edu/**, select the **Projects** drop down menu, then 'Annotation Resources' and the 'GEP UCSC Browser Mirror.' Click on 'Genome Browser' in the left-hand column of the next window. Then, select "*D. grimshawi*" genome, "Mar. 2009 (GEP/Simple Annot. Problem)" assembly, enter contig11 in the position box and click "submit" (Fig. 7a.)



D. grimshawi (*Drosophila grimshawi*) Genome Browser Gateway

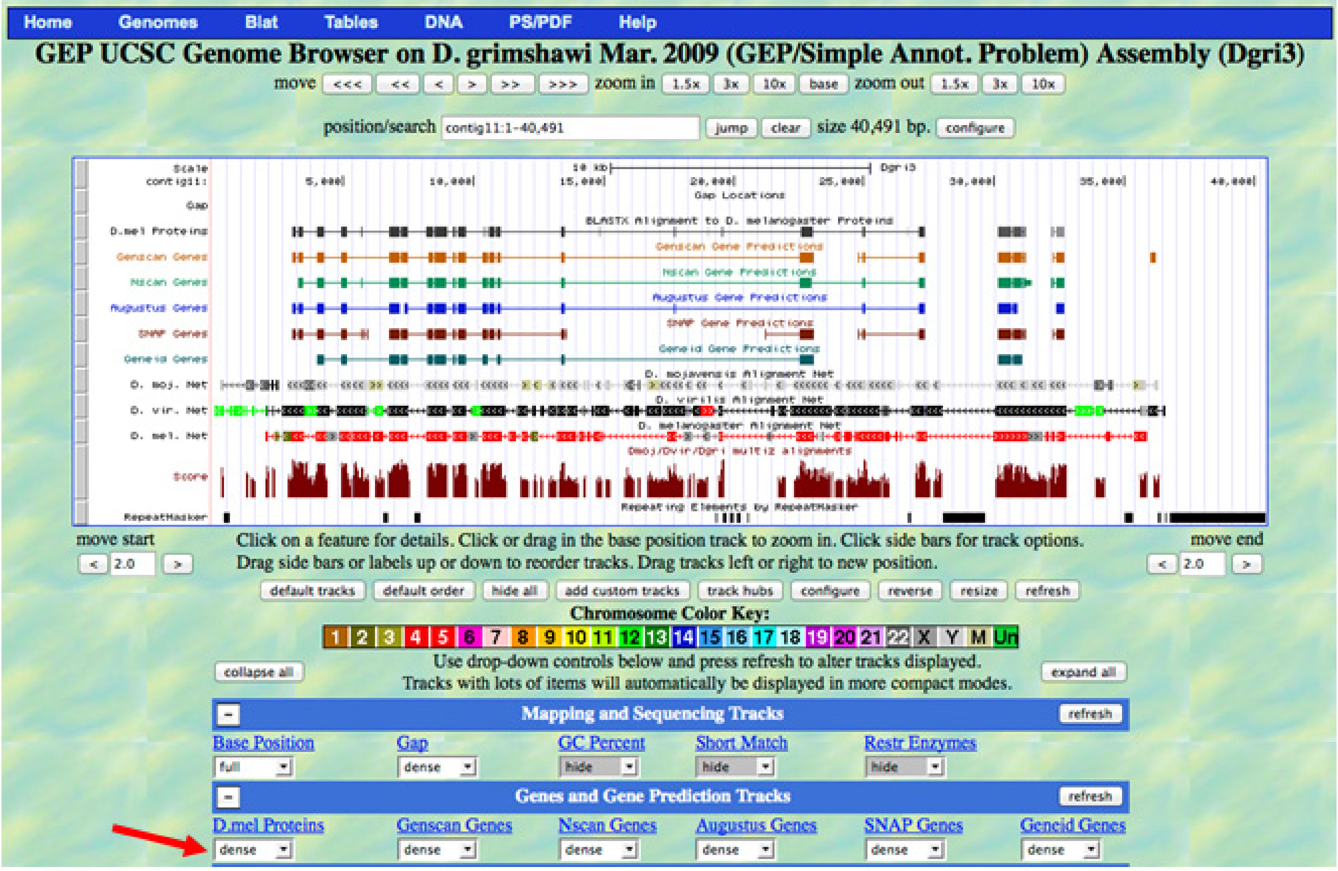
The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term
Insect	D. grimshawi	Mar. 2009 (GEP/Simple Annot. Problem)	contig11

Click here to reset the browser user interface settings to their defaults.

Figure 7a. GEP UCSC Genome Browser mirror. Note that the fosmid name and/or number in the above window (in the 'position or search term' box) will change depending upon which fosmid is selected for study. The position field will also change as you navigate to different parts of the contig. For example, "contig11:1-40,491" indicates that the genome browser is displaying the first 40,491 base pairs (i.e., the entire length) of contig11.

Do not be scared by the next window that appears, which is pretty complicated (Fig. 7ba)



GEP UCSC Genome Browser on D. grimshawi Mar. 2009 (GEP/Simple Annot. Problem) Assembly (Dgri3)

position/search contig11:1-40,491 size 40,491 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.
Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

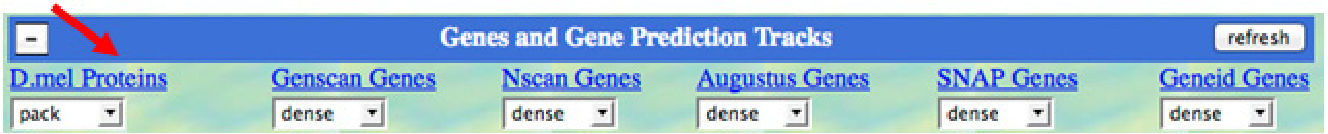
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks: Base Position full; Gap dense; GC Percent hide; Short Match hide; Restr Enzymes hide

Genes and Gene Prediction Tracks: D.mel Proteins dense; Genscan Genes dense; Nscan Genes dense; Augustus Genes dense; SNAP Genes dense; Geneid Genes dense

Figure 7b. The GEP UCSC Genome Browser view for contig11 in *Drosophila grimshawi*

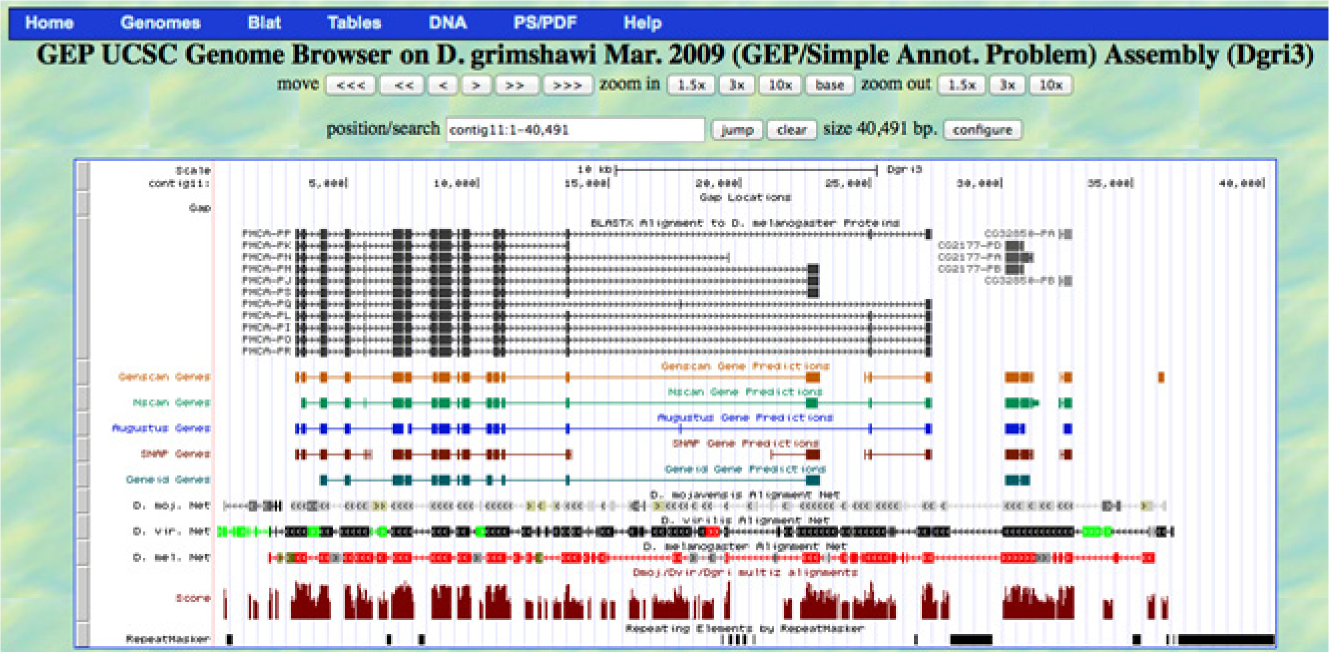
Each colored entry in the top white box represents a different "track" generated by a different sequence analysis software program. The names (on the left) of several different gene predictor tracks are repeated in the second row of menu bars (Genes and Gene Prediction Tracks). When 'dense' is selected below the name of each track (see red arrow above), the track in the image above simply indicates that the software analysis has been run. However, to see the specific data that were generated by that software program, you need to adjust the tracks to change how or what data are displayed. To adjust the settings of any track, click on the title of the track in the bottom section of the page to be taken to a page that contains more information on the track and tools for adjusting any settings. To set the D. mel Proteins track to match the images in Fig. 9, click on the title "*D. mel Proteins*" (Fig. 8).



