Drosophila Annotation Goals:
Final Presentation and Written Reports

Chris Shaffer, Wilson Leung, and Sarah C.R. Elgin

This document describes the primary annotation goals to be included in the final oral presentation and written report for students enrolled in the Bio4342 course at Washington University in St. Louis. The final will be made up of two parts, a 10-minute oral presentation and a written report.

Oral Presentation

Give a final presentation in which you briefly discuss the most interesting features of your *Drosophila* project(s), potentially including some (probably not all) of the putative genes and non-coding RNAs, the assignment of potential Transcription Start Sites (TSS), distribution of repetitious elements, repeat element analysis, synteny with *Drosophila melanogaster*, expansion analysis, and any other curious observations found in your region. Your paper will discuss your best determination of the exact nature of each feature (possible gene) and the reasoning that led you to this conclusion. However, given the 10-minute limit on your talk you will probably want to focus on those genes and/or TSS assignments that were particularly interesting or difficult to annotate. Remember there is no right answer; this is the first time this sequence has ever been looked at by anyone. The goal is to come to the most reasonable interpretation of the data based on observations, and be able to articulate the reasoning behind your hypothesis. Show us your evidence! Critique of the protocols we used, are not required but certainly of interest. Remember that you can use this time to seek feedback on your analysis which may be useful in the final written report.

You may use PowerPoint (recommended) or other presentation tools but you will need a final file that you can submit to a Canvas assignment. Figures of some kind are required for your presentation, as they always play an important role in showing your evidence and getting your point across to your audience.

Written Report

The major part of the final will be your complete written report. Several major topics (grouped as you see fit) are recommended: Abstract, Introduction, Genes, TSS Estimates, Repeats, Synteny, Expansion Analysis, and Discussion/Conclusions. TSS estimates can be reported at the end of the report on the intron/exon structure of the gene, rather in a separate section. Each section should end with a brief statement and/or table and/or figure summarizing your findings and conclusions. The annotation of the first gene you worked on should be reported in detail; subsequent cases can be reported succinctly, with reference back to the methods used in the first case, but should include the results of your search for conservation of coding sequences (table or possibly other), a table with your final exon coordinates and characteristics (reading frame, splice site phases, etc.), and Gene Model Checker output (dot plot AND aa alignment). Note that the final report should include revised versions of the first gene report and the TSS/Repeats report you have submitted earlier in the semester. We also ask that you submit an **Appendix** with appropriate data files (see below). Your report should be submitted to the Canvas assignment as a file upload of a Word document. You may also include a PDF version of the report as part of the submission (see the "**Addendum on Writing**" section for details).

Remember that your job is to communicate your findings clearly, and to convince the reader of the validity of your conclusions! The report should integrate text with figures, making a single, coherent document. The reader must be able to **see** the data that you present, and to understand the source. Zoom in as needed, and add arrows, boxes, etc. to the figure to highlight the features of interest for the reader. (Remember that you can use the "configure" functionality in the UCSC Genome Browser to increase the font size and reduce the image width in order to make the Browser images easier to read.)

When you discuss BLAST results, remember to specify the **BLAST program, the query sequence, and the subject sequence** used in the search; this information should also be in the Figure legend where you show results. Don't leave out supportive data — cite all sources, particularly in describing the general procedure (your first gene). Be as specific and precise as possible when you describe the evidence used to support your final gene model. This document will be the basis for future research on this topic by following Bio 4342 students, will be the source of information for our joint publication, and may be used (with your permission) to demonstrate what undergraduates can accomplish in a challenging research endeavor.

Abstract

The abstract should provide a brief statement of the goals and a report of your findings. This should be similar to and about the same length as a typical abstract found at the beginning of a scientific paper (~300 words). Your abstract should be a summary of your results, **NOT the process by which you achieved those results**. For the *D. kikkawai* annotation project, be sure to include the size of the project, number of Genscan predictions, protein-coding genes documented, and other verified features (*e.g.*, novel repeats, non-coding RNAs) or unusual findings (*e.g.*, changes in exon structure, changes in the number of isoforms), etc. Be specific in reporting what you have found and your resulting conclusions.

