Research Problem, Spring 2018

The scientific problem we will tackle this year is a question of comparative genomics: can we distinguish heterochromatic and euchromatic genes based on sequence characteristics and organizational features of the genes in these different environments? In particular, can we discriminate characteristics of the transcription start sites? What insights can we gain to elucidate how genes can be expressed when resident in a heterochromatic environment, normally a condition that promotes silencing? In *Drosophila* (the fruit fly), the small fourth chromosome (sometimes called the "dot" chromosome or F element) is unusual in that it appears to be entirely heterochromatic - packaged in a relatively condensed form, replicated late in S phase, exhibiting no meiotic recombination, etc. An examination of the DNA sequence of *D. melanogaster* indicates that the 1.3 Mb long arm has a normal gene density (~80 genes), but a three-fold higher frequency of repetitious sequences (thought to be targets for heterochromatin formation) than the other chromosome arms, which are euchromatic. Most of the genes on the fourth chromosome are associated with silencing marks (enriched histone H3K9me2/3, presence of HP1), yet these genes are expressed in this heterochromatic environment. So how do these genes function?

Students enrolled in Bio 4342/434W, working with the Genomics Education Partnership, have sequenced and annotated the F elements (and comparison D element domains) of several Drosophila species across 40 million years of evolution. We have published three papers on the evolution of the F element, with all contributing students as co-authors (Leung et al, 2010 Genetics 185: 1519-34; Leung et al 2015 G3 5: 719-40; Leung et al 2017 G3 7: 2439-2460). We are now improving and annotating F elements from a group of species close to *D. melanogaster*, which will help us to search for regulatory motifs. Each student in Bio 4342 will take on the challenge of finishing ~100 kb, from *D. takahashii* this year, and annotating a ~40 kb project from the F element of *D. takahashii*. We will be able to compare our results with data from the other fly species, looking for conserved motifs – "phylogenetic footprints" – in collaboration with Dr. Jeremy Buhler from Computer Science. The comparative analysis should tell us much more about this interesting chromosome and the genes that reside there, potentially giving insights into gene regulation.

In "finishing" (checking the DNA sequence assembly), your challenge will be to spot and correct potential sequencing errors, calling additional sequencing reactions as needed. In your annotation project, you will want to address the following questions: Which genes and/or pseudogenes are present in your project? What models can you construct – can you find homologues for all of the isoforms reported in *D. melanogaster*? What is the gene density? What types of repetitious elements are present, at what density? What can be said from a comparison of this region (defined by the genes present) with other Drosophila species? Is the region syntenic with *D. melanogaster*? In our subject species, can we predict the transcription start sites of the genes? Can any non-coding conserved sequences be detected – potential regulatory elements? As time permits, we may want to look at codon bias and DNA melting profiles, as we have found these to differ between regions with high rates of recombination (euchromatin) and regions with low rates of recombination (heterochromatin). If we can, we will also want to expand our search for potential regulatory elements, in particular noting similarities among the 5'upstream regions of fourth chromosome genes, based on the hypothesis that genes that function within a heterochromatic environment might exhibit special characteristics. The

answers to these questions should tell us more about the relationship between DNA sequence organization, chromatin packaging, and gene regulation. As time permits, we will also explore other characteristics of a gene of your choice, including protein folding.