Genes and Gene Prediction Tracks: D.mel Proteins pack; Genscan Genes dense; Nscan Genes dense; Augustus Genes dense; SNAP Genes dense; Geneid Genes dense

Figure 8. Click on any title to be taken to a page about that track

On the page that comes up next, select "pack" on the menu next to "Display mode" and enter **400** in the box labeled "Show only items with score at or above:" Now click the 'Submit' button. This will take you back to the main browser image where the track has been expanded to give more detailed information (Fig. 9). You can set any of the other gene prediction tracks to 'pack' using the drop down menu under each name, then hit 'refresh' in the blue bar above.



GEP UCSC Genome Browser on D. grimshawi Mar. 2009 (GEP/Simple Annot. Problem) Assembly (Dgri3)

position/search contig11:1-40,491 size 40,491 bp.

Figure 9. Expanded gene prediction track for BLASTX alignment

The “BLASTX alignment to *D. melanogaster* Proteins” track shows the location of all the BLAST alignments that resulted from using the translated genomic DNA sequence (contig11) to search the database of all known *D. melanogaster* proteins. It is used to demarcate regions in the *D. grimshawi* contig that have sequence similarity to proteins in *D. melanogaster*. You may find when working with DNA from other species that this track is replaced by a track called “Refseq genes”. These two tracks are both used to show regions of similarity with *D. melanogaster* proteins but use slightly different search algorithms. For the purposes of annotation the differences are trivial and unimportant. *For either track, there are a few things to note.*

1. The black boxes represent regions of sequence similarity between the translated *D. grimshawi* sequence and some known *D. melanogaster* protein-coding region. The longer the black block, the longer the region of sequence similarity.
2. If more than one alignment block comes from the same protein, the blocks are connected with a thin, horizontal line covered with arrowheads. The thin lines are usually intron sequences in the *D. grimshawi* gene, which are not present in mature mRNA molecules.
3. Sometimes these alignment tracks show directionality (i.e., which DNA strand contains the coding information for the protein), as the arrowheads point to either the right or the left. Be aware that the arrow showing directionality is in general unreliable, especially in cases where the matches are DNA to DNA. If directionality is important, it should be confirmed by other methods (e.g. BLAST).
4. Note that the alignment blocks are only showing regions of high similarity. These tracks will only mark an entire exon if conservation extends across the whole exon. Thus, the extent of the alignment will mostly depend on how closely related the species is to *D. melanogaster*. Therefore, it is not possible to infer gene structure (number and placement of exons) based on these tracks, since exons with little or no conservation may be missed while large exons with multiple conserved domains may be broken into multiple smaller alignment blocks.
5. If multiple isoforms exist for a BLASTX feature, this is represented by a unique letter at the end of their names.

From the BLASTX alignment, it appears that this region of *D. grimshawi* genomic DNA has three features that show sequence similarity to putative *D. melanogaster* protein-coding genes. From the above screenshot, one can see that the symbols for these three genes are *PMCA, CG2177* and *CG32850*, respectively. If one examines the different gene predictor tracks below the BLASTX track, multiple ‘hits’ appear in their tracks on the Genome Browser window (numbered in the physical order, from left to right, that they appear along the DNA sequence of the fosmid), which show where the putative genes are as predicted by each software package. Note the differences between the gene predictors!

Question 1
Which gene predictor matches the best to the BLASTX output?

Question 2
What D. melanogaster *protein-coding gene in the BLASTX track appears to match well to the left-most* D. grimshawi *feature, as predicted by Genscan or Nscan? Write the symbol for this gene below.*

Type this gene symbol (case sensitive) into the **http://flybase.org/** *Quick Search box. What is the full name of this gene?*

Which chromosome is it on in D. melanogaster*?*

Why does the fact that the D. melanogaster *gene is on this particular chromosome (and not on a different one) strengthen the case for this gene being an ortholog to the* D. grimshawi *gene?*

Now that we are done examining the BLASTX track in detail, you may wish to change the setting back to "dense" to simplify the display. While the 'BLASTX alignment to *D. melanogaster* Proteins' track provides us with the identity of the probable ortholog for this gene region in *D. grimshawi*, this should be confirmed with a blastp search using FlyBase, as below.

Obtain the predicted amino acid sequence of contig 11.001.1 in the Nscan Gene Predictions track by (1) making sure 'pack' is selected, then (2) clicking directly on any of the predicted exons for this contig (medium green in color). Click on the 'Predicted Protein' sequence link in the next window. Now, copy this protein sequence and use it to search a database of all *D. melanogaster* proteins using the search engine at **http://flybase.org/blast**. Select the 'Annotated Proteins (AA)' database and the 'blastp: AA-> AA' program.

Question 3
Examine the list of the top 25 hits. How do the Scores and E-values of the PMCA isoforms compare to the other hits?

If you click on the various alignment scores, the window will jump to the actual alignment for each gene as compared to the genomic DNA from *D. grimshawi*. Note how much better and longer the alignments are with the *PMCA* isoforms than with the other hits. Thus, we will proceed with our analysis with the assumption that this region does indeed contain the *D. grimshawi* ortholog for the *PMCA* gene of *Drosophila melanogaster* and not any of the other genes identified by this BLAST search.

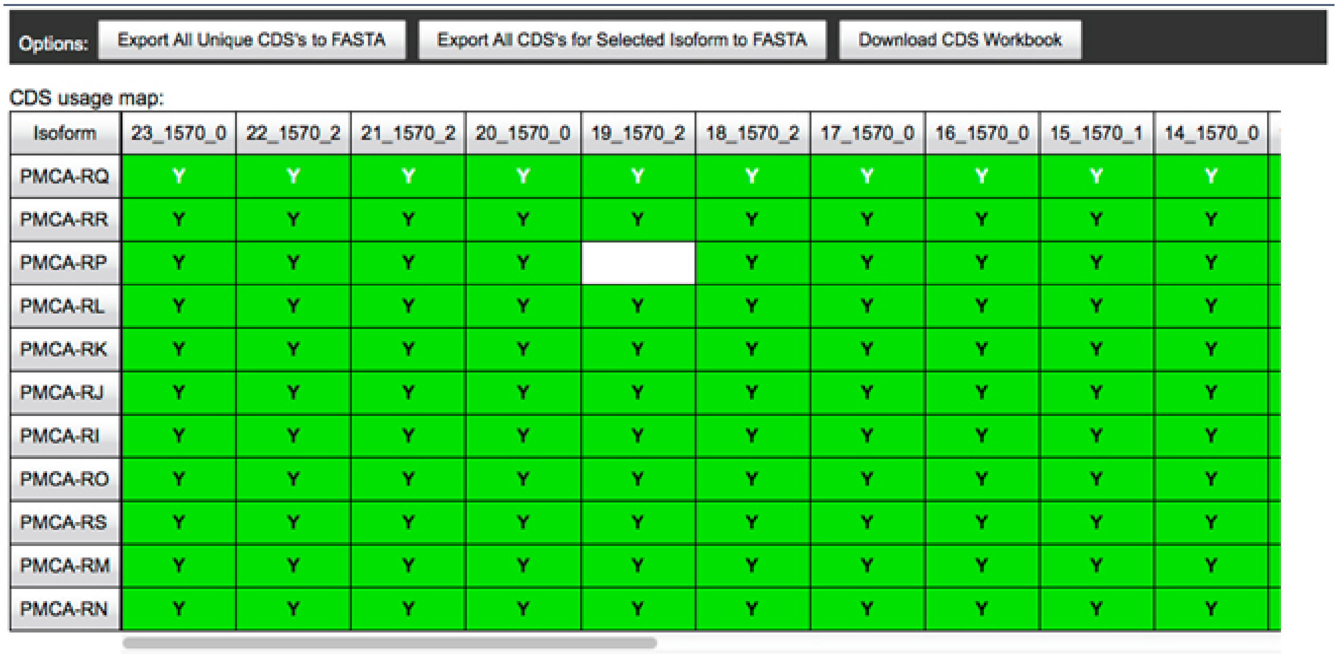
Coding DNA Sequence (CDS) by CDS Searches

While it is true that the Nscan prediction has significant matches to *PMCA*, that does not mean that the prediction is in anyway "correct". In fact, published accuracy rates for most *ab initio* gene prediction algorithms are in the range of 20-30%. It is therefore more likely than not that the Nscan prediction is actually wrong (i.e. not perfect). Without detailed analysis of the alignment between the contig 11.001.1 Nscan prediction and *PMCA*, we have no way of knowing. Common errors generated by Nscan and the other *ab initio* gene prediction algorithms are skipped exons and errors involving the ends of the gene (split genes, fused two genes, etc.). For now, all we know is that BLAST has aligned at least some of contig 11.001.1 with at least some of *PMCA* and the total sum of all the alignments gives us a good E value. Therefore, the similarity searches have convinced us that the *PMCA* ortholog is probably in the fosmid DNA somewhere and it overlaps with the Nscan prediction to a large enough extent to give a very good BLAST E value. The next step then is to use BLAST searches to find the best matches to the individual *D. melanogaster PMCA* coding sequences (CDS), as these matches will be the best evidence we can gather as to the structure of this gene in *D. grimshawi*. By doing individual CDS-by-CDS searches, we avoid confusion that often comes from the tendency of BLAST to overextend the alignment beyond the CDS boundary.