Introduction

One to three paragraphs: Why are we studying this problem? Why is comparative genomics a powerful approach for analyzing the features of genes and chromosomes? What kinds of results are we generating right now, and what do we hope to derive in the future aggregate analysis? Try to keep this section concise and informative but give some context — what are the goals of the overall study, and what are your goals in particular. Include a figure covering the entire sequence of your project(s) that indicates the size and position of each putative feature you will investigate. A screenshot of the GEP UCSC Genome Browser for your project region may provide a good starting point for this figure. Box and number the features that you will investigate/report on for ease of communication.

Genes

This section will be significantly larger (reflecting the number of features in your project), and it should be subdivided with a sub-section for each putative feature. You should present here a more detailed analysis of any genes, pseudogenes or partial genes you find in your project. Use the Genscan predictions (or similar) as an organizer for your presentation. Remember that while in *D. melanogaster* pseudogenes are rare, this may not be the case in these regions of high expansion so keep an eye out for evidence suggesting the presence of pseudogenes in your project. Because the projects are partitioned into sub-regions, you may find only part of a gene at one end or the other of your project. Also be aware that because we are annotating the draft whole genome assemblies, sequencing errors may also occur. Note that Genscan and other gene finders may miss some candidate genes, so do not limit your investigation to these predictions; look at the other evidence tracks (homology search with *D. melanogaster*, RNA-seq, etc.) and examine the orthologous region of the *D. melanogaster* genome as annotated by the curators at FlyBase. In particular, Genscan will not predict non-coding RNAs (ncRNAs); be sure to investigate those posted for *D. melanogaster*, looking for conserved sequences in your species (*e.g.*, via a *blastn* search of the non-coding RNA against your contig). When investigating a feature, you should use all of the information available to you: gene predictions, results from BLAST searches, RNA-seq data (*e.g.*, splice junction predictions and assembled transcripts), and conservation tracks for the other *Drosophila* species (see the walk-throughs and practice examples).

For each gene in your *D. kikkawai* project, start by confirming the *D. melanogaster* orthologue; include the FlyBase ID and the name of any *D. melanogaster* gene that you consider likely to be orthologous in your report. After making approximate assignments for each exon using the “Align two or more sequences” functionality in BLAST, determine the exact location of each coding exon using all available evidence. Construct a model for each putative isoform. Use the Gene Model Checker to confirm your basic decisions; be sure to check both the dot plot and the peptide sequence alignment comparison to *D. melanogaster.* Every difference between your model and the *D. melanogaster* orthologue represents an assumption of evolutionary change that underlies your gene model. Based on the GEP annotation protocol, the best gene model should minimize these assumptions; as such, you should be able to account for every difference and convince yourself that there is no other acceptable gene model with fewer differences. [If possible, these differences should be supported by biological evidence (*e.g.*, RNA-Seq data) or sequence conservation with other *Drosophila* species.] The dot plot and peptide sequence alignment, with a discussion of important differences (for each isoform), should be part of your report.

Note that information from high-throughput sequencing of RNA, the RNA-Seq track and its derivatives, will be very useful. RNA-Seq sequences were generated by producing many short reads (100-125 bases) from expressed RNAs (in many cases using total RNA from embryos or from adult females or males) using the Illumina sequencing technology. These reads have been mapped back to the genome sequences and the number of reads that overlap each base plotted in the "RNA-Seq Coverage" tracks. In addition to using the RNA-Seq results to help you with your annotations, you should also compare the RNA-Seq data with your final gene model. Relevant questions include: How well does the RNA-Seq coverage track correspond to the locations of the individual exons in your annotation? Can you reliably use RNA-Seq coverage to identify untranslated regions (5' and 3' UTRs) for each of your genes? Of all of the introns in your gene model, how many were supported and how many were unsupported by results in the splice junction and transcript tracks? Be sure to investigate all regions that have high RNA-Seq coverage, especially in regions that do not correspond to the genes you have annotated. (However, some of the regions with high RNA-Seq coverage could be spurious if they overlap with transposon remnants or other repeats.) Note that genes that are expressed only at low levels or in a restricted set of tissues or developmental stages may not display significant RNA-Seq data for our species. You can assess the expression status of the orthologous gene in detail from the data available for *D. melanogaster* on https://flybase.org.

