Annotation of *Drosophila*

Wilson Leung and Chris Shaffer

Last Updated: 12/26/2021

**Agenda**

- Overview of the GEP annotation project
- GEP annotation strategy
  - Types of evidence
  - Analysis tools
  - Web databases
- Annotation of a single isoform (walkthrough)

**Annotation:** Adding labels to a sequence

- **Genes:** Novel or known genes, pseudogenes
- **Regulatory Elements:** Promoters, enhancers, silencers
- **Non-coding RNA:** tRNAs, miRNAs, siRNAs, snoRNAs
- **Repeats:** Transposable elements, simple repeats
- **Structural:** Origins of replication
- **Experimental Results:**
  - DNase I Hypersensitive sites
  - ChIP-chip and ChIP-Seq datasets (e.g., modENCODE)

**Muller element nomenclature**

<table>
<thead>
<tr>
<th>Species</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. simulans</em></td>
<td>X</td>
<td>2L</td>
<td>2R</td>
<td>3L</td>
<td>3R</td>
<td>4</td>
</tr>
<tr>
<td><em>D. sechellia</em></td>
<td>X</td>
<td>2L</td>
<td>2R</td>
<td>3L</td>
<td>3R</td>
<td>4</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>X</td>
<td>2L</td>
<td>2R</td>
<td>3L</td>
<td>3R</td>
<td>4</td>
</tr>
<tr>
<td><em>D. yakuba</em></td>
<td>X</td>
<td>2L</td>
<td>2R</td>
<td>3L</td>
<td>3R</td>
<td>4</td>
</tr>
<tr>
<td><em>D. erecta</em></td>
<td>X</td>
<td>2L</td>
<td>2R</td>
<td>3L</td>
<td>3R</td>
<td>4</td>
</tr>
<tr>
<td><em>D. ananassae</em></td>
<td>X</td>
<td>3R</td>
<td>3L</td>
<td>2R</td>
<td>2L</td>
<td>4/R</td>
</tr>
<tr>
<td><em>D. pseudoobscura</em></td>
<td>X</td>
<td>4</td>
<td>3</td>
<td>2R</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><em>D. persimilis</em></td>
<td>X</td>
<td>4</td>
<td>3</td>
<td>2R</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><em>D. willistoni</em></td>
<td>X</td>
<td>2R</td>
<td>2L</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><em>D. mojavensis</em></td>
<td>X</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>D. vitripennis</em></td>
<td>X</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><em>D. grimshawi</em></td>
<td>X</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
Nomenclature for *Drosophila* genes

- Drosophila gene names are **case-sensitive**
  - Lowercase initial letter = recessive mutant phenotype
  - Uppercase initial letter = dominant mutant phenotype
- Every *D. melanogaster* gene has an annotation symbol
  - Begins with the prefix CG (Computed Gene)
- Some genes have a different **gene symbol** (e.g., *mav*)
- Suffix after the gene symbol denotes different isoforms
  - mRNA = -R; protein = -P
  - mav-RA = Transcript for the A isoform of mav
  - mav-PA = Protein product for the A isoform of mav

GEP annotation strategy

- Technique optimized for projects with a **moderately close, well annotated** neighbor species
  - Example: *D. melanogaster*
- Need to apply different strategies when annotating genes in other species:
  - Example: corn, parrot, parasitoid wasps

GEP annotation goals

- Identify and annotate all the genes in your project
  - For each gene, identify and precisely map (accurate to the base pair) all coding exons (CDS)
  - Do this for ALL isoforms
  - Annotate the initial transcribed exon and transcription start site (TSS)
- Optional curriculum not submitted to GEP
  - Clustal analysis (proteins, promoter regions)
  - Non-coding genes

Evidence-based annotation

- Human-curated analysis
  - Much higher accuracy than standard ab initio and evidence-based gene finders
- **Goal**: collect, analyze, and synthesize all the available evidence to create the best-supported gene model
  - Example: 4591-4688, 5157-5490, 5747-6001

