Annotating your *Drosophila* contig: final presentation and written report

The final will be made up of two parts, a 10-minute oral presentation and a written report.

The oral presentation:

Give a presentation in which you briefly discuss each of the features (putative genes) of your sequence(s), the distribution of repetitious elements, and synteny with *D. melanogaster*. Your paper will discuss your best determination of the exact nature of each feature (possible gene) and your reasoning that led you to this conclusion. However given the 10 minute limit on your talk you will probably want to mention some genes only briefly to give yourself more time to focus on those genes which were particularly interesting or difficult to annotate. Remember there is no right answer; this is the first time that this sequence has ever been looked at by anyone. The goal is to come to the most reasonable interpretation of the data and be able to articulate the reasoning behind your hypothesis. While we are interested in all features of genome organization, especially as they relate to euchromatin vs. heterochromatin, we are giving special attention this year to estimating the Transcription Start Sites (TSS) of our genes, and searching for regulatory motifs or common features. Thus you will also want to discuss your findings on this problem for those genes where you were able to estimate a TSS. Note that *D. biarmipes* and *D. elegans* are fairly closely related to *D. melanogaster*, so the more extensive analysis of gene expression that has been done with *D. melanogaster* will give us clues.

You may use PowerPoint (recommended), overheads or simply draw on the blackboard, but figures of some kind are required, as they always play an important role in getting your point across to your audience.

The written report:

The other part of the final will be your written report. Eight main sections (subdivided as you see fit) are recommended: Abstract, Introduction, Genes, TSS Estimates, Gene Evolution, Repeats, Synteny, and Discussion/Conclusions. We also ask that you include an Appendix with appropriate data files (see below). Your report should be submitted both as hard copy and in electronic form. The report should integrate text with figures, making a single, coherent document. This document will be the basis for future research on this topic by following Bio 4342 classes, will be the source of information for our joint publication, and will 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). 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 evolution of genes and chromosomes? For searching for regulatory motifs? Attempt 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 of the entire sequence you are assigned that indicates the size and position of each putative feature you will

investigate. The Genscan prediction figure (preferred) or a simple UCSC genome browser screenshot can serve as a starting point. Number the features that you will investigate for ease of communication.

Genes

This section will be significantly larger (reflecting the number of genes in your project), and should be subdivided with a section for each putative feature. You should present here a more detailed analysis of any genes, pseudo-genes or partial genes you find in your project. Use the Genscan predictions as an organizer for your presentation. Remember that in flies (based on results from *D. melanogaster*), pseudogenes are rare, so it is unlikely (but not impossible) that you will find pseudogenes in your project. Since the ends of your project were generated based on available predictions, it would not be surprising if the ends land in the middle of a gene and thus you will only find a partial fragment. Despite our finishing efforts, 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 as well (homology search with *D melanogaster*, RNA-seq, etc.). In particular, Genscan will not predict RNA-only genes. 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. TopHat splice junction predictions, Cufflinks and Oases assembled transcripts), and conservation tracks for the other Drosophila species. (Utilizing data from additional species may be helpful.) For each gene in your project include the FlyBase accession number and the name of any *D. melanogaster* gene that you consider likely to be orthologous. For each gene you should determine as closely as you can the exact location of each coding exon. 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 peptide sequence alignment comparison to *D melanogaster*. Every difference between your model and the *D. melanogaster* ortholog represents an assumption of evolutionary change that underlies your gene model. The best gene models will always 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. The dot plot itself, and a discussion of important differences, should be part of your report.

In addition to sequence conservation, you should also integrate information from high-throughput sequencing of mRNA into your gene annotations, working from the RNA-Seq track. 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" track. In addition to using the modENCODE 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' UTR) for each of your genes? Of all of the introns in your gene model, how many were supported and how many were unsupported by TopHat or Cufflinks predictions? 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. Note that genes that are expressed only at low levels or in a restricted set of tissues may not display significant RNA-Seq data for our species. You can assess the expression status of a given gene in detail from the data available for *D. melanogaster* (consult the "High-Throughput Expression Data" section [under "Expression Data"] of the FlyBase Gene Report).