First Gene Analysis

As described above, you should **provide a detailed description of your general approach** with the description of the first gene discussed in the paper. Be sure to include screenshots documenting the following:

- Verification of the *D. melanogaster* orthologue by performing a *blastp* search of the Genscan prediction against the "Annotated proteins" database at FlyBase;
- Exon-by-exon *blastx* (or *tblastn*) search results;
- Close-up of the splice donor and acceptor sites (~20 nts), demarcating the possible splice sites you have considered and your final annotation, indicating the reading frame and phase (remember to verify that the orientation of the Base Position track is correct);
  - If your first gene contains more than 6 introns, you can show in detail the supporting evidence for the annotation of the first 3 introns. For the remaining introns, you can focus on any introns that are more challenging to annotate, but be sure to include all CDS in your final model table (see below) and in the checking of your model in the gene model checker.
- Translation start and stop sites.

Include the final model in a table, showing the exact exon coordinates, size, percent identity with *D. melanogaster*, reading frame, phase of the donor and acceptor sites. While you should **include a dot plot and protein alignment** from Gene Model Checker, a screenshot of **the Gene Model Checker checklist (i.e. green check marks) should not be shown if every test passes.** Figures like this do not hold informational content that cannot simply be described in the text (i.e. "all tests passed in Gene Model Checker") and is evident from the dot plot and protein alignments. However, you should include the Gene Model Checker checklist if it contains any warnings or errors (*e.g.*, non-canonical splice donor), along with the evidence you used to support these exceptions in your final model. Be sure to include a detailed justification in your report if the proposed gene model for your species differs substantially from the *D. melanogaster* orthologue. There will be differences — evolution happens! But you should be able to explain any substantial gaps in the dot plot, particularly large vertical or horizontal gaps (which correspond to large insertions or deletions compared to the *D. melanogaster* orthologue, respectively) and discuss why your model is better than any alternatives.

Once you have established the general approach, you can summarize your results for subsequent genes. Focus on anything unusual or different in the results of subsequent annotations. Each gene summary should include a summary table as described above and a final dot plot and protein alignment. **Do not include multiple dot plots and alignments if the coding exons are identical, just describe which isoforms are identical and mention that the dot plot and alignment figures represent the results for all of them. For highly similar isoforms you do not need to show the full final results from the Gene Model Checker in the main text, just focus your discussion on the variable sections and include the dot plot and alignment files in the appendix.**

You will need to report the exact position of each coding exon in every gene as a table in an Appendix, and as a separate file in the format described below. Remember when trying to precisely place intron/exon boundaries that introns (almost) always start with the two bases, 'GT', and end with 'AG'. Non-canonical splice sites (*e.g.*, GC splice donor site) should only be used if they are supported by conservation with *D. melanogaster*, or conservation with neighboring species (*e.g.*, see the "*Drosophila* Conservation (36 Species)" track in the genome browser for *D. melanogaster*), or with RNA-Seq data (*e.g.*, splice junctions).

Finally, report on the one *Drosophila willistoni* gene you annotated following the instructions above for the single F element gene.

Transcription Start Sites (TSS)

As we have discussed, we would like you to use the TSS protocol to define the TSS position(s) and TSS search region(s) for the genes in your *D. kikkawai* project region. Begin your TSS analysis by characterizing the shape of the promoter for the orthologous gene in *D. melanogaster* based on the TSRchitect results. Define the genomic region to search for the TSS's of the *D. kikkawai* gene based on the gene structure of the ortholog in *D. melanogaster*, the *blastn* search result comparing the initial exon from *D. melanogaster* against the *D. kikkawai* F element scaffold, and the available *D. kikkawai* RNA-Seq data.

Examine the strand-specific RAMPAGE data in the genomic region you have defined. Use the TSRchitect RAMPAGE peaks and RAMPAGE read density evidence tracks to define the shape of the promoter, the TSS position(s), and the TSS search region(s) if possible. Depending on the RAMPAGE results, this step might be straightforward or it could be quite difficult. Remember the goal is not to give a historical record of the analysis, but to describe the evidence that either support or refute your TSS annotations in a way that clearly demonstrates your analysis. Negative results should, at a minimum, be discussed briefly in your report to assure the reader that all sources of evidence were examined in your analysis.

Repeats

Following the repeat analysis protocol, report on the repeat distributions in your project region in general, as well as the results of your analysis of the F element comparison region you analyzed. As described, group members should perform the repeat analysis on different adjacent regions of about half a million bases. The results of the D element should then be compared to the region you selected from the F element. Summarize the repeat class distributions, and the most common transposons (by size) for your analysis regions and then compare them.

