Annotation of *Drosophila* Primer

August 2019

Wilson Leung and Chris Shaffer

Outline

- Overview of the GEP annotation projects
- GEP annotation workflow
- Practice applying the GEP annotation strategy

GEP *Drosophila* Annotation Projects

- Intronant Species
- GEP Publications
- Motif Project
- Expanded F Project

The Pathways Project analyze genes from 27 *Drosophila* species

Tree scale: 0.1

Muller element nomenclature

Gene structure nomenclature

GEP annotation goals

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

Evidence for gene models
(in general order of importance)

1. Conservation
   - Sequence similarity to genes in *D. melanogaster*
   - Sequence similarity to other *Drosophila* species (Multiz)
2. Expression data
   - RNA-Seq, EST, cDNA
3. Computational predictions
   - Open reading frames; gene and splice site predictions
4. Tie-breakers of last resort
   - See the “Annotation Instruction Sheet”

Basic annotation workflow

1. Identify the likely *D. melanogaster* ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. Verify the model using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

Annotation workflows available under the “Introducing Genes” section of the GEP web site

Four main web sites used by the GEP annotation strategy

1. GEP UCSC Genome Browser (http://gander.wustl.edu)
2. FlyBase (http://flybase.org)
   - Tools ➔ Genomic Tools ➔ BLAST
   - Jump to Gene ➔ Genomic Location ➔ GBrowse
3. Gene Record Finder (http://gep.wustl.edu)
   - Projects ➔ Annotation Resources
   - BLASTX ➔ select the checkbox: Align two or more sequences

Annotation workflow: Step 1

1. Identify the likely *D. melanogaster* ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. Verify the model using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

Two different versions of the UCSC Genome Browser

- Official UCSC Version
  - http://genome.ucsc.edu
  - Published data; lots of species, whole genomes; used for “Chimp Chunks”
- GEP Version
  - http://gander.wustl.edu
  - GEP projects, *Drosophila* genome assemblies; used for *Drosophila* annotations
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

GEP annotation strategy

- Use *D. melanogaster* as reference
- *D. melanogaster* is very well annotated
- Use sequence similarity to infer homology

- Minimize changes compared to the *D. melanogaster* gene model (parsimony)
  - Coding sequences evolve slowly
  - Exon structure changes very slowly

Overview of NCBI BLAST

- Detect local regions of significant sequence similarity between two 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/
- FlyBase BLAST (Drosophila and other insects)  
  - http://flybase.org/blast/

Accessing TBLASTX at NCBI

Annotation workflow: Step 2

1. Identify the likely D. melanogaster ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. Verify the model using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

Gene Record Finder – Observe the structure of D. melanogaster genes

- Retrieves CDS and exon sequences for each gene in D. melanogaster
- CDS and exon usage maps for each isoform
- List of unique CDS
- Designed for the exon-by-exon annotation strategy

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., ey)
- Suffix after the gene symbol denotes different isoforms
  - mRNA = -R, protein = -P
  - ey-RA = Transcript for the A isoform of ey
  - ey-PA = Protein product for the A isoform of ey
Be aware of different annotation releases

- *D. melanogaster* Release 6 genome assembly
  - First change of the assembly since late 2006
  - Most modENCODE analysis used the Release 5 assembly
- Gene annotations change much more frequently
  - Use FlyBase as the canonical reference
- GEP data freeze
  - GEP materials are updated before the start of semester
- Potential discrepancies in results and screenshots
  - See the archived BLAST results in the exercise package
  - Let us know about major errors or discrepancies

**DEMO**: Determine the gene structure of the *D. melanogaster* gene CG31997

---

**Annotation workflow: Step 3**

1. Identify the likely *D. melanogaster* ortholog
2. Observe the gene structure of the ortholog
3. **Map each CDS** to the project sequence
4. Determine the exact coordinates of each CDS
5. Verify the model using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

**BLAST** parameters for CDS mapping

- Select the “Align two or more sequences” checkbox
- Settings in the “Algorithm parameters” section
  - Verify the Word size is set to 3
  - Turn off compositional adjustments
  - Turn off the low complexity filter