In order to do CDS-by-CDS searches, we will need the sequence of each CDS. This information is most easily obtained from the “Gene Record Finder”. This can be found in the ‘Projects’ drop down menu on the **http://gep.wustl.edu** website, then select ‘Annotation Resources’ to get to the Gene Record Finder. (Note: an on-line user guide for the Gene Record Finder can be accessed at the GEP Website at **http://gep.wustl.edu/repository/documentations/gene_record_finder_guide.pdf.**) Enter ‘PMCA’ into the search box (case sensitive) to obtain the information on this gene. As you type, note that the name of the gene appears in a drop-down bar below the search box, with the number of mRNA isoforms (11) and the total number of exons (26!) and CDS (20) for this gene. Now, click the “Find Record” button.

As you saw in the previous BLAST activity, there are three regions of information in the next window that appears: Gene Details, mRNA Details and Transcript Details/Polypeptide Details. Note that the names of the isoforms have changed from PMCA-PO (in the previous database windows) to PMCA-RO, etc. *The ‘R’ in the Gene Record Finder has simply been substituted for the ‘P’ in these other databases, and you can ignore this difference*. The term CDS refers to DNA sequences that are transcribed, present in a mature mRNA molecule *and* code for amino acids. In most cases, the term CDS can be used synonymously with ‘exon.’ The exceptions are those exons that include 5’ or 3’ untranslated regions, which do not code for amino acids. Also, due to alternative splicing, not all exons of a gene may appear in a particular mature mRNA. Hence, the windows for Transcript Details and Polypeptide Details reflect these situations and are not the same, with the numbered CDS’s referring to the specific DNA sequences that are translated in a particular mRNA molecule.

If you click on the Transcript Details tab, you will see from the exon usage chart that next appears that there are eleven different mRNA isoforms for this gene. However, when there are differences only in the 5’ and/or 3’ untranslated regions, different mRNA isoforms may in fact code for the same polypeptide. Click on ‘Polypeptide Details’ and examine the different polypeptide isoforms (Fig. 10), by moving the scroll bar at the bottom of the CDS usage map from left to right.


Options:
Export All Unique CDS's to FASTA
Export All CDS's for Selected Isoform to FASTA
Download CDS Workbook

CDS usage map:

Isoform	23_1570_0	22_1570_2	21_1570_2	20_1570_0	19_1570_2	18_1570_2	17_1570_0	16_1570_0	15_1570_1	14_1570_0
PMCA-RQ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RP	Y	Y	Y	Y		Y	Y	Y	Y	Y
PMCA-RL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RK	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RJ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RO	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RM	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RN	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Figure 10. “Polypeptide Details” for the eleven mRNA isoforms of PMCA (left side of screen)

Question 4

Hint: Examine both sides of the Polypeptide Details window to answer these questions.

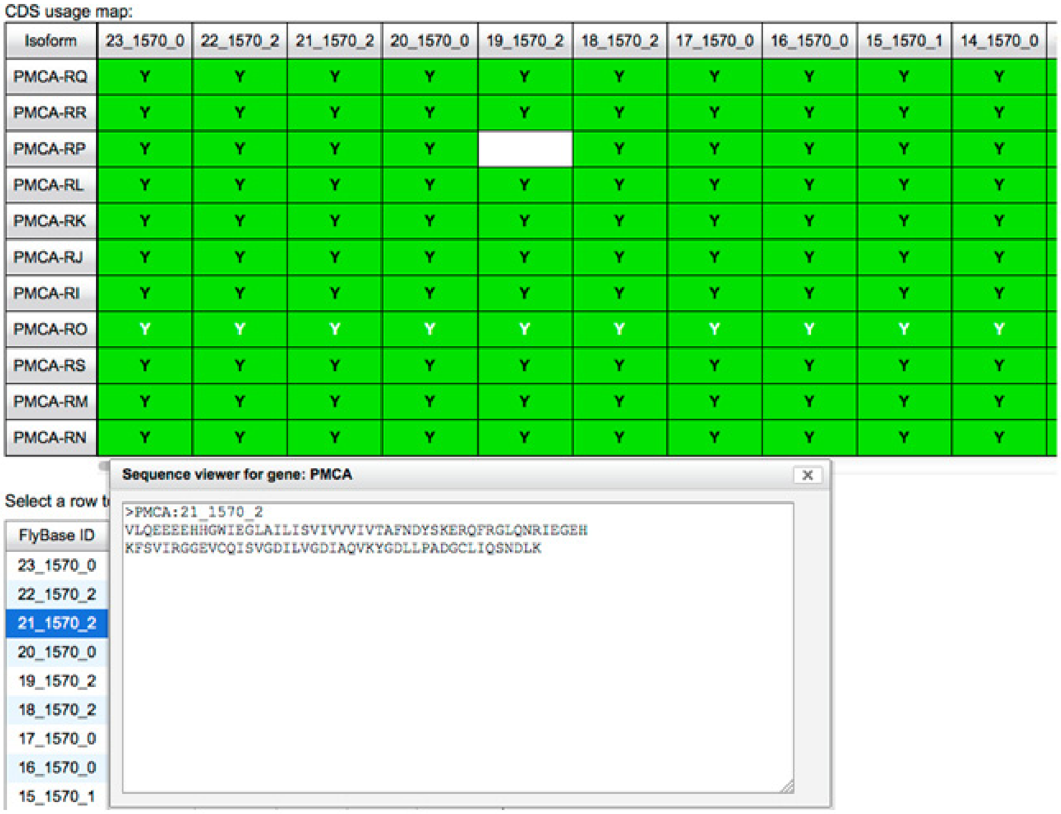
How many different protein isoforms exist for this gene?

Which mRNA molecules encode for identical protein sequences?

What might the different protein isoforms tell you about the (minimum) number of stop codons that are used in the expression of this gene?

Given the above answer, why do you think there is not a protein isoform that includes all of the CDS?

If you click on any of the CDS blocks of a particular protein isoform, information specific to the CDS's that are in that isoform appear in the table below the CDS usage chart. If you then click on a particular row of that table, the amino acid sequence of the specific CDS appears in a new pop-up window (Fig. 11). *Pay attention to the number (under Length) of amino acid residues in the Drosophila melanogaster isoform encoded by each CDS.*



CDS usage map:

Isoform	23_1570_0	22_1570_2	21_1570_2	20_1570_0	19_1570_2	18_1570_2	17_1570_0	16_1570_0	15_1570_1	14_1570_0
PMCA-RQ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RP	Y	Y	Y	Y		Y	Y	Y	Y	Y
PMCA-RL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RK	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RJ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RO	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RM	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PMCA-RN	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Select a row t

FlyBase ID
23_1570_0
22_1570_2
21_1570_2
20_1570_0
19_1570_2
18_1570_2
17_1570_0
16_1570_0
15_1570_1

Sequence viewer for gene: PMCA

```
>PMCA:21_1570_2
VLQEEEEHHGWIEGLAILISVIVVVIVTAFNDYSKERQFRGLQNRIEGEH
KFSVIRGGEVCQISVGDILVGDIAQVKYGDLLPADGCLIQSNDLK
```

Figure 11. Amino acid sequence of CDS #21_1570_2 (the third exon) for the 'O' isoform of *PMCA*

To map each putative CDS position, we will do searches that use the BLAST algorithm to compare two sequences (bl2seq). In this case, we will compare the entire translated *D. grimshawi* fosmid to each *D. melanogaster* CDS, to search for any *D. grimshawi* DNA sequence that could code for this CDS. Go to the NCBI BLAST search site (**http://blast.ncbi.nlm.nih.gov/Blast.cgi**) and compare the protein sequence of the first coding exon of *PMCA* (CDS #23_1570_0) to the entire fosmid DNA sequence by doing a blastx search. Obtain the contig11 fosmid DNA sequence by clicking 'DNA' in the header of the UCSC Genome Browser page (with contig11 entered in the position/search box). Then, paste this sequence (40,491 bp long) into the 'Query Sequence' box. Then, click on the 'Align two or more sequences' box and paste the amino acid sequence of CDS #23_1570_0 (copied from the pop-up window of the 'Polypeptide Details' section of the Gene Record Finder window) into the 'Subject Sequence' box. Click on 'Algorithm parameters, and be sure that the "Low complexity regions" box is NOT checked and that neither masking technique is being used. Click 'Show results in a new window', then click the BLAST button.

