

Annotation of *Drosophila* Primer

December 2023

Wilson Leung and Chris Shaffer

1

Outline

- Overview of the GEP annotation projects
- Annotation workflow for the F Element Project
- Practice applying the annotation strategy

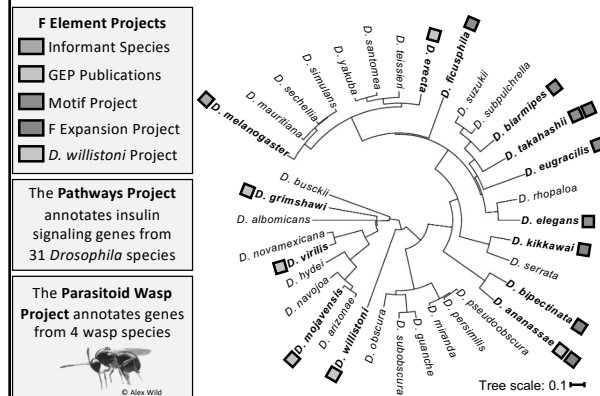
2

AAACAACAAATCATAAATAGAGGAAGTTTTCGGAATATACGATAAGTGAAATATCGTTCT
TAAAAAGAGCAAGACAGTTTAAACATTGAAAAAAGATTATCCAAATAGCCGTAAAG
GTTCAATTAATGACAATGACCATGGCGGCAAGTCGATGAAGGACTAGTCGGAACCTGGA
AATAGGAATGCGCCAAAGCTAGTCAGCTAAACATCAATTGAAACAAGTTTGTACATC
GATGCGCGGAGGCGCTTTTCTCTCAGATGGCTGGGGATGCCAGCAGTTAATCAGGAT
ACCAATTGAGGAGTGGCCAGCTCAGCTAGAGCGGCAATAGGACCCATCGGGGG
GCCGTTATGTGAAGCAACATTAAACATAGGCAACCGATTGTGGGAATCGAATT
TAACAACGGCGGTGAGCAACCGCTCAACAGTGCAGAGCCATCTGGGGGCAATAG
CCTTCATCAAAATTTGGCGGAACTTGGGGCGAGGACGATGATGGCGCCGATAGCACCGAG
CGTTTGGACGGGTGAGTATCCACATATGCACAACTCTGGTTGCACTGGGTGGCA
TAGCGCTGGCGGTGGCGCGCTGCTGGTCCCTAATGGGGACAGGCTGTTGCTGTTGG
TGTGGAGTCGGAGTTGCTTAACTCGACTGGAATAACAATGCGCCGGCAACAGGAG
CCCTGCTGCGGTGGCTCGTCCGAAATGTGGGGACATCATCTCAGATTGCTCACAATC
ATCGCCGGAATGNTAANGAATTAATCAAAATTTGGCGGACATAATGNGCAGATTGAGA
ACGTATTAAACAAATGGTGGCGCCGTTGTTAGTGCAACAGGTCAAATATCGCAAGOT
CAATATTGGCCCAAGCGGTGTTGGTTCCGTATCCGTAATGTCGGGGCACAAATGGGGA
GCCACACAGCCCGCTGGGGGCCCCAAGGATTTTCAAGCAAACTACTGGATGGGAGGA
ACCAACATCAGATTAGAAATTAACAAATGGTGGCGCCGCTGTTATGGATAAAAAA
TTTGTGCTTCTGACGGAGATTATGTTGTAATCAATTTTAAAGATTTTAAATAAA
TATGTGACCTTTACGAGAAATTTGCTTACCTTTTCGACACACACACTTATACAGACA
GGTAATAATACCTTTGAGCAATTCGATTTTCATAAAATATACCTAAATCGCATCGTC

Start codon Coding region Stop codon
Splice donor Splice acceptor UTR

3

GEP Annotation Projects



4

Annotation Projects Under Development

Puerto Rican Parrot

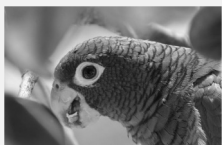


Photo by Tom Mackenzie

- Critically endangered species found only in Puerto Rico
- Reduced eggshell strength leads to loss of offspring
- Annotate genes that are differentially expressed in the uterus during eggshell deposition

Detoxification Genes



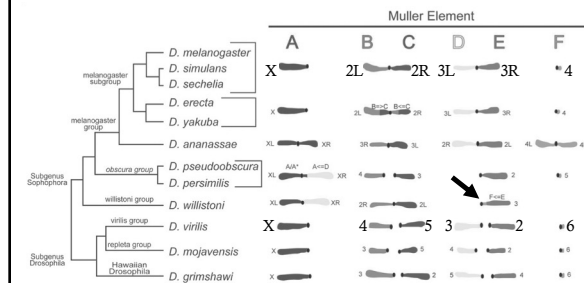
Scott Chialvo et al., 2019

Photo by Darren Obbard

- Evolution of toxin tolerance in mushroom-feeding *Drosophila*
- Focus on species from the immigrans – tripunctata radiation
- Annotate genes from six gene families that contribute to xenobiotic detoxification

