Annotation for *D. virilis*

Chris Shaffer July 2012

Last Update: 12/2021

Annotation

- “Big Picture” of annotation and then one practical example
- This technique may not be the best with other projects (e.g., parasitoid wasps, corn)
- The technique is optimized for projects with a moderately close, well annotated neighbor species (i.e., *Drosophila melanogaster*).

Goals for F Element Project

- Identify & annotate genes in your project
  - for each gene identify and precisely map (accurate to the base pair) all CDS’s
  - do this for ALL isoforms
- Optional curriculum not sent to GEP
  - Clustal analysis (protein, promoter)
  - Non-coding genes

Evidence Based Annotation

- Human curated analysis
- Much better outcome than standard *ab initio* gene finders
- Goal: Collect, analyze and synthesize all evidence available to create most likely gene model:
  4591-4688, 5157-5490, 5747-6001

Collect, Analyze, and Synthesize

- Collect
  - Genome Browser (https://gander.wustl.edu)
  - Conservation (BLAST searches)
- Analyze
  - Reading browser tracks
  - Interpreting BLAST results
- Synthesize
  - Annotation Instruction Sheet
  - Annotation Strategy Guide

Evidence for Gene Models

Interpret in the context of basic biology (in general order of importance):

1. Expression data
2. Conservation
3. Computational predictions
4. Tie-breakers of last resort
Basic Biology (context)

- Known biological properties
  - Coding regions start with a methionine (M)
  - Coding regions end with a stop codon (TAA, TAG, or TGA)
  - Genes are only on one strand of DNA
  - Exons appear in order along the DNA
  - Introns start with GT (or rarely GC)
  - Introns end with an AG
- CCTAGA GT ACCAGAG...

Expression: RNA-Seq

- Positive result very helpful
- Negative result less informative
  - No transcription ≠ no Gene
- RNA-Seq evidence tracks:
  - RNA-Seq read coverage
  - Junction prediction (e.g., TopHat, regtools)
  - Noisy, Noisy Data

Conservation

- Coding sequences evolve slowly
- Exon structure changes VERY slowly
- Similarity implies homology
- D. melanogaster very well annotated
- Use BLAST to find amino acid conservation
- As evidence accumulates indicating the presence of a gene, more time/effort is justified when looking for conservation

Computational Evidence

- Assumption: there are recognizable signals in the DNA sequence that the cell uses; it should be possible to detect these computationally
  - Many programs designed to detect these signals
  - These programs do work to a certain extent, the information they provide is better than nothing; high error rates

Computational Evidence

- DNA sequence analysis:
  1. *ab initio* gene predictors – exons vs. genes
  2. Splice site predictors – low quality
- Summary / organization of observations
  1. Multi-species conservation – low specificity
  2. Blastx alignment (WARNING! Students often over interpret this data, use with caution or ignore)

Genome Browser

- Lots and lots of evidence!
- Organized for your use
Browser is Incomplete

• Most evidence has already been gathered for you and is shown in the genome browser tracks

• However, some conservation data must be generated by the annotator at the time of annotation

1. Assignment of orthology
2. Look for regions of conservation using BLAST
   - Collect the locations (position, frame and strand) of conservation to orthologous protein
   - If no expression data available, these searches may need to be exhaustive (labor intensive)

Typical Annotation Workflow

1. Identify the likely ortholog in *D. melanogaster*
2. Find gene model of ortholog, view structure and get exon sequences (FlyBase & Gene Record Finder)
3. Use blastx to find conservation to exons; search exon by exon, note position and frame
4. Based on data generated above, plus other data in browser create gene model; identify the exact base location (start and stop) of each CDS (coding exon) for each isoform
5. Confirm your model using Gene Model Checker and Genome Browser

Basic Procedure (graphically)

To annotate, find ends of exons (i.e., start codon, donor (GT) and acceptor (AG) splice sites, and stop codon). Use conservation and browser info to create model making sure to select compatible splice donor and acceptor sites in order to maintain an open reading frame that encodes conserved amino acids after splicing.

Once these have been identified, write down the exact location of the first base and last base of each exon. Use these numbers to check your gene model in the Gene Model Checker.

Evidence for Gene Models

Evidence:
1. Expression data
2. Conservation
3. Computational predictions

• 1 & 3 are already done for you (Genome Browser)
• 2 is generated by annotator
Searching for Conservation

- Search for conservation “to do” list
  1. Use protein BLAST searches to assign the putative ortholog based on sequence similarity
  2. Use FlyBase to view the gene structure of the *D. melanogaster* ortholog
    - Gene structure is usually conserved among different *Drosophila* species
  3. Use BLAST to find conserved amino acids at single exon level

Example Annotation

- Open 4 internet browser tabs:
  1. Genome Browser: thegep.org/browser
     - *D. virilis* - Mar. 2005
  2. FlyBase: https://flybase.org
     - Info from most recent *D. melanogaster* genome annotations
  3. Gene Record Finder: thegep.org/finder
     - Info from release 6.43 of the *D. melanogaster* annotations
     - BLAST home page → blastx → click the checkbox:

Example Annotation from *D. virilis*

- Browse/Select Species: *D. virilis*
- Assembly: Mar. 2005 (GEP/Annot. *D. virilis* ppt)
- Position/Search Term: chr10
  - chr10 is a fosmid from 2005
- Click on GO

