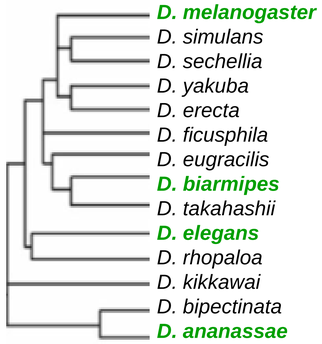


Drosophila phylogenetic tree. Green indicates species of interest. (Adapted from W. Leung, GEP)

Different Species, Same Genes?:

Constructing Gene Models in Related Drosophila Species

Max Blumenthal ‘16, Tim DeRosa ‘16, Mark Scheuerman ‘14, Dr. Amy Hark

Funded by: Biology Department
With: The Genomics Education Partnership (GEP)

Introduction

We identified and described orthologous coding sequences in *D. elegans, D. ananasse,* and *D. biarmipes* by comparison to *D. melanogaster.* (Shown above in green)

- ◆Collect
  - Working with a 50kb section of sequence at a time, we used the BLAST algorithm to identify conserved regions.
- ◆Analyze
  - We constructed gene models for the identified conserved regions. Our generated gene model data was checked via a Dot Plot viewer.
- ◆Synthesize
  - Our findings are currently used for a larger goal of identifying key features of heterochromatic domains across eukaryotic species.

Material and Methods

Identify the likely ortholog in *D. melanogaster*. Determine the gene structure of the ortholog.

↓

Map each exon of the ortholog to the project sequence using NCBI BLAST.

↓

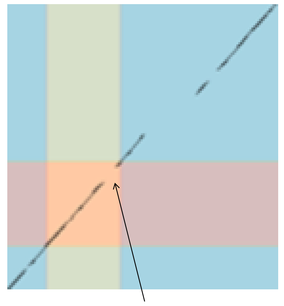
Use mapping data to construct a gene model to identify the start and stop positions and splices sites for each coding sequence.

↓

Use the Gene Model Checker to verify the gene model; to confirm no in-frame stop codons, non-canonical splice sites, etc.

Results

To check for accuracy of the gene model we created a Dot Plot through Gene Model Checker.

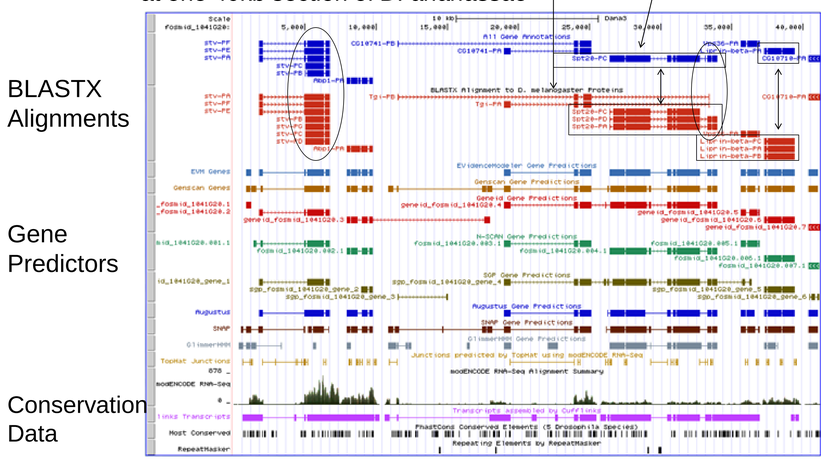


Gaps at exons intersections might indicate an improper splice site.

The Dot Plot used our gene model and compared it to *D. melanogaster*. When analyzing the Dot Plot, we had to consider any deviations from the axis along with gaps at exon intersections.

Computer Generated Genes **Student Generated Genes**

The UCSC genome browser, looking at one 40kb section of *D. ananassae*



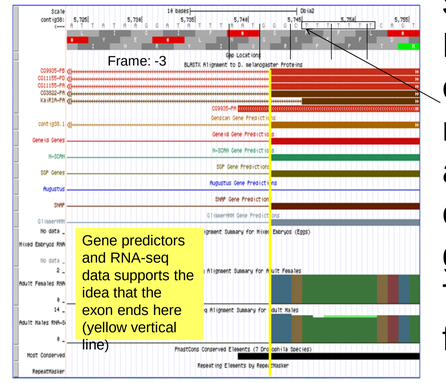
As shown, the BLASTX (computer generated) track can show too many or too few exons (ovals) as well as too many or too few isoforms (rectangles).

Conclusion

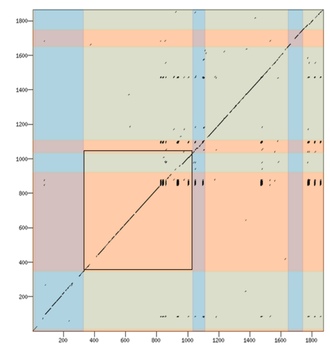
A final report was drafted and sent to the Genome Education Project at Washington University to have our work compared to other students who annotated the same contigs. Once reviewed, data is then compiled to produce a final annotated sequence for publishing.

Annotation Challenges

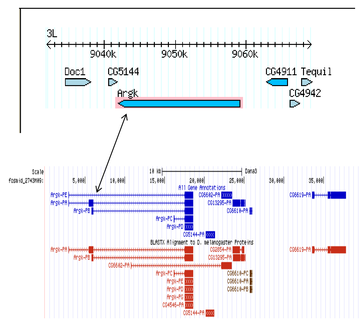
Problems during annotation can include:



Sequencing Error: Pyrosequencing error creates too many T nucleotides and shows no stop codon at end of gene. Removing a T shifts the reading frame “adding stop.”



Two Exons For One: In this instance, we see that *D. ananassae* (potted on vertical axis) has two exons for the one exon *Drosophila melanogaster* has (plotted on horizontal axis).



Chromosomal Inversion: The gene Argk in *Drosophila ananassae* is surrounded by different genes than in *Drosophila melanogaster,* suggesting a chromosomal inversion.

Acknowledgments

The Genomics Education Partnership (GEP)
Available at: http://gep.wustl.edu
Funded by: HHMI and NSF