The best (first) alignment from this search should look as below. Note that the other two alignments below this one in the BLAST results window are from other regions of the fosmid and are so weak that they can be ignored.

```
>lcl|11553 PMCA:23_1570_0
Length=52

Score = 99.0 bits (245),  Expect = 4e-28
Identities = 45/52 (87%), Positives = 49/52 (94%), Gaps = 0/52 (0%)
Frame = +2

Query
3035    MATIDGRPAQYGVSLKQLRDIMEHRGREGIAKINEYGGIHELCKKLYTSPNE 3190 (Dgri DNA)
        MATIDGRPAQYG+SLKQLR++MEHRGREG+ KI E GGIHELCKKLYTSPNE
Sbjct 1 MATIDGRPAQYGISLKQLRELMEHRGREGVMKIAENGGIHELCKKLYTSPNE   52 (Dmel CDS)
```

By using the whole fosmid in our search, we can read the base coordinates of the alignment directly. We note that the beginning of the coding region of this gene appears to be highly conserved between *D. melanogaster* and *D. grimshawi* and is located starting at base 3035 in the fosmid sequence. The alignment begins at base 3,035 with the codon for methionine (remember your Basic Biology rules!) and ends at 3190. We also note that these bases were translated in frame +2 to obtain the similar amino acids. Finally, we note that the entire 52 amino acids of the *D. melanogaster* CDS (Sbjct line) align to this region of the fosmid.

Question 5.

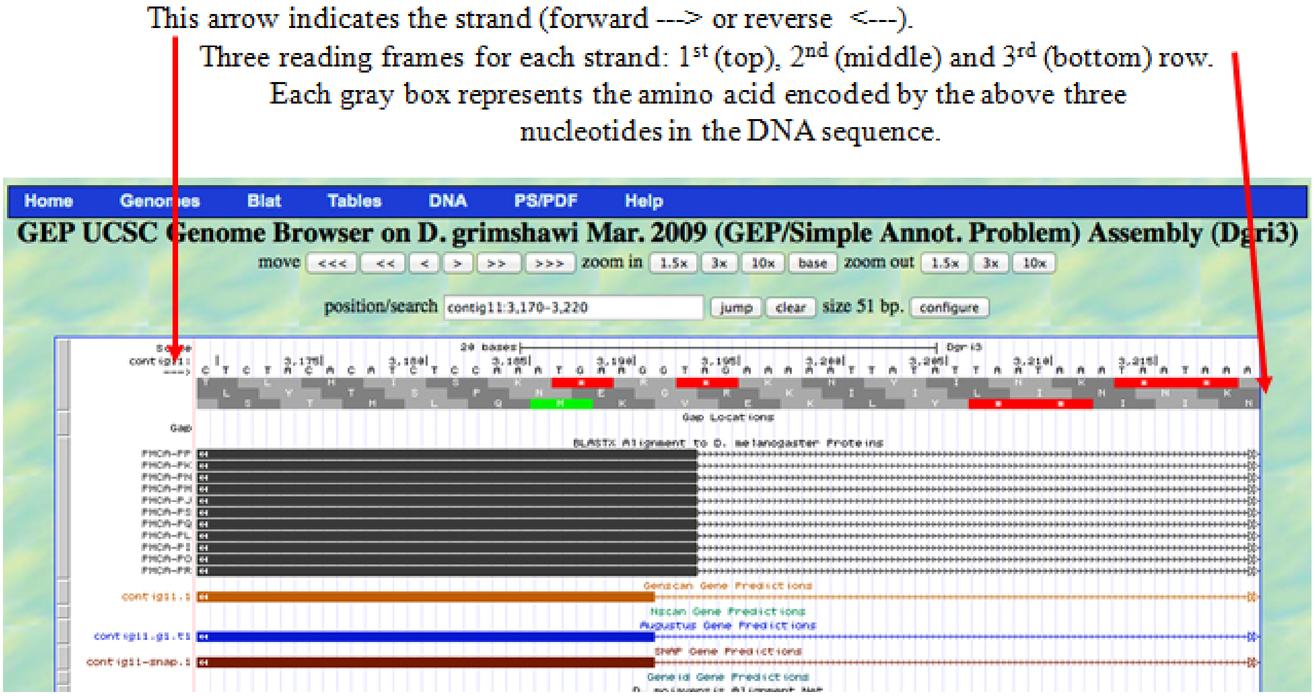
Repeat the same blastx searches with the next two CDS's (#22_1570_2 and #21_1570_2); copy and paste the best alignments into a Word document (when copying alignments, be sure to include the Score, etc. header information and shrink the margins and/or font to keep the sequences in alignment). Highlight the DNA base coordinates of the beginning and end of each alignment. Also hightlight the frame number that was translated to generate the amino acid sequence for the alignment.

For this worksheet, we will only attempt to find the exact boundaries for the first few CDS of the gene. If you were going to annotate the entire gene, you would continue down the line doing blastx searches for sequences similar to each CDS. Full annotation of this gene would require mapping ALL coding regions for ALL isoforms and creating viable gene models of each isoform.

Annotating CDS Boundaries

Since the BLASTX alignments show conserved amino acid residues between *D. melanogaster* and *D. grimshawi*, extra nucleotides between the first or last complete codon of an exon and intron splice sites may not be apparent in these alignments (if, for instance, a triplet codon has been split by an intron-exon boundary). Thus, to carefully annotate a protein-coding gene, we must find the exact mRNA splice positions that would create a processed mRNA that links these exons together to create a continuous coding block (i.e., an open reading frame).

Use the UCSC Genome Browser to navigate to the end of the first CDS. From the alignment above, we suspect the end to be very close to base 3,190. To jump directly to this region, we can enter the coordinates "contig11:3,170-3,220" in the 'position/search' box, then click 'jump.' Also make sure 'full' is selected under the 'Base Position' link in the 'Mapping and Sequencing Tracks' box and 'pack' is selected under 'D. mel Proteins' in the box below that. The new window is shown in Fig. 12. (You may need to zoom in a bit to clearly see the nucleotide bases.)



This arrow indicates the strand (forward ---> or reverse <---).

Three reading frames for each strand: 1st (top), 2nd (middle) and 3rd (bottom) row.

Each gray box represents the amino acid encoded by the above three nucleotides in the DNA sequence.

GEP UCSC Genome Browser on D. grimshawi Mar. 2009 (GEP/Simple Annot. Problem) Assembly (Dgri3)

position/search contig11:3,170-3,220 size 51 bp.

Figure 12. Close-up look at the region around the end of the first CDS alignment

From the blastx alignment just before Question 5, we know that the similar amino acids are found in frame +2. We should examine this region for potential intron donor splice sites, since we are at the beginning of the first intron. In this case, there is really only one potential "GT" donor site, which is at bases 3,192-3. This donor site therefore maps the end of the first exon at base 3,191. We can see by careful inspection that a cleavage here would leave a single base (a G) after the last complete codon in frame +2 (the codon for E, in frame 2, which is the second row of gray letters above). We use the term "phase" to describe these remaining bases; in this case, the 3,191 exon end is said to be in phase 1 because one base is present after the last complete codon and before the exon ends. To make a complete mRNA, we must find an acceptor site at the end of this intron, such that this one base at the end of the exon will join with two other bases in the next exon to make a complete three-base codon (Fig. 13). For now, we simply note that the only acceptable donor site in this region is in phase 1.

When you are deciding where to splice adjacent exons together, the acceptor site that you select for an exon must have a phase that is compatible with the phase of the donor site for the previous exon, e.g., phase 0 acceptor with phase 0 donor, phase 1 acceptor with phase 2 donor, or phase 2 acceptor with phase 1 donor (as in Fig.13). If the acceptor site has a phase that is incompatible with the donor, the extra bases will create a frame shift. The downstream CDS will then be translated in an incorrect reading frame, creating an abnormal polypeptide.

Use the Genome Browser to navigate to the region where you suspect the second exon begins based on the alignments you found in Question 5. Remember that intron acceptor sites have the invariant sequence 'AG' just before the first base in the following exon. *When you are trying to decide between or confirm intron splice sites*, change the 'Predicted Splice Sites' setting (at bottom of Genome Browser mirror window, under 'Experimental Tracks') to dense or full to see the position and ranking of potential donor and acceptor splice sites (see explanation of this track in the **Computational evidence** section of the Student Outline). Note the pink bar for a high acceptor splice site directly underneath an 'AG' (red arrow in Fig. 14). This pink bar marks the end of intron 1, and it directly abuts the start of the tan, green, blue, and brown bars that show the beginning of exon 2 in the gene predictor tracks.