Investigate 10.4

- We will focus on 10.4 in this example
- To zoom in on this gene enter: *chr10:15,000-21,000* in the position box, then click on the “go” button

Find Ortholog

- If this is a real gene, it will probably have at least some homology to a *D. melanogaster* protein
  - Step 1: do a BLAST search with the predicted protein sequence of 10.4 to all proteins in *D. melanogaster*
Find Ortholog

- Click on one of the exons in gene 10.4
- On the Genscan report page, click on the "Predicted Protein" link
- Select and copy the sequence
- Do a blastp search of the predicted sequence to the D. melanogaster "Annotated proteins" database
  - Tab 2 / FlyBase

Results of Ortholog Search

- The alignment looks right for D. virilis vs. D. melanogaster — regions of high similarity interspersed with regions of little or no similarity
- Same chromosome supports orthology
- We have a probable ortholog: "mav"

Gene Structure Model

- We need information on "mav"; What is mav? What does its gene look like?
- Use the Gene Record Finder
  - Under "Resources & Tools" on the F Element project page (https://thegep.org/felement/)
  - Search for "mav" and click on “Find Record”

Gene Structure Model

- Click on the “View in Gbrowse” link under the "Gene Details" section to view the mav gene in FlyBase Gbrowse

Gene Structure Model

- The results show a significant hit to the gene “mav” (the PA names the isoform)
  - Note the large step in E value from mav to next best hit gbb; good evidence for orthology

Gene Structure Model

- We need information on “mav”; What is mav? What does its gene look like?
- If this is the ortholog, we will also need the amino acid sequence of each exon.
- Combination of two sites:
  - Use GEP’s Gene Record Finder (Tab 3)
    - TABLE list of isoforms; exon sequences
  - Use FlyBase (Tab 2)
    - GRAPHIC representation of gene; relationship among exons
Gene Structure Model

• The CDS usage map of the Gene Record Finder shows which isoforms use each CDS:

  Gene Record Finder table format

Gene Structure Model

Terminology:
- Gene span
- Primary Exons
- UTR's
- CDS's
- mRNA
- Protein

Gene Structure Model

• We now have a putative gene model (one isoform with 2 CDS's)

• For your final report you would annotate all isoforms for all the genes you identify in your project

Investigate Exons

• Use the coding exon sequence to search the project
  - Where in this project are sequences which code for amino acids that are similar?

  - Best to search entire project DNA sequence (easier to keep track of positions) with the amino acid sequences of exons

Investigate Exons

• Click on the exon to get amino acid sequence of each CDS

  - Under Tab1 (Genome Browser displaying chr10 of D. virilis)
    - Click on the “DNA” button under the “View” menu
    - Change the “Position” box to “chr10” (if you want the entire chr10 sequence, not just the 15000-21000 region)
    - Click on the “get DNA” button

  - Under Tab 3 (Gene Record Finder), make sure you have the peptide sequence for exon 1 of the D. melanogaster mav gene

  - These two tabs now have the two sequences you are going to compare
Investigate Exons

• Copy and paste the genomic sequence from Tab 1 into the top box (Query) of Tab 4
• Copy and paste the peptide sequence of exon 1 from Tab 3 into bottom box (Subject) of Tab 4
• Since we are comparing a DNA sequence on top to an amino acid sequence on the bottom, we need to run blastx
• Open the “Algorithm parameters” section:
  o Turn off the low complexity filter and compositional adjustments; leave other values at default
• Click on the “BLAST” button to run the comparison

Investigate Exons

• Sometimes the blastx search will not find any significant matches
• If so, change the Expect threshold to 1000 or larger and run the blastx search again
• Keep raising the Expect threshold until you get hits, then evaluate hits by position
• This will show very weak similarity which can be better than nothing

Investigate Exons

• We have a weak alignment (50 identities and 94 similarities), but we have seen worse when comparing single exons from these two species
• Notice the location of the hit (16866-17504) and reading frame (+3)
• Note that the first 92aa of the CDS are missing

Investigate Exons

• A similar search with the mav exon 2 sequence gives a location of chr10:18476-19747 and frame +2 with all but the first amino acid aligning
• For larger genes continue with each exon, searching with blastx for similarity (adjusting the Expect threshold if needed) noting location, frame and any missing amino acids
• Very small exons may be undetectable, move on and come back later; find all areas of conservation before making the final model
• Draw these out if this will help
  o Note unaligned amino acids as well

Evidence for mav Annotation

Start Codon
Evidence for mav Annotation

Create a Gene Model

• Pick ATG (met) at start of gene, first met in frame with coding region of similarity (+3)

• For each putative intron/exon boundary compare location of blastx result to locate exact first and last base of the exon such that the conserved amino acids are linked together in a single open reading frame

Create a Gene Model

• For many genes, the locations of donor and acceptor sites will be easily identified based on the locations and quality of the alignments of the individual exons and how these regions compare with evidence of expression from RNA-Seq.

• However, when amino acid conservation is absent, other evidence must be considered. See the handout "Annotation Instruction Sheet" for more help.

Confirm Gene Model

• We will do this later once you have a real model of your own to check

• As a final check, do all three:
  1. Enter coordinates into Gene Model Checker to confirm it is a valid model
  2. Use custom tracks (magnifying glass) to view model and double check that the final model agrees with all your evidence
  3. Examine dot plot to discover possible errors