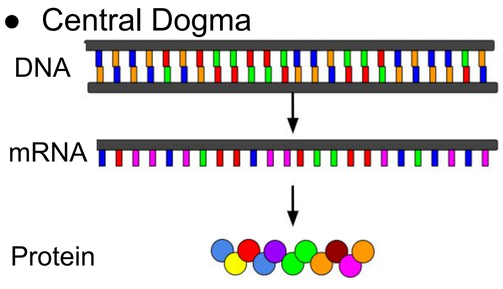
Gene Annotation: Finding and characterizing genes in genomes

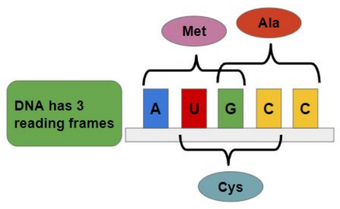
Malcolm Arnott ‘22, Blake Tellinghusen ‘23, Vy Lam ‘22, Jack Allen ‘21, Sam Alper ‘19, Juliana Choza ‘20, and Sami Zimmerman ‘19, *Department of Biology[1], BCMB[2], Psychology[3], Lewis & Clark College.*

Introduction

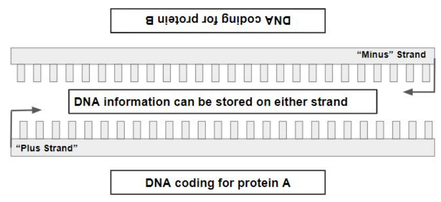
- Central Dogma



- Reading Frames & codons



- DNA strands



- Gene structure
  - Both have:
    - Beginning, middle, and end
    - Irrelevant components.
      - Introns are like commercials. They can be removed without altering the “story”

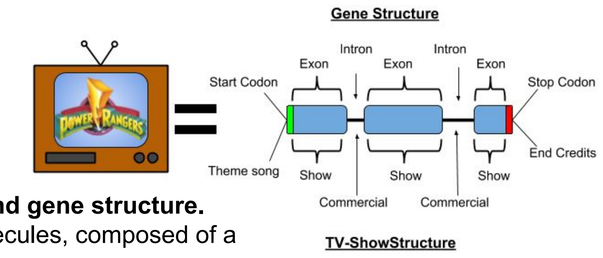


Figure 1. Schematic of the central dogma in biology and gene structure. Genomic information in an organism is stored in DNA molecules, composed of a sequence of nucleotides. Genes have a defined structure.

Terminology

- *Open Reading Frame (ORF):* Sequence of DNA without a stop codon
- *Exon:* Segment of a gene expressed as a protein
- *Intron*: Gene segment removed during splicing
- *Splicing:* Removal of introns
- *Ortholog:* Homologous (similar) gene sequences

Important Sequences

- Start codon: ATG
- Stop codons: TAA, TGA, TAG
- Splice sites: DNA sequences prompting intron removal
  - Donor: GT
  - Acceptor: AG

Methods

Gene annotation process:

1. Select the gene prediction
2. Identify the ortholog in *D. melanogaster*
3. Review structure of ortholog
4. Determine approximate locations of exons using BLAST/RNA-Seq
5. Identify splice sites (GT/AG) and determine phases
6. Verify splice sites with tophat where possible
7. Verify annotation with Gene Model Checker