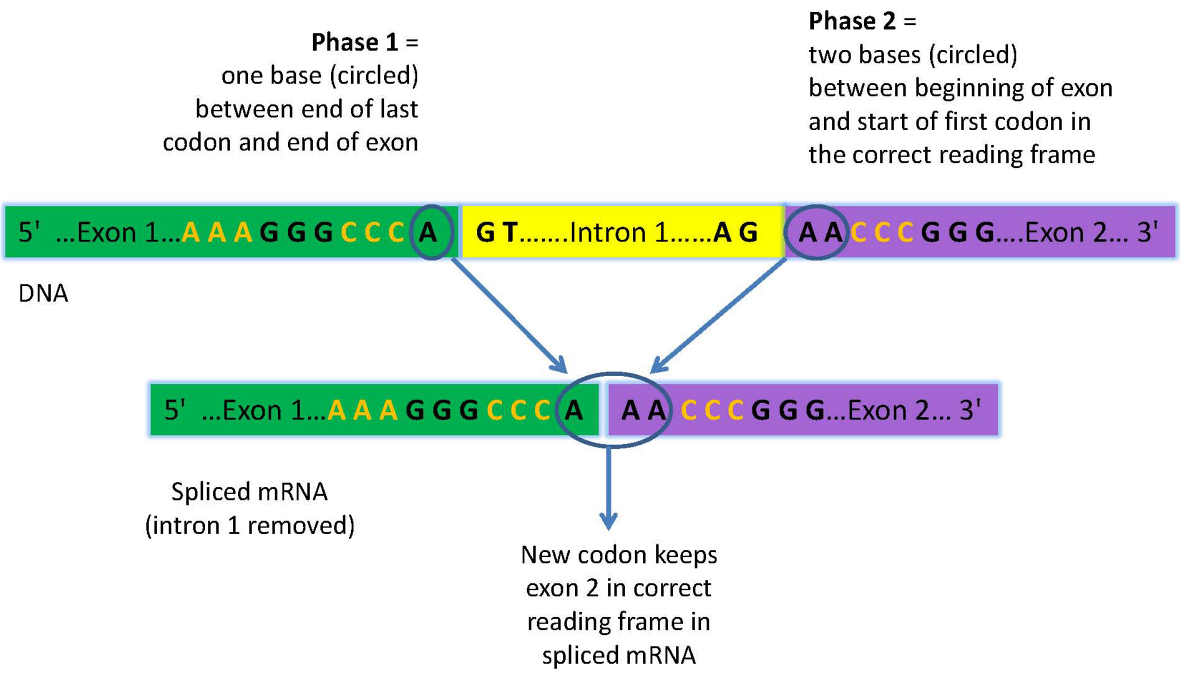
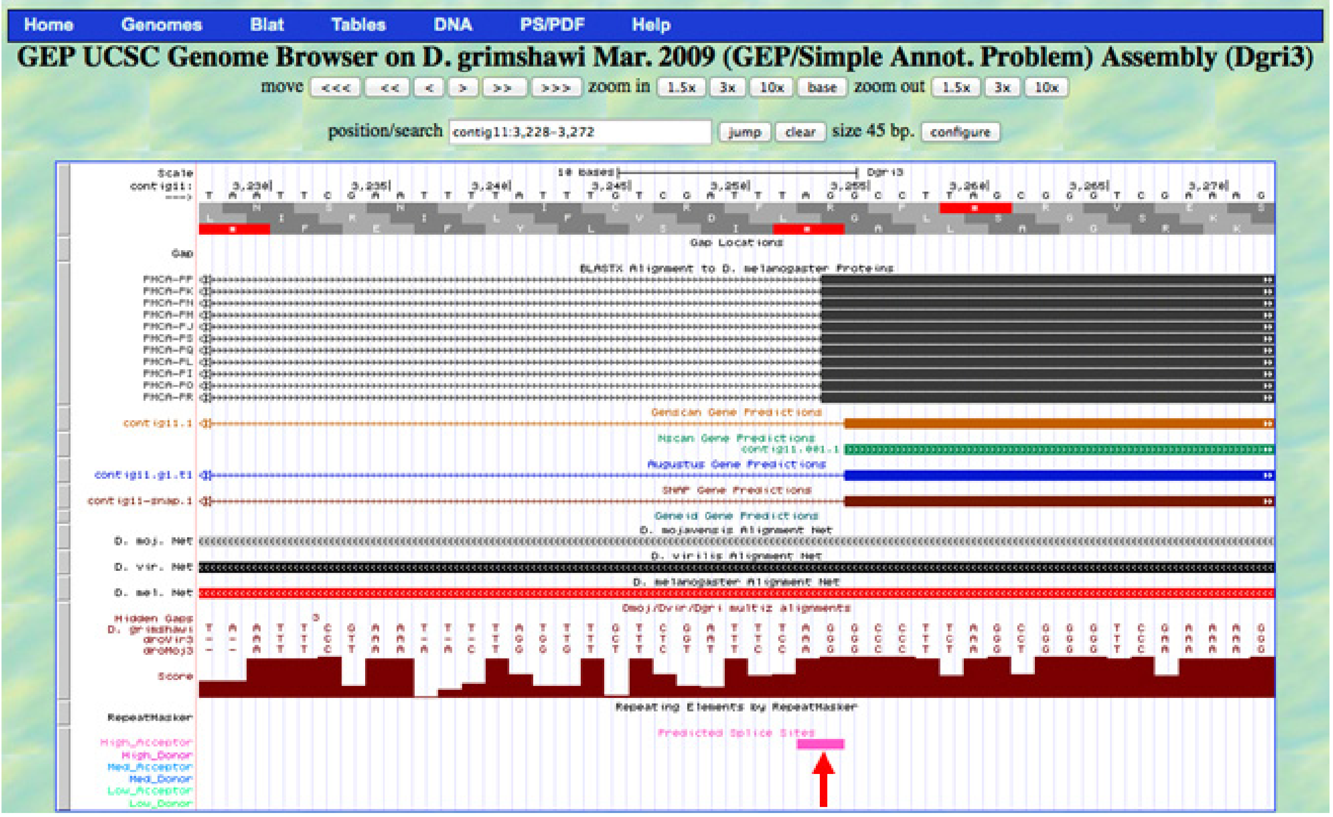

Phase 1 =
one base (circled)
between end of last
codon and end of exon
Phase 2 =
two bases (circled)
between beginning of exon
and start of first codon in
the correct reading frame
5' ...Exon 1...A A A G G G C C C A
G T.......Intron 1......A G
A A C C C G G G....Exon 2... 3'
DNA
5' ...Exon 1...A A A G G G C C C A
A A C C C G G G...Exon 2... 3'
Spliced mRNA
(intron 1 removed)
New codon keeps
exon 2 in correct
reading frame in
spliced mRNA

Figure 13. Example of compatible phases at the end and beginning of two adjoining exons



GEP UCSC Genome Browser on D. grimshawi Mar. 2009 (GEP/Simple Annot. Problem) Assembly (Dgri3)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search contig11:3,228-3,272 size 45 bp.

Figure 14. The results of the 'Predicted Splice Sites' program for the end of intron 1

Question 6

Look around the region where the alignment to CDS 22_1570_2 (the second exon) begins. How many possible acceptor sites can you find?

Considering the frame of the conserved amino acids you found in question 5, what is the phase of each putative acceptor site you find?

Using just phase information, which if any of these acceptor sites is/are usable to maintain the proper translation frame throughout the first two exons?

Itemize what other evidence you could consider if you have two or more possible donor/acceptor pairs.

Finally, record the base coordinates for the first exon and the beginning of the second exon as deduced from your complete analysis.

Full annotation of this gene would proceed to each subsequent exon, putative donors and acceptors would be analyzed for phase and all putative combinations would be compared to find the donor/acceptor pair with the most support.

Question 7

Use the results of the alignment of the second and third exons in question 5 to locate the 3' end of the second exon and the beginning (5' end) of the third exon.