Collect, analyze, and synthesize

- **Collect**:
  - Genome Browser
  - Conservation (BLAST searches)
- **Analyze**:
  - Interpreting Genome Browser evidence tracks
  - Interpreting BLAST results
- **Synthesize**:
  - Construct the best-supported gene model based on potentially contradictory evidence

Evidence for gene models (in general order of importance)

1. Conservation
   - Sequence similarity to genes in *D. melanogaster*
   - Sequence similarity to other *Drosophila* species (ROAST)
2. Expression data
   - RNA-Seq, EST, cDNA
3. Computational predictions
   - Open reading frames (ORFs), gene and splice site predictions
4. Tie-breakers of last resort
   - See the "Annotation Instruction Sheet"
Expression data: RNA-Seq
- Positive results are very helpful
- Negative results are less informative
  - Lack of transcription ≠ no gene
- Evidence tracks:
  - RNA-Seq coverage (read depth)
  - Splice junction predictions
  - Assembled transcripts (Cufflinks, Oases)
- GEP curriculum:
  - RNA-Seq Primer (PowerPoint presentation)
  - Browser-Based Annotation and RNA-Seq Data

Basic annotation workflow
1. Identify the likely ortholog in *D. melanogaster*
2. Determine the gene structure of the ortholog
3. Map each CDS of ortholog to the project sequence
   - Use BLASTX to identify conserved region
   - Note position and reading frame
4. Use these data to construct a gene model
   - Identify the exact start and stop base position for each CDS
5. Use the Gene Model Checker to verify the gene model
6. For each additional isoform, repeat steps 2-5

Gene structure - terminology
- mRNA
- Protein
- Gene span
- Exons
- UTR's
- CDS's

Annotation workflow (graphically)
- BLASTX search of each *D. melanogaster* CDS against the contig
- BLASTP search of feature against the *D. melanogaster* proteins database
- BLAT search of each *D. melanogaster* CDS against the contig
- Use the Gene Model Checker to verify the final CDS coordinates

UCSC Genome Browser
- Provide a graphical view of genomic regions
  - Sequence conservation
  - Gene and splice site predictions
  - RNA-Seq data and splice junction predictions
- BLAT – BLAST-Like Alignment Tool
  - Map protein or nucleotide sequence against an assembly
  - Faster but less sensitive than BLAST
- Table Browser
  - Access raw data used to create the graphical browser
UCSC Genome Browser overview

Genomic sequence

BLASTX alignments
Gene predictions
RNA-Seq
Repeats
Comparative genomics

Evidence tracks

Two different versions of the UCSC Genome Browser

Official UCSC Version
https://genome.ucsc.edu/
Published data, lots of species, whole genomes; used for species with UCSC Assembly Hubs (e.g., parasitoid wasps)

GEP Version
https://gander.wustl.edu
GEP projects, Drosophila genome assemblies; used for Drosophila annotations

Additional resources for the UCSC Genome Browser

- Training section on the UCSC website
  - https://genome.ucsc.edu/training.html
  - Video tutorials
  - User guides
  - Mailing lists

- Biostars
  - https://www.biostars.org/t/ucsc/
  - Questions with the “ucsc” tag

Four websites used by the GEP annotation strategy

Open 4 tabs on your web browser:

1. GEP UCSC Genome Browser (https://gander.wustl.edu)
   - Genome Browser
   - D. virilis – Mar. 2005 – chr10

2. FlyBase (https://flybase.org/)
   - Tools ➔ Genomics Tools ➔ BLAST

3. Gene Record Finder (https://gander.wustl.edu/~wilson/dmelgenerecord/)
   - GEP website ➔ Projects ➔ F Elements
   - Information on the D. melanogaster gene structure

   - BLASTX ➔ select the checkbox: Align two or more sequences

Navigate to the Genome Browser for the project region

- Configure the Genome Browser Gateway page:
  - Browse/Select Species: D. virilis
  - D. virilis Assembly: Mar. 2005 (GEP/Annot. D. virilis ppt)
  - Position/Search Term: chr10