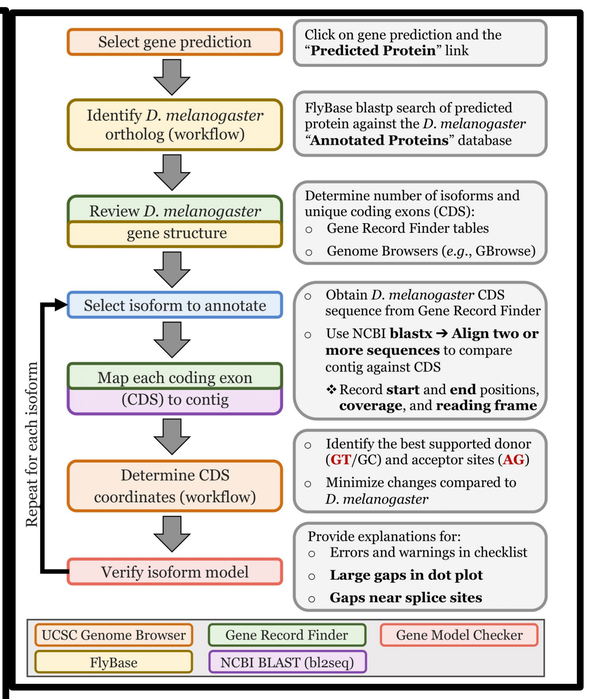


Image Credit: Genomics Education Partnership

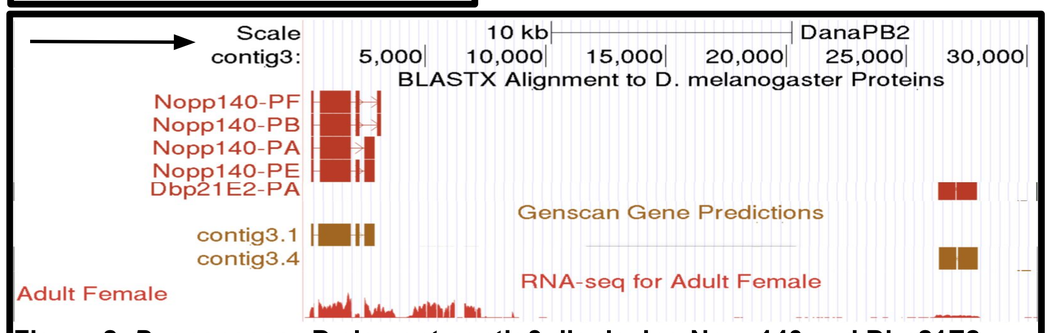
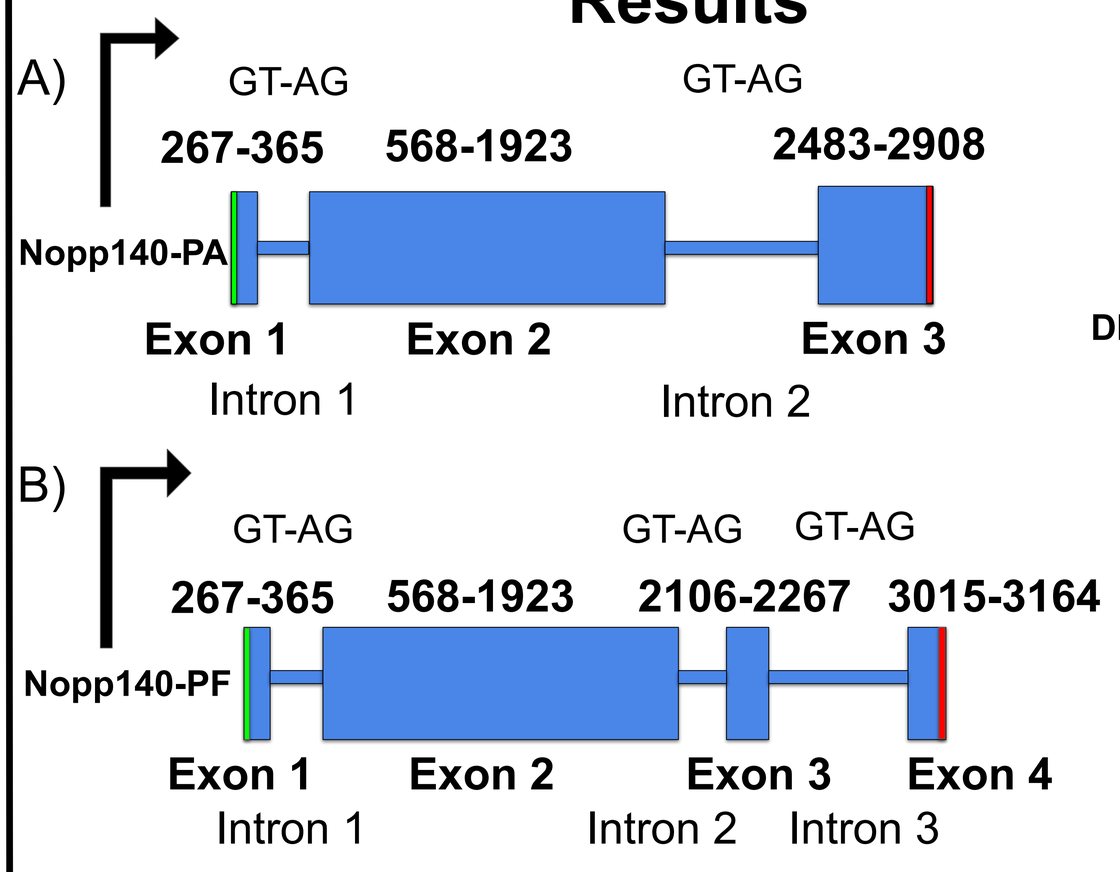