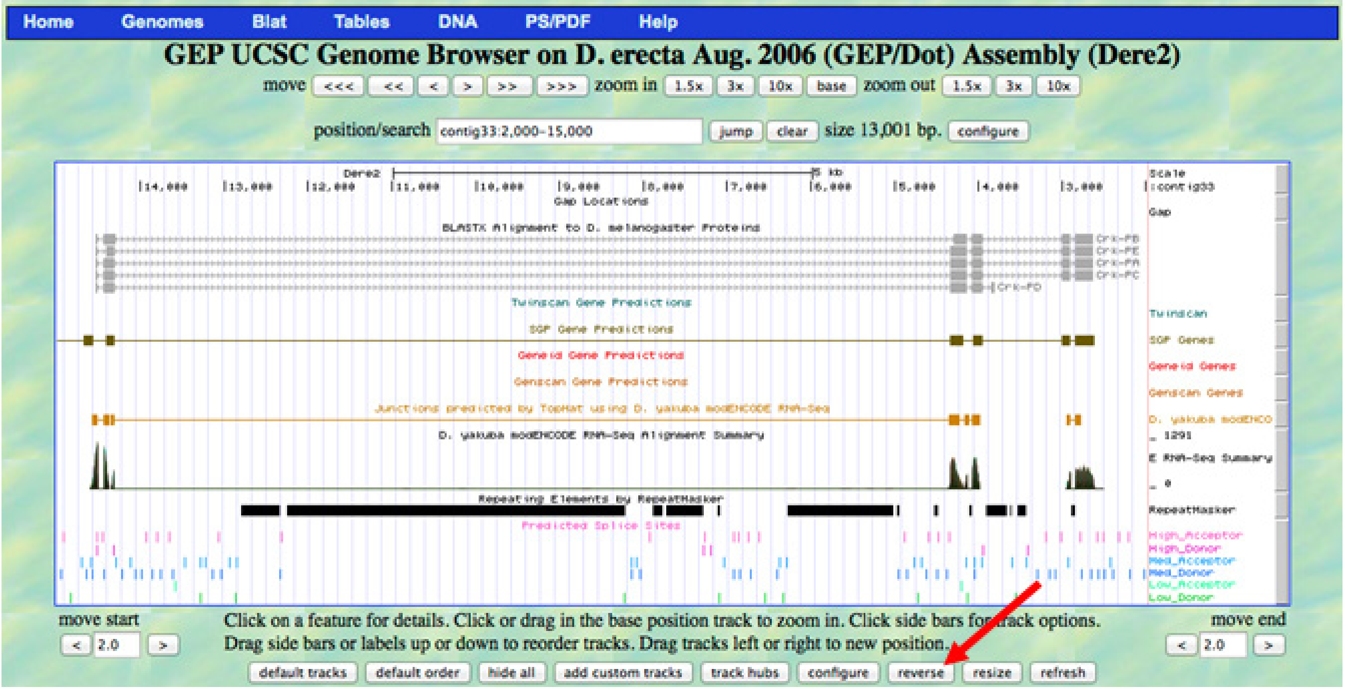
Addendum to a Sample Annotation Problem

Overview

Once you have carefully mapped the base pair coordinates of the beginning and end of each coding sequence (CDS) of your assigned gene(s) or isoform(s), the next step in the process of making a gene model is to check whether or not these coordinates code for a full-length polypeptide chain. To do so, you will use the GEP's Gene Model Checker. Detailed instructions for using this software can be found in the on-line User Guide at **http://gep.wustl.edu/repository/documentations/checker_guide.pdf**.

We think that working with the Gene Model Checker now will help you to better understand the specific goal of your annotation research projects. To do so, you will input the coding sequence coordinates from a gene model that was completed by a previous GEP undergraduate student. This student worked on the annotation of the *Drosophila erecta* ortholog of the *D. melanogaster Crk* gene. The *Crk* gene is found on the fourth (dot) chromosome and codes for five different isoforms in *D. melanogaster*. You will examine the model for the protein encoded by the B isoform in *D. erecta*. Figure 15 shows the fosmid region from *D. erecta* that contains the *Crk* ortholog. The BLASTX and the SGP Gene Predictor tracks both suggest that there are six coding regions in the *D. erecta Crk*-PB ortholog.

Table 1 lists the base pair coordinates (which were obtained by the GEP student using the UCSC Genome Browser) for the beginning and end of each CDS ortholog in the *D. erecta* fosmid. The CDS coordinates are listed in the 5' to 3' order that the corresponding coding regions would appear in the mature mRNA molecule. Note that the numbering of the coordinates goes down as one goes along the gene, indicating that it is the bottom (-) strand of the fosmid DNA that contains the coding information for the *Crk* gene. The "reverse" button (red arrow above) was clicked in the UCSC Genome Browser window to make the (-) strand go from left to right. Thus, in the SGP Gene track in Fig. 15, CDS 1 is the left-most box and CDS 6 is the right-most box, respectively, as conventionally drawn with 5' to 3' going left to right. Note that the base positions in the fosmid (top line of numbers) go in descending order as one reads from left to right.



GEP UCSC Genome Browser on D. erecta Aug. 2006 (GEP/Dot) Assembly (Dere2)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump
clear
configure

position/search contig33:2,000-15,000 size 13,001 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.
Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Figure 15. Site of probable Crk ortholog on contig33 of *D. erecta*

Table 1. Coordinates of *Crk* CDS in contig33 of *Drosophila erecta*.

	Base Pair Coordinates
CDS 1	14539-14508
CDS 2	14435-14318
CDS 3	4292-4137
CDS 4	4076-3948
CDS 5	2957-2904
CDS 6	2844-2623
Stop codon	2622-2620

Using the Gene Model Checker

To confirm the accuracy of a gene model using the Gene Model Checker, proceed as follows:

1. Select 'Gene Model Checker' from the Projects -> Annotation Resources drop-down menu at **http://gep.wustl.edu/**.
2. Download a copy of the fasta sequence file for this particular fosmid (*D. erecta* dot, contig33) on to the desktop following the directions of your instructor. Upload this sequence file into the first box on the left of the Gene Model Checker window.
3. Type in the name of the *D. melanogaster* ortholog (Crk-PB, case-sensitive) in the second box– watch for this gene to appear in the drop-down menu box as you type.
4. Enter the base number of the beginning and end of each coding sequence, in the order they appear in the table above, in the following format: # - #, # - #, # - #, etc.
5. *Do not include the stop codon in the last CDS since the stop codon does not code for an amino acid.* Instead, enter the stop codon coordinates in # - # format in the box lower down in the Gene Model Checker window.
6. The information for this gene is on the (-) strand, so click the circle in front of 'Minus' following 'Orientation of Gene Relative to Query Sequence.'

7. We think all the gene's coding sequences are accounted for, so click the circle in front of 'Complete' in the next row.
8. Use the drop-down menu to select the Project Group (*D. erecta* Dot) and type in the Project Name (contig33) for this fosmid.
9. *Red boxes will appear around any entries that the Gene Model Checker deems incorrectly entered. Fix these before going on.*
10. Click on the 'Verify Gene Model' box at the bottom of the window.
11. A summary of how your model did now appears on right side of the Gene Model Checker window. If all the information was typed correctly, you should have passed with flying colors.
12. *If you passed the Gene Model Checker on the first try, intentionally change some of the coordinates to see what a failure would look like!*
13. If there are failed parts of the Gene Model Checker, click on the small + box to the left of the failed part to get more information on the problematic sequence. A simple misreading or mis-typing of intron-exon boundary coordinates is responsible for many model failures.
14. If your problem was "Fail with premature stop codons', click on the Peptide Sequence tab in the upper-right menu bar next to "Checklist.' The symbol * in the peptide sequence that appears next will show you where these premature stop codons are. You can then use this information to find the coordinates of the problematic CDS using your saved blastx alignments (as you will learn more about during the next lab).
15. Re-check your work and the typing of the coordinates until you pass *ALL* parts of the Gene Model Checker.
16. Click on the 'Peptide Sequence' tab to see the amino acid sequence created by your gene model. You should also click on the 'Dot Plot' tab in the Gene Model Checker output window, which will show you how well this sequence aligns with the amino acid sequence of the putative *D. melanogaster* ortholog. If the two models are completely homologous, you will see a solid diagonal line; you should be able to explain all discrepancies from this outcome.

For your own projects, you will then use BLAST to compare the predicted amino acid sequence to the *D. melanogaster* protein sequence for a final check of your gene model.

After checking your answers against the answer key (which will go live on the course Web site on Friday at 5 PM), write down any questions you have on the activity and bring them to lab next week.

We will begin next week's lab by going over answers to your questions.

Materials

AV Support/Room Requirements

These exercises require a classroom or computer lab with a LCD projector driven by an instructor's computer and desktop computers (PC or Mac) with high-speed Internet connection. There should be at least one computer for every two students and ideally one computer per student. Every computer should also have the Microsoft Office software package (with MS Word and MS PowerPoint) and a simple text editor (such as WordPad).

On-Line Documents

Materials developed for use by GEP members and other interested educators can be viewed and/or downloaded from the GEP web site (**http://gep.wustl.edu**). Electronic files for all documents referenced in the Implementation section are available on Dr. Julie Emerson's GEP Wiki page at **http://gep.wustl.edu/wiki/index.php/Julia_Emerson_Amherst_College**. The materials are also available through the main GEP web site under Curriculum -> Course Materials -> GEP Partners (**http://gep.wustl.edu/curriculum/course_materials_GEP_partners/**). Detailed help guides to the various GEP Web sites are accessible via the main GEP Web site: **http://gep.wustl.edu/**; go to Help -> Documentations -> Web Framework. Finally, an instructor's answer key to *A Sample Annotation Problem* and a FASTA file with the contig33 DNA sequence from *D. erecta* (for the *Addendum to A Sample Annotation Problem*) are also posted on Dr. Emerson's GEP Wiki page.