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 electronic file in the format provided. Remember when trying to precisely place intron/exon boundaries that introns (almost) always start with the two bases, 'GT', and end with 'AG'.

Transcription Start Sites

As we will discuss in class, we would like you to estimate the Transcription Start Sites for as many of your genes as possible. Relevant data will include conserved upstream regions, RNA-seq data, the presence of known conserved motifs, and the like. Start by mapping the untranslated 5' exons using more sensitive BLAST parameters and examining the match to RNA-seq data. Search for short sequence matches to known promoter motifs. As time permits, we might also use the MEME suite to look for conserved motifs.

Gene Evolution

Do a Clustal analysis on at least one gene, and comment on the evolution of your gene. In analyzing a gene using Clustal, find orthologous genes from a variety of species, run a Clustal analysis on the protein sequence from at least four different species, and report on the results. What can you say about the evolution of the gene? Depending on the success of your efforts to estimate a TSS, try the Clustal analysis with DNA sequences, using a region that encompasses the putative TSS and any suspected regulatory motifs, and the first coding exon. In this case, you will want to use much more closely related species, specifically Drosophila in this subgroup. Comment on your results.

Repeats

The analysis of repeats is most interesting with respect to the questions of heterochromatin and euchromatin. Prepare a table that lists all the large repeats (>500bp) that you found within your project. Include both those repeats picked out by RepeatMasker (generally remnants of repetitious elements; use the reference set tailored for the species you are working on) as well as any sequences that you identified by a BLAT search against the whole genome assembly. You should also calculate and report the percentage of repetitive DNA. In some species of Drosophila, a subset of the repeats identified by RepeatMasker has been found to be derived from genomic fragments of *Wolbachia* that have been integrated into the genome. Check to see if that has happened in your species / project and include the results of that check in your report.

Synteny

Generate a map of your contig including genes and large repetitious elements, and then compare this with a map of the *D. melanogaster* genome over the same number of kb centered around each of the genes in your project. Remember that the orientation of your project is arbitrary (based on computer generated files) and be sure to consider both orientations of your project when analyzing synteny. Note which Muller element the gene is on in *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 be of one map to the other (two lines). 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 lines, stacked up). Determine the frequency of genes and of repetitious elements (number per kb) in these regions of *D. melanogaster*, as well as looking for evidence of events (such as inversions, transpositions, etc) that occurred during evolution, if possible.

Discussion
One to three paragraphs. Put your work in context. How do your findings support or differ from prior analyses? Refer in particular to the results of the Drosophila Twelve Genomes Consortium (2007) and to the prior findings of the GEP (Leung et al, 2010; 2014). If your region has perfect synteny (minimizing the discussion above), you could include your final map in this section.

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 screen shots with key evidence supporting your conclusions. Use highlighting, circles and arrows in your figures to focus the readers attention to what's relevant. Give details on exon splicing for your first gene, then only for unusual cases.

Note: Given that every gene is different, your ability to use this format may vary. Please consult the instructors concerning any questions about your particular sequence on how best to present your observations and conclusions. Good luck and good hunting!

Appendix
For your appendix you should include various sequence files that will be needed in subsequent analysis of the compiled results.

For any predicted gene 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. You can combine the individual files for each annotated gene into a single file using the Annotation Files Merger tool available through the GEP web site (under Projects → Annotation Resources). In addition to the appendix files, you should complete the annotation report form for each gene you have annotated. The annotation report form is included in annotation project package. A separate form is used to report putative Transcription Start Sites.

For any new repeat include a fasta formatted nucleic acid file of the sequence of the repeat you found. For your Clustal analysis within Drosophila, include the collection of fasta sequence you used in your analysis.

<u>These files are not needed for printed copies</u> given to additional readers, but must be submitted to the GEP system and provided to Shaffer/Leung.