---

**Strategies for CDS mapping**

- Start by mapping the **largest** CDS
  - The first and last CDS tend to be smaller than internal CDS in *Drosophila*
  - Continue mapping CDS by size in descending order
- Defer mapping small or weakly conserved CDS
  - Use placements of adjacent CDS to define the search region
  - Use the splice donor and acceptor **phases** of adjacent CDS as additional constraints

---

**Strategies for finding small CDS**

- Examine RNA-Seq coverage and TopHat junctions
  - Small CDS is typically part of a larger transcribed exon
- Use Query subrange to restrict the search region
- Increase the **Expect threshold** and try again
  - Keep increasing the Expect threshold until you get matches
  - Also try decreasing the word size
- Use the **Small Exon Finder**
  - Minimize changes in CDS size
  - Available under Projects ➔ Annotation Resources
- See the “Annotation Instruction Sheet” for details
**EXERCISE:**
Map each CDS to the project sequence

- Use *BLASTX* to determine the approximate locations for the three CDS of *CG31997* on contig10
- Consult with each other
- The “Annotation of a *Drosophila* Gene” document provides a step-by-step walkthrough

---

**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

---

**modENCODE RNA-Seq data**

- RNA-Seq evidence tracks:
  - RNA-Seq coverage (read depth)
  - *TopHat* splice junction predictions
  - Assembled transcripts (*Cufflinks*, *Oases*)
- Positive results very helpful
- Negative results less informative
  - Lack of transcription ≠ no gene
- GEP curriculum:
  - RNA-Seq Primer
  - Browser-Based Annotation and RNA-Seq Data

---

**Discussion and coffee break**

Carolina Ponce: [https://flic.kr/p/otHbqV](https://flic.kr/p/otHbqV)

---

**Annotation workflow: Step 4**

1. Identify the likely *D. melanogaster* ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the **exact coordinates** of each CDS
5. Verify the model using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

---

**DEMO:** Map CDS 3_10767_1 of *CG31997* against contig10 with *BLASTX*
Overview of RNA-Seq (Illumina)

5’ cap
Processed mRNA
RNA fragments (~250bp)
Library with adapters
5’ 3’ 5’ 3’ 5’ 3’
Paired end sequencing
RNA-Seq reads

DEMO: Use RNA-Seq coverage to support the placement of the start codon

Can use the TopHat splice junction predictions to identify splice sites

EXERCISE: Confirm the placement of the stop codon for CDS 3_10767_1

A genomic sequence has 6 different reading frames

Frame: Base to begin translation relative to the start of the sequence

A codon could be derived from nucleotides in adjacent exons
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 of the donor and acceptor sites must be compatible

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

Incompatible donor and acceptor phases result in a frame shift

DEMO: Use RNA-Seq to annotate the intron between CDS 1_10767_0 and 2_10767_2 of the *CG31997* ortholog

EXERCISE: Determine the coordinates for CDS 2_10767_2 and 3_10767_1 of the *CG31997* ortholog
Annotation workflow: Step 5

1. Identify the likely D. melanogaster ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. **Verify the model** using the Gene Model Checker
6. **Repeat steps 2-5** for each additional isoform

Verify the final gene model using the Gene Model Checker

- Gene model should satisfy biological constraints
- **Explain errors or warnings** in the GEP Annotation Report
- Compare model against the D. melanogaster ortholog
- Dot plot and protein alignment
- See “Quick check of student annotations”
- View your gene model as a custom track in the genome browser
- Generate files required for project submission

Annotation workflow: Step 6

1. Identify the likely D. melanogaster ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. **Verify the model** using the Gene Model Checker
6. **Repeat** steps 2-5 for each additional isoform

DEMO: Verify the proposed gene model for the ortholog of CG31997

Next step: practice annotation

*D. biarmipes* Aug. 2013 (GEP/Dot) assembly

- Annotation of a *Drosophila* Gene
- onecut on contig35
- cy on contig40
- CG1909 on contig35
- Arl4 and CG33978 on contig10

Questions?