Notes for the Instructor

Membership in the GEP enables biology faculty/instructors and their students to participate in the research activities of the GEP, by improving DNA sequences and/or annotating newly-sequenced DNA. The GEP is recruiting new members, and interested individuals should contact Professor Sarah Elgin (**selgin@biology.wustl.edu**) at Washington University in St. Louis for more information.

DNA sequence improvement and/or gene annotation is done at many institutions as part of an upper-level course or as an independent research project (see GEP Wiki at **http://gep.wustl.edu/wiki/index.php/Table_of_Faculty**). In these situations, a major goal is for students to be active participants in the GEP research enterprise. Critical to achieving this goal is a small class size and low student to instructor ratio (ideally, no greater than 6:1). Experienced students make excellent peer instructors, coaching novices in the use of the computer-based tools used for annotation. In these courses, students receive significantly more training and have more time to figure out problems they encounter when working on their own projects.

At Amherst College, we are doing gene annotation with large numbers (>100) of introductory students and have a student to instructor ratio of up to 8:1 per lab section. Thus, we developed the *How-To Guide for Gene Annotation* (see Emerson GEP Wiki page), which provides students with start-to-finish, step-by-step instructions for annotating a single isoform of a newly-sequenced *Drosophila* fosmid.

One caveat of providing this document is that some students may be successful in following directions, but do not really understand what they are doing and why. Therefore, it is important that the instructional staff frequently ask students to explain the rationale for the specific steps as they work through the procedure.

Another concern of annotating many different, newly-sequenced genes in large enrollment courses is that each gene may present different challenges, and it is difficult to train all instructors and TAs for all possible scenarios. As a result, we have decreased the number of new features that students annotate each year and pre-screen all of them prior to the TA training sessions. In this way, we know what most of the challenges will be ahead of time and are better able to help students when they get confused.

Finally, the number of weeks we devote to gene annotation ranges from four weeks (as described in this paper) to just one week, as it depends in part upon the preferences of the faculty members who rotate through the Biology 191 course. In the one-week lab, we de-emphasize the research objective of this work to focus on introducing students to bioinformatics tools while reinforcing concepts of eukaryotic gene structure, mRNA splicing and translation reading frames. Students work through a new worksheet that we created by combining and modifying parts of the GEP's An Introduction to NCBI BLAST, the Sample Annotation Problem and the How-To-Guide. Students in each lab section work as a group to annotate the same gene and create a gene model (see the 2012 files linked to the Emerson GEP Wiki page). Although students may miss out on the excitement of working on their "own" uncharted genes in the one-week lab, we believe that they are nonetheless gaining a richer understanding of eukaryotic gene structure and how to use on-line databases, which will serve them well in their upper-level biology course work, future research experiences and post-graduate studies in the biological or biomedical sciences.

Acknowledgements

The GEP is supported by Howard Hughes Medical Institute grant #52005780 to Sarah C.R. Elgin at Washington University in St. Louis, Missouri.

Literature Cited

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B. and G. Rätsch. 2008. Tutorial: Support vector machines and kernels for computational biology. **http://svmcompbio.tuebingen.mpg.de/splicing.html**

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B. and J. Merrick. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science* 269: 496–512

Genomics Education Partnership. 2012. **http://gep.wustl.edu.**

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., H. Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and S. G. Oliver. 1996. Life with 6000 genes. *Science* 274: 546, 563–567

Guo, M., Lo, P. C. and S. M. Mount. 1993. Species-specific signals for the splicing of a short *Drosophila* intron *in vitro. Molecular and Cellular Biology* 13: 1104-1118.

McManus, C. J., Duff, M. O., Eipper-Mains, J. and B. R. Graveley. 2010. Global analysis of trans-splicing in *Drosophila. Proceedings of the National Academy of Sciences USA* 107: 12,975-12,979.

Painter, T. S. 1934. Salivary chromosomes and the attack on the gene. *Journal of Heredity* 25: 465-476.

Talerico, M. and S. M. Berget. 1994. Intron definition in splicing of small *Drosophila* introns. *Molecular and Cellular Biology* 14: 3434-3445.

Walter, C. and M. Wilkerson. 2006. Annotation for Amateurs website, **http://www.plantgdb.org/tutorial/annotatemodule/index.html**

About the Authors

Dr. Julie Emerson is the Lab Coordinator for the Department of Biology at Amherst College. She teaches in the Biology 181 (Adaptation and the Organism) and Biology 191 (Molecules, Genes and Cells) courses, in which students obtain a comprehensive introduction to many key biological concepts and underlying scientific approaches, methodologies and pedagogical approaches to learning. Dr. Emerson joined the GEP in June, 2009.

Dr. S. Catherine Silver Key is an Assistant Professor of Biology at North Carolina Central University. She teaches undergraduate Genetics (BIOL3100), Cell and Molecular Biology (BIOL2200), Introduction to Research (BIOL4400), and Inquiries in Developmental Biology (BIOL4100). She also mentors graduate students at the NCCU Master's Program in her Drosophila-based research lab. Dr. Silver Key joined the GEP in June, 2009.

Dr. Consuelo Alvarez is an Associate Professor in Biological and Environmental Sciences at Longwood University and the Faculty Athletic Representative. She teaches Biochemistry 412, Modern Genetics 426, Evolution 399, Genetics 324, Biology 121 for majors and Biology 101 as well as Chemistry 101 for non-majors. All her courses have a component that introduces the students to bio-techniques such as DNA microarray, Genomics analysis and Synthetic Biology. Dr. Alvarez joined the GEP in June 2007.

Dr. Stephanie Mel is a Lecturer in the Division of Biological Sciences at the University of California at San Diego (UCSD). UCSD is a large, public institution with approximately 6,000 biology majors. Dr. Mel has used GEP materials in a small, research-focused course Undergraduate Research Explorations in Genomics, which was limited to twenty students. Students in the course appreciated the opportunity to work closely with faculty on a research-based problem. Dr. Mel joined the GEP in 2009.

Dr. Gerard McNeil is an Associate Professor and Chair of the Department of Biology at York College, The City University of New York (CUNY). He is also an active member of the graduate faculty of the Biology and Biochemistry Doctoral Programs at The Graduate Center/CUNY. He teaches Genetics, Cell Biology, Bioinformatics, and General Biology II and has used GEP resources in several of these courses in different ways. Dr. McNeil joined the GEP in June, 2007.

Dr. Kenneth Saville is a Professor of Biology at Albion College in Albion, Michigan. He is a geneticist whose primary research organism is the fruit fly *Drosophila melanogaster*. He teaches multiple courses in genetics and molecular biology and has integrated GEP materials into a stand-alone course in genomics. Dr. Saville joined the GEP in 2007.

Mr. Wilson Leung graduated from Washington University in 2005 with a BA in Biology. He is the Chief Technical/Teaching Assistant for the GEP, and he assists in the biannual GEP training workshops for new members, writes documentation, manages the GEP Web site, and assists current GEP members with questions as they implement GEP materials into courses at their home institutions. Mr. Leung also performs computational research analyses of various Drosophila genomes.

Dr. Christopher Shaffer is a Lecturer in the Department of Biology at Washington University in St. Louis, Missouri. As Technical Director of the GEP, he instructs in the biannual GEP training workshops for new members, writes documentation (including the GEP's A Simple Annotation Problem) for adoption by GEP members, sets up GEP-related software, and assists current GEP members with questions.

Dr. Sarah Elgin is the Viktor Hamburger Professor of Arts & Sciences and Professor of Biology at Washington University in St. Louis. Dr. Elgin's research interests focus on the role that chromatin structure plays in gene regulation, which her lab investigates using the fruit fly *Drosophila melanogaster* and related species. Dr. Elgin is a founding member of the GEP, which was created in 2006, and the GEP Program Director. Dr. Elgin uses GEP materials in her upper-level course, Research Explorations in Genomics, in which each student completes a research project in both DNA finishing and gene annotation.

Appendix

Incorporation of GEP Materials and Curricula into Courses at Other Institutions

North Carolina Central University (Cathy Silver Key):

Currently, Cathy Silver Key at NCCU uses the GEP resources in two of her courses: Introduction to Research (BIOL4100) and Genetics (BIOL3100). For the Intro to Research course, students meet at least 10 hours a week in groups of one to four students. For the first 3 weeks, students are assigned the tutorials and exercises available through the Curriculum tab on the GEP website. They then progress into annotating a fosmid or contig for the remainder of the 15-week semester. Students are able to successfully annotate one or two contigs by the end of the semester and present a poster at our NCCU Research Day.