Synteny

Generate a simplified map of your *D. kikkawai* project region, including genes and possibly any other features you have found. Analyze the genes from your project, and assess synteny with *D. melanogaster*. Using one of the *D. kikkawai* D element genes in your project region, identify the location of the corresponding ortholog on the *D. melanogaster* F element. Generate a map of the *D. melanogaster* genome centered around this gene which covers approximately the same number of protein-coding genes as your *D. kikkawai* project region. Note which Muller element each gene is on in *D. melanogaster*.

Using these two maps, compare the relative gene order and orientation in *D. kikkawai* and *D. melanogaster*. Remember that **the orientation of your *D. kikkawai* project is arbitrary** (based on computer generated files), so you should focus on the **relative** orientations of the genes when analyzing synteny. (The project region is considered to be syntenic with *D. melanogaster* if the reverse complement of the *D. kikkawai* project has the same gene order and orientation as *D. melanogaster*.)

If synteny has been preserved, the genes in your project will all be from the same region of the *D. melanogaster* genome, and your comparison will show a one-to-one map of the orthologs from one genome to the other (two horizontal lines). If the regions are completely syntenic, the genes will be in the same relative order and orientation in the two species. If synteny has not been preserved, please include a comparison based on each gene in your project to the same gene and flanking regions in *D. melanogaster* (several horizontal lines, stacked up). Look for evidence of events (such as inversions, transpositions, etc.) that could have easily occurred during evolution that could explain the changes in synteny. If there is no set of simple events, then consider the changes as “complex” and report as such.

Expansion Analysis

Since the *D. kikkawai* F element is substantially larger than the *D. melanogaster* F element, we would like you to explore how the expansion affects gene characteristics by comparing the sizes of the genes in your D element project with the genes in the region you analyzed in the repeat analysis of the F element. To simplify this analysis, focus on the “coding span” (i.e., the region between the start of translation and the stop codon) of one isoform for each gene. If a gene has multiple isoforms, pick the isoform which has the largest number of CDS's for your analysis. Calculate the mean size of the coding regions and the introns across all genes and compare these to the mean size of the orthologous genes from *D. melanogaster*. This should be done for both the genes in your region and a similar number of genes from the F element found within the region you analyzed in your repeat comparison. For the *D. willistoni* F element genes, you can use the gene models in the "BLAT Alignments of NCBI RefSeq Genes" track. Finally, interpret your results. Do you see evidence for expansion of the *D. willistoni* F element genes compared to the *D. melanogaster* F element genes? What about expansion of the D element genes in your region? When considering CDS regions independent of introns, which of these subregions have expanded? One of them or all of them? How do the expansion rates compare in the two Muller elements?

Further Investigations

As time permits, we encourage you to explore one of the genes or features in your project in more detail. What is the presumed function(s) of one of the genes? Can you generate evidence in support of these functions? Are key regions of the predicted protein conserved? Looking at the "Gene Ontology" and "Protein Domains" sections of the FlyBase report, and the Clustal Omega analysis results may be informative. Is this gene part of a known biological pathway in *D. melanogaster* (see the "Pathways" section of the FlyBase Gene Report)? Anything is possible here but not required.

Discussion

One to three paragraphs. Start with a "final map" in this section. Use custom tracks to show the entire project region with all the final annotations of all CDS-based gene models for all members of the group. Point out the genes and TSS's you were directly involved in annotating. Summarize the accomplishments of the group. In subsequent paragraphs, put your work in context. How do your findings support or differ from prior analyses? Refer in particular to the results of Leung *et al.*, 2017 and questions raised in the current grant proposal.

Note: Given that every gene is different, your ability to use the above format may vary. If your particular results are not compatible with the above recommended format, consult with the instructors on how best to adjust your writing to fit your results. If you have any questions regarding the best way to present your observations and conclusions for your particular project be sure to ask. Good luck and good hunting!

Addendum on Writing

As always, high standards are expected. Clarity of exposition is most important. The reader should be able to follow your reasoning without difficulty. Include screenshots with key evidence supporting your conclusions. Use highlighting, circles and arrows in your figures to focus the reader's attention on what's relevant. Remember that logic, not history, should dictate your presentation! Avoid "stream of consciousness" writing. Scientific writing has little in common with Faulkner. Show us the evidence you need to convince us your conclusions are correct. You do not need to show us experiments that failed, or describe ideas that did not yield interesting results, unless that excursion resolves an obvious question — but do present interesting challenges that you were able to resolve. Use the past tense to describe your actions, but the present tense to describe your project, and the *D. kikkawai* genome — they are still in existence. Use words correctly and precisely. BLAST is not a verb! Avoid the temptation to make up a new verb to describe a whole process. Avoid slang. Remember that some of your readers do not speak English as their first language.