- Click “Go”
Control how evidence tracks are displayed on the Genome Browser

- Five different display modes:
  - **Hide**: track is hidden
  - **Dense**: all features appear on a single line
  - **Squish**: overlapping features appear on separate lines
    - Features are half the height compared to full mode
  - **Pack**: overlapping features appear on separate lines
    - Features are the same height as full mode
  - **Full**: each feature is displayed on its own line
    - Set “Base Position” track to “Full” to see the amino acid translations

- Some evidence tracks (e.g., RepeatMasker) only have a subset of these display modes

Initial assessment of fosmid project chr10

- Seven gene predictions (features) from GenScan
  - Need to investigate each feature if one were to annotate this entire project

 Investigate gene prediction chr10.4

- Enter chr10:15000-21000 in the position box and click go to navigate to this region
- Click on the feature and select “Predicted Protein” to retrieve the predicted protein sequence
  - Select and copy the sequence

Computational evidence

- Assumption: there are recognizable signals in the DNA sequence that the cell uses; it should be possible to detect these signals computationally

- Many programs designed to detect these signals
  - Use machine learning and Bayesian statistics

- These programs do work to a certain extent
  - The information they provide is better than nothing
  - The predictions have high error rates

Accuracy of the different types of computational evidence

- DNA sequence analysis:
  - ab initio gene predictors – exons vs. genes
  - Splice site predictors – high sensitivity but low specificity

- Multi-species alignment – low specificity

- BLASTX protein alignment:
  - WARNING! Students often over-interpret the BLASTX alignment track (D. mel Proteins); use with caution

Data on the Genome Browser is incomplete

- Most evidence has already been gathered for you and they are shown in the Genome Browser tracks
- Annotator still needs to generate conservation data
  1. Assign orthology
  2. Look for regions of conservation using BLAST
    - Collect the locations (position, frame, and strand) of conservation to the orthologous protein
    - If no expression data are available, these searches might need to be exhaustive (labor intensive)
Detect sequence similarity with BLAST

- **BLAST** = Basic Local Alignment Search Tool

- Why is BLAST popular?
  - Provide statistical significance for each match
  - Good balance between sensitivity and speed

- Identify local regions of similarity

Common BLAST programs

- Except for BLASTN, all alignments are based on comparisons of protein sequences
  - Alignment coordinates are relative to the original sequences

- Decide which BLAST program to use based on the type of query and subject sequences:

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database (Subject)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Nucleotide → Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Protein</td>
<td>Nucleotide → Protein</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Nucleotide → Protein</td>
<td>Nucleotide → Protein</td>
</tr>
</tbody>
</table>

Where can I run BLAST?

- **NCBI BLAST** web service

- **EBI BLAST** web service
  - [https://www.ebi.ac.uk/Tools/sss/](https://www.ebi.ac.uk/Tools/sss/)

- **FlyBase BLAST (D. melanogaster)**
  - [https://flybase.org/blast/](https://flybase.org/blast/)

Accessing TBLASTX at NCBI

Detect conserved *D. melanogaster* coding exons with BLASTX

- Coding sequences evolve slowly
- Exon structure changes very slowly
- Use sequence similarity to infer homology
  - *D. melanogaster* very well annotated
  - Use BLAST to find similarity at amino acid level
- As evidence accumulates indicating the presence of a gene, we could justify spending more time and effort looking for conservation

FlyBase – Database for the *Drosophila* research community

- Key features:
  - Lots of ancillary data for each gene in *D. melanogaster*
  - Curation of literature for each gene
  - Reference *D. melanogaster* annotations for other databases (including NCBI, EBI, and DDBJ)
  - Fast release cycle (~6-8 releases per year)
Web databases and tools

- Many genome databases available
  - Be aware of different annotation releases
    - e.g., release 6 versus release 5 assemblies
  - Use FlyBase as the canonical reference