Figure 2. *D. ananassae* D element contig3 displaying Nopp140 and Dbp21E2. Pictured above is an edited image of the genome browser containing the main evidence used to annotate genes. The first set of evidence tracks display BLASTX alignments to annotated proteins in *D. melanogaster*. The second set of evidence tracks display an example of computer generated gene prediction alignments. The third evidence track displays RNA-Seq data displaying gene expression data in mRNA

Results



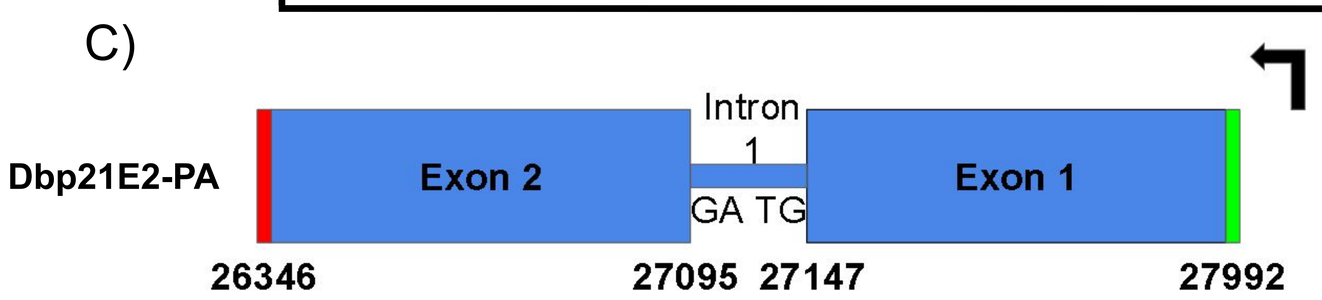
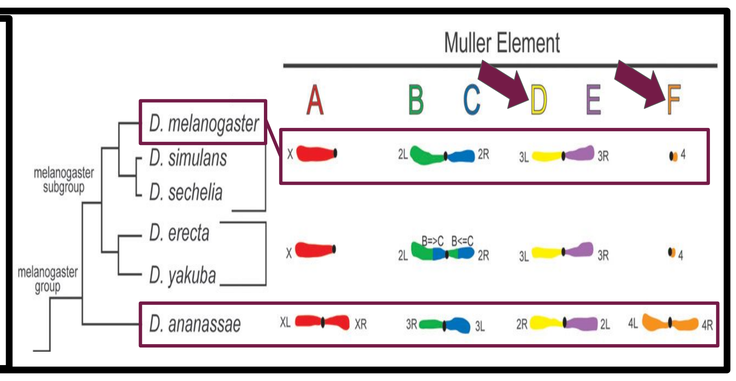


Figure 3. *D. ananassae’s* Nopp140 and Dbp21 Gene Models. Gene models including labeled introns (standard) and exons (bold), splice donor (GT) and acceptor (AG) sites, start (green box) and stop codons (red box), and coordinates for each exon. Arrows denote beginning and direction of transcription. A) Nopp140-PA gene model. B) Nopp140-PB gene model. C) Dbp21E2-PA gene model.

Future directions

- Dot plots of our annotations (comparison of sequence similarity to *D. melanogaster*) did not come out as expected. This highlights the need for independent corroboration.
- The genes P5CDh1 and CG14565 still require annotation, as well as other contigs of *D. ananassae*. Contig 12 of Muller D element and contig 59 of Muller F element will be annotated next.

Figure 4. Muller D and F- element of *D. melanogaster* and *D. ananassae*. The image indicates the F-element of *D. ananassae* is extended compared to the F-element of *D. melanogaster*; this suggests the gene structures would be different. Identifying those differences is the goal of this project.



Conclusions: Why does this matter?

- The process of Gene Annotation can be used to find and characterize genes in previously uncharacterized genomes.
- Applications:
  - These data allow us to answer questions about evolutionary conservation and protein function across species (including in humans).
    - For example: If there is a human protein of unknown function involved in a disease we can search the genomes of other species to find a similar protein (an ortholog) to make an educated guess as to what the human function is. Gene annotation allows for these educated guesses.

Acknowledgements

- Genomic Education Partnership. Prof. Norma Velázquez-Ulloa.