We use figures and tables — not graphs, charts, etc. Tables have rows/columns of numbers, and everything else is classified as a figure. Make sure that the figures are legible, specifically that one can read the labels for each axis and other printed information that is important to interpreting the figure without magnification. Include a figure title and an informative figure legend. A figure showing results of a BLAST search should **always** identify the query and the subject of the search, as well as the BLAST program (e.g., *blastp*, *tblastn*) used. Figure legends should be very specific; the readers should NOT have to guess what they are looking at!

Please upload the MS Word version of your report to Canvas. You might want to also upload a PDF version of the report, as multi-component figures (such as those with added arrows or boxes) can shift or come apart depending on the platform and version of Microsoft Word. Alternatively, you can create the multi-component figures in a different program (*e.g.*, PowerPoint, Adobe Illustrator), and then insert the composite image into the Word document.

Please use the following format for references:
Smith, JD, EF Jones, and JJ Green (2010) All the world's a phage. Science 210: 33 – 44.

Write out the names of all authors up to 12 before switching to *et al.* References can be cited in the text by number, or by *Smith et al. (2010*), or by (*Smith et al. 2010*). Find the papers you need on PubMed or cited in your readings; **cite the journal** as above**; do NOT cite PubMed or the URL**.

Specific Style Guide for this Paper

1. Fly species **and** gene names are always italicized, protein names are not.
2. Fly gene names are case sensitive — make sure that the gene names match the official FlyBase gene symbol! Fly protein names are not italicized and should be capitalized in all cases irrespective of the gene name.
3. Spell out the abbreviation the first time (*e.g.*, "*Drosophila melanogaster*"), and then you can use abbreviations thereafter (*e.g.*, *D. melanogaster*). The abstract and text are separate for this purpose.
4. Specify the species for the sequence you are using.
5. Be very careful with the use of exon names, particularly in cases of alternative splicing. I prefer "2a, 2b…" etc. rather than "3_10903_1," although admittedly the latter is more precise, and may be necessary if the gene has a large number of alternative exons and isoforms.
6. Please use "splice donor" and "splice acceptor" rather than designating 5' or 3', since the latter can be confused between introns and exons and can be confounded by the strand.
7. Each page should have your last name and the page number in either the header or the footer.

In Scientific Writing in General

1. Do not be general when you can be specific (say "4 out of 5 cases" not "most cases').
2. Do not use "value-loaded" language; words like "good" or "bad" have little meaning in science — support your argument with facts and observations, with estimates of error as needed.
3. Keep the language impersonal.
4. Avoid slang, jargon, and other language shortcuts. Remember that when you publish in the scientific literature, many of your readers will be people who do not speak English as their first language.

Appendix

In addition to your written report, you should also complete the GEP Annotation Report that provides the documentations for your coding regions and TSS annotations, as well as the Repeat Analysis Worksheet. The GEP Annotation Report and the Repeat Analysis Worksheet are available on Canvas.

You should also submit the combined GFF, transcript, and peptide sequence files for all the gene models in your project as part of the appendix. These files will be needed in the subsequent analysis of the compiled results. There will be a separate assignment on Canvas where you can submit these appendix files.

For any predicted gene isoform, you should have three files:

1. A fasta formatted file of the protein sequence.
2. A fasta formatted file with the nucleic acid sequence which codes for the protein (make sure it translates!)
3. A tab-delimited file in GFF format for every isoform of every gene found in your project.

All three files are generated by the Gene Model Checker (under the "Downloads" tab) for each isoform. Remember that you should generate GFF, transcript, and peptide sequence files for **all** isoforms, irrespective of whether they have the same coding regions. You can combine the individual files for each annotated gene into a single file using the Annotation Files Merger tool available through the "Annotation Files Merger" link on the "Links to GEP Annotation Tools" page on Canvas. If you believe you have found an error in the consensus sequence, please discuss with Wilson how to use the Sequence Updater to generate a VCF file and include this file in the appendix.

These files should be submitted to the appendix files assignment on Canvas.