- Web databases are being updated constantly
  - Update GEP materials before semester starts
  - Potential discrepancies in exercise screenshots
  - Minor changes in search results
  - Let us know about errors or revisions

Finding the ortholog

BLASTP search of the chr10.4 gene prediction against the set of D. melanogaster proteins at FlyBase

BLASTP results show a significant hit to the mav gene

<table>
<thead>
<tr>
<th>Description</th>
<th>Species</th>
<th>Strain</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mav-PA</td>
<td>Drosophila</td>
<td>A</td>
<td>5.80E-43</td>
</tr>
<tr>
<td>gbb-PA</td>
<td>Drosophila</td>
<td>B</td>
<td>1.79E-15</td>
</tr>
</tbody>
</table>

Note the large change in E-value from mav-PA to the next best hit gbb-PB; good evidence for orthology

FlyBase naming convention:
- <gene symbol>-[RP]<isoform>
  - R = mRNA, P = Protein
- mav-PA = Protein product from the A isoform of the gene mav

Results of ortholog search

- Degree of similarity consistent with comparison of typical proteins from D. virilis versus D. melanogaster
  - Regions of strong similarity interspersed with regions of little or no similarity
  - Same Muller element supports orthology
  - We have a probable ortholog: mav

Constructing a gene model

- We need more information on mav:
  - Determine the gene function and structure in D. melanogaster

- If this is the ortholog, we also need the amino acid sequence of each CDS

- Use two websites to learn about the gene structure:
  - FlyBase
    - Graphical representation of the gene; detailed information on each gene
  - Gene Record Finder
    - Table list of isoforms, retrieve exon sequences

Obtain the gene structure model

Using the Gene Record Finder and FlyBase to determine the gene structure of the D. melanogaster gene mav
Retrieve the Gene Record Finder record for mav

1. Type mav into the search box, then press [Enter]

2. Click on the “View in GBrowse” link to get a graphical view of the gene structure on FlyBase

Gene structure of mav

1. Both the graphical and table view shows mav has a single isoform (mav-RA) in D. melanogaster
   - This isoform has two transcribed exons
   - Both exons contain translated regions

Investigate exons

1. Search each CDS from D. melanogaster against the project sequence (e.g., fosmid / contig)
   - Identify regions within the project sequence that code for amino acids similar to the D. melanogaster CDS

2. Best to search with the entire project DNA sequence
   - Easier to keep track of positions and reading frames

Retrieve CDS sequences

1. In the Polypeptide Details section of the Gene Record Finder, select a row in the CDS table
2. The corresponding CDS sequence will appear in the Sequence viewer window

Retrieve the fosmid sequence

1. Go back to the Genome Browser (first tab) and navigate to chr10
2. Click on the DNA link under the “View” menu, then click on the “get DNA” button
Compare the CDS against the fosmid sequence with BLASTX
- Copy and paste the genomic sequence from tab 1 into the “Enter Query Sequence” textbox
- Copy and paste the sequence for the CDS 1_9489_0 from tab 2 into the “Enter Subject Sequence” textbox
- Expand the “Algorithm parameters” section:
  - Verify the Word size is set to 3
  - Turn off compositional adjustments
  - Turn off the low complexity filter
- Click “BLAST”

Adjust the Expect threshold
- E-value is negatively correlated with alignment length
  - Difficult to find small exons with BLAST
- Sometimes BLAST cannot find regions with significant similarity:
  - Use the Query subrange field to restrict the search region
  - Change the Expect threshold to 1000 and try again
    - Keep increasing the Expect threshold until you get hits
  - This strategy will identify regions with very weak similarity, but they can be better than nothing

Examine the BLASTX alignment
- BLASTX result shows a weak alignment
  - 50 identities and 94 similarities (positives)
  - Low degree of similarity is not unusual when comparing single exons from these two species
  - Location: 16866-17504, frame: +3, missing first 92 aa