In Genetics (BIOL3100), Dr. Silver Key uses the GEP resources in a 6-week module. The current approach is to start with a modified 'Basics of BLAST' followed by the first two BLAST exercises created by Dr. Hui-Min Chung (University of West Florida) and also an exercise based on the *Flybase for Undergraduates* You Tube video also created by Dr. Chung. The students then receive Genomics lectures from Chapter 20 in our class textbook: *Genetics: a Conceptual Approach*, 4th Ed. by Benjamin Pierce, followed by Washington University exercises titled *A Simple Introduction to BLAST* by Wilson Leung and *A Simple Annotation Problem* by Chris Shaffer. I also find the 'Splice Site Flow Sheet' and the 'Genomics Glossary' very helpful. Groups of four students then begin annotation of *Drosophila* contig projects downloaded from the GEP website and are required to submit a completed submission sheet at the end of the semester. Genetics has one 2-hour lab and two 1 hour and 20 min lectures per week. The first 3 weeks, Genomics is the topic in both lecture and lab. For the remaining 3 weeks, Genomics studies occur in lab only. Usually, subsets of students become the leaders in these Genetics groups and have some success with annotation.

Longwood University (Consuelo Alvarez):

Longwood University is a small liberal arts university in central Virginia. We do not have graduate programs in the sciences; thus our teaching is mainly oriented to undergraduate students. GEP materials are used in a stand-alone class taught every-other year as an elective course for biology majors. I recruit students by visiting my peers' classes before registration for the spring semester starts, as well as by posting flyers around the science building and inviting students to read the GEP poster outside my office. We do both finishing and annotation in the course, and due to the small number of students taking the class, each one works on his/her own project. The class meets weekly in 3-hour blocks, during which we have a mix of lecture, discussion of papers students read in advance, troubleshooting discussions, and then each student continues working individually.

Assessment is an important component, and we do both the GEP assessment as well as an internal assessment that is used to improve the course. The course is designed as a written and speaking course because students orally present their findings in both finishing and annotation projects, in addition to submitting a written paper of each project. I love the overall idea of these projects that students need to "prove by evidence your findings." This is a good skill that students in this course develop as the semester progresses. In one instance (a fall semester), I have done internships with only annotation work.

University of California, San Diego (Stephanie Mel):

The University of California San Diego is a large public university with more than 23,000 undergraduate students and over 4,000 Biology majors. In the spring of 2010, I created a stand-alone course using GEP materials. The 20 students accepted to the class had to have completed a molecular biology lecture course and received at least a B in that class. We met twice a week for 10 weeks in a Mac computer lab and each class lasted 2 hours.

The class format included short lectures, worksheets, and computer exercises on both finishing and annotation. After gaining a solid introduction to the tools, students then claimed annotation projects. The short lecture topics included an Introduction to BLAST, DNA sequencing, background on Consed and finishing, and annotation. We used some GEP handouts and exercises, but also created worksheets of our own. A total of 21 annotation projects were submitted at the end of the class. Students were required to write a report and give an oral presentation on their annotation results.

The course was very well received. Student comments included "Should be the model that all science classes are built upon" and "This is a class that other classes should be modeled after. It takes what students have learned in other classes and really applies it to a research question. It differs from other lab classes as the data and research produced from this class is actually new data and will be used in future research."

York College, The City University of New York (Gerard McNeil)

York College is a senior college of the City University of New York located in Queens, NY. One of the major challenges of the Department of Biology is to provide a research opportunity to all majors (over 400). With only nine active research faculty, this is a daunting challenge. To address this, I have tried to incorporate a real research experience in several of my courses using The Genomics Education Partnership as a tool. Advantages of the curriculum of this partnership include only the need for computers, low cost, training for faculty and TAs, ability to generate novel data for a large genomics question, access to many students without the need of a wet lab, and its versatility in implementation.

I have incorporated this curriculum into a stand-alone bioinformatics course (Btec 352), as part of a cell biology lab (Bio320), and as both an individual and group independent study course (Bio490-493). In Btec352, we have the students perform both DNA sequencing improvement (7 weeks) and annotation (7 weeks), meeting 4 hours per week. In Bio320, we incorporated annotation in the lab interspersed with several wet lab experiences. Independent study has been done in my research lab along with wet lab experiments (all my undergraduates do this), either individually or as a group of four separate from my research lab. Meeting times varied from 3-6 hours per week for 14 weeks. Other members, including other authors of this paper, have successfully used shorter versions of the GEP curriculum.

Albion College (Kenneth Saville)

Albion College is a small, private, liberal arts college in south-central Michigan. I have taught GEP materials in two different settings. First, I have taught a stand-alone genomics course with an average of 10-12 students. In this course, I begin with general bioinformatics background information, primary literature on heterochromatin in general and the fourth chromosome of *Drosophila* in particular. We then spend approximately half of the semester doing sequence improvement or ‘finishing’, using the Consed program, and the remainder of the semester doing sequence annotation. In this class format, each student is assigned one fosmid for finishing and another for annotation. This format is ideal and students really get a sense of doing research, solving problems, and contributing to the overall project.

The second format in which I have taught this class is as part of a genetics lab connected to a more traditional genetics course. These classes typically have about 32 students with 16 students per lab section. In this scenario, about 6 weeks of the semester is dedicated to annotation. Typically, students work in pairs to annotate one or two genes, with the entire class completing a few fosmids. While students report some sense of doing ‘real’ research, the primary outcome is a solid understanding of gene structure and the sequence information contained within genes. However, the prospect of being part of a publication is exciting to these students as well.

Finally, I have worked with several students one-on-one in a directed study setting. These students have done both finishing and annotation and seem to enjoy process.

Washington University in St. Louis (Christopher Shaffer, Wilson Leung and Sarah Elgin):

Washington University in St. Louis is a mid-sized research university with over 10,000 students in undergraduate, professional and Ph.D. programs. Many Biology undergraduates go on to either medical or graduate school. Most GEP materials were originally developed for Biology 4342 *Research Explorations in Genomics*, a full-semester 8-hour/week lab course devoted completely to GEP projects. Students who take this class are usually juniors or seniors majoring in biology plus a few students from the Department of Computer Science in the College of Engineering. The class meets twice a week for 3.5 hours and once a week for 1 hour. The two large time blocks are used for training and student work while the 1-hour sessions are typically used for lectures by Washington University faculty that cover the use of genomic analysis in other ongoing research projects.

Biology 4342 is divided into two sections: students undertake a sequence improvement project during the first 6 weeks of the course and work on annotation projects during the following 8 weeks. The sequence improvement section begins by using GEP training material for 2 weeks. We start by asking students to read *A Guide to Consed* outside of class and follow the scripted walk-though “Using *Consed* Graphically”, “A Simple *Drosophila* Fosmid” and “A Complex *Drosophila* Fosmid” during class. Following the initial training, students are assigned the *Drosophila Finishing Problem Set* as a homework assignment. Students then spend 4 weeks working on their own projects. Students have three opportunities to design sequencing reactions and receive back results to provide any additional data needed for their project. Most students will complete one or more projects during this 6-week period. We ask our students to present a 10-minute talk and produce a paper (~10-20 pages) documenting the progress they have made on their projects.

Students undertake an annotation project during the last 8 weeks of the semester. Training begins with an ungraded scripted walk-though using *A Simple Introduction to NCBI BLAST*. This is followed by two graded exercises: *Exercise #1: Detecting and Interpreting Genetic Homology* and *Exercise #2: Browser-Based Annotation and RNA-Seq Data*. We follow this initial training with a 2-week unit on the analysis of primate genomes, starting with the *Worksheet: Chimp BAC Analysis: Genes and Pseudogene* followed by groups of 2-4 students working together on the annotation of a portion of the chimpanzee genome that contains at least one gene and at least one pseudogene. The students complete their training with the scripted walk-though *A Simple Annotation Problem*, which describes specific annotation strategies and issues in GEP annotation projects. Students work on their own projects for 4 weeks, generating optimal gene models based on available evidence. In addition, they examine the characteristics of their fosmid (repeat density and distribution, synteny with *D. melanogaster*, etc.), carry out a Clustal analysis, and explore any other issues. Students present the results of their analysis as both an oral presentation and a written report. Examples of written reports are available on the GEP webpage.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Emerson, J.A., S.C.S. Key, C.J. Alvarez, S. Mel, G. McNeil, K.J. Saville, W. Leung, C.D. Shaffer and S.C.R. Elgin. 2013. Introduction to the Genomics Education Partnership and Collaborative Genomics Research in *Drosophila*. Pages 135-165. in *Tested Studies for Laboratory Teaching,* Volume 34 (K. McMahon, Editor). Proceedings of the 34th Conference of the Association for Biology Laboratory Education (ABLE), 499 pages. http://www.ableweb.org/volumes/vol-34/?art=6

ABLE strongly encourages individuals to use the exercises in this proceedings volume in their teaching program. If this exercise is used solely at one's own institution with no intent for profit, it is excluded from the preceding copyright restriction, unless otherwise noted on the copyright notice of the individual chapter in this volume. Proper credit to this publication must be included in your laboratory outline for each use; a sample citation is given above.