Map the second CDS (2_9489_0)
- Perform similar BLASTX search with CDS 2_9489_0
  - Location: chr10:18,476-19,747, Frame: +2
  - Alignment includes the entire CDS except for first amino acid
- For larger genes, continue mapping each exon with BLASTX (adjust Expect threshold as needed)
  - Record location, frame, and missing amino acids
  - BLAST might not find very small exons
    - Move on and come back later
  - Plotting the location of each CDS might help
    - Note the parts of the CDS missing from the alignments

Identify missing parts of CDS
- BLASTX CDS alignments for mav on D. virilis
  - Location ranges:
    - Frame +3: 16,866-17,504
    - Frame +2: 18,476-19,747

Building a gene model
Determine the exact start and end positions of each exon for the putative mav-PA ortholog
Basic biological constraints (inviolate rules*)

- Coding regions start with a methionine
- Coding regions end with a stop codon
- Gene should be on only one strand of DNA
- Exons appear in order along the DNA (collinear)
- Intron sequences should be at least 40 bp
- Intron starts with a GT (or rarely GC)
- Intron ends with an AG

* There are known exceptions to each rule

---

Find the missing start codon

- Only a single start codon (16,515-16,517 in frame +3) upstream of conserved region before the stop codon
  - Start codon at 16,977-16,979 would truncate conserved region
- CDS alignment starts at 16,866 but 92 amino acids missing
  - 16,866 - (92 * 3) = 16,590

---

A genomic sequence has 6 different reading frames

- Frame: Base to begin translation relative to the first base of the sequence

---

Splice donor and acceptor phases

- Phase: Number of bases between the complete codon and the splice site
  - Donor phase: Number of bases between the end of the last complete codon and the splice donor site (GT/GC)
  - Acceptor phase: Number of bases between the splice acceptor site (AG) and the start of the first complete codon
- Phase depends on the reading frame of the CDS

---

Phase depends on the reading frame

- Extra nucleotides from donor and acceptor phases will form an additional codon
  - Donor phase + acceptor phase = 0 or 3

---

Phase of donor and acceptor sites must be compatible

- CCA AAT GT ... AG CTC GAT
  - Translation: P N V L D

---
Incompatible donor and acceptor phases results in a frame shift

Phase 0 donor is incompatible with phase 2 acceptor; use prior GT, which is a phase 1 donor.

Picking the best acceptor site

Reading frame (+2) dictated by BLASTX result

Two splice acceptor candidates:
- 18,471-18,472 (phase 0)
- 18,482-18,483 (phase 1), remove conserved amino acids
18,471-18,472 is the better candidate
Pick different candidate if no phase 0 donor in previous exon

Picking the best donor site

Reading frame (+3) dictated by BLASTX result

Two splice donor candidates:
- 17,505-17,506 (phase 0)
- 17,518-17,519 (phase 1)
Phase 0 donor (17,505-17,506) is compatible with phase 0 acceptor (18,471-18,472)

Create the final gene model

Pick ATG (M) at the start of the gene
For each putative intron/exon boundary, use the Genome Browser to locate the exact first and last base of the exon
- After splicing, conserved amino acids are linked together in a single long open reading frame
- CDS coordinates for mav: 16515-17504, 18473-19744

Additional strategies for identifying splice donor and acceptor sites

For many genes, the location of donor and acceptor sites can be identified easily based on:
- Locations and quality of the CDS alignments
- Evidence of expression from RNA-Seq
  - Splice junction predictions

When amino acid conservation is absent, other evidence and techniques must be considered. See the following documents for help:
- Annotation Instruction Sheet
- Annotation Strategy Guide

Verify the final gene model using the Gene Model Checker

Examine the checklist and explain any errors or warnings in the GEP Annotation Report
View your gene model in the context of the other evidence tracks on the Genome Browser
Examine the dot plot and explain any discrepancies in the GEP Annotation Report
- Look for large vertical and horizontal gaps
- See the "Quick Check of Student Annotations" document on the GEP website