5

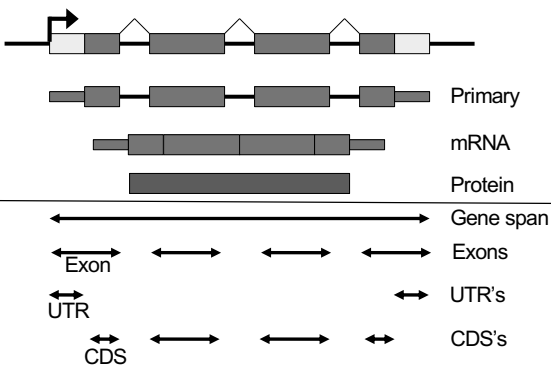
Muller element nomenclature



Schaeffer SW et al., 2008. Polytene Chromosomal Maps of 11 *Drosophila* Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. *Genetics*. 2008 Jul;179(3):1601-55

6

Gene structure nomenclature



7

Annotation goals for the F Element Project

- ⊛ 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
- ⊛ Optional analyses not submitted to GEP
 - ⊛ Annotate the initial transcribed exon and transcription start site (TSS)
 - ⊛ Clustal analysis (proteins, promoter regions)
 - ⊛ Non-coding genes

8

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"

9

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 in the "Curriculum" section of the GEP website

10

Four websites used by the annotation strategy for the F Element Project

1. GEP UCSC Genome Browser (thegep.org/browser)
2. FlyBase (<https://flybase.org/>)
 - ⊛ Tools → Genomics Tools → BLAST
 - ⊛ Jump to Gene → Genomic Location → JBrowse
3. Gene Record Finder (thegep.org/finder)
 - ⊛ Under "Resources & Tools"
4. NCBI BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)
 - ⊛ *blastx* → select the checkbox: ☒ Align two or more sequences

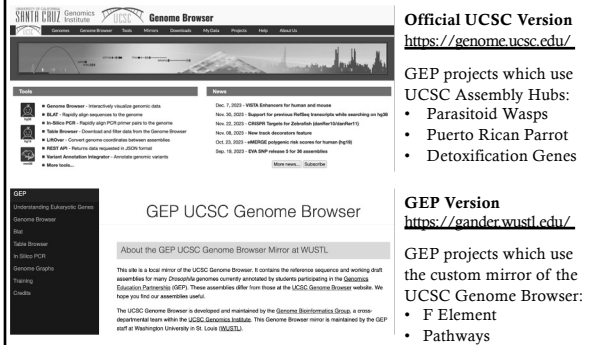
11

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

12

Two different versions of the UCSC Genome Browser



Official UCSC Version
<https://genome.ucsc.edu/>

GEP projects which use UCSC Assembly Hubs:

- Parasitoid Wasps
- Puerto Rican Parrot
- Detoxification Genes

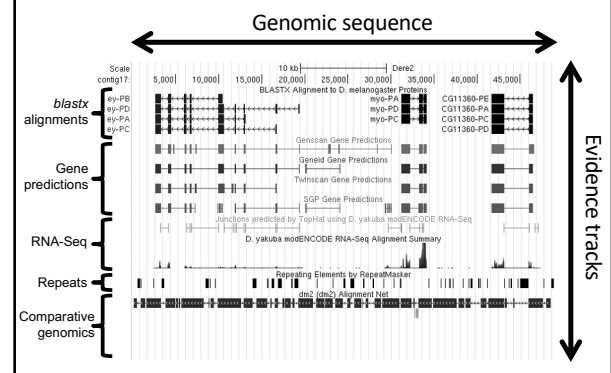
GEP Version
<https://gander.wustl.edu/>

GEP projects which use the custom mirror of the UCSC Genome Browser:

- F Element
- Pathways

13

UCSC Genome Browser overview

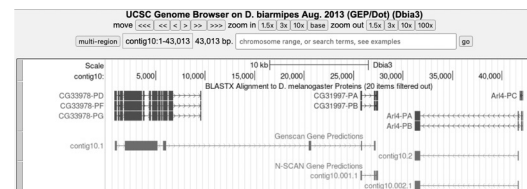


14

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

15



DEMO: Examine contig10 in the *D. biarmipes* Aug. 2013 (GEP/Dot) assembly with the GEP UCSC Genome Browser

16

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

17

FlyBase – Database for the *Drosophila* research community



- Lots of ancillary data for each gene in *D. melanogaster*
- Curation of literature for each gene
- Reference for *D. melanogaster* annotations for all other databases
 - Including NCBI, EBI, and DDBJ
- Fast release cycle (6-8 releases per year)

18

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:

Program	Query	Database (Subject)
<i>blastn</i>	Nucleotide	Nucleotide
<i>blastp</i>	Protein	Protein
<i>blastx</i>	Nucleotide → Protein	Protein
<i>tblastn</i>	Protein	Nucleotide → Protein
<i>tblastx</i>	Nucleotide → Protein	Nucleotide → Protein

19

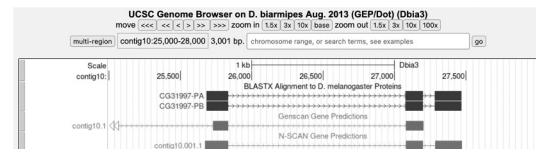
Where can I run BLAST?

- ☛ NCBI BLAST web service
 - ☛ <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- ☛ EBI BLAST web service
 - ☛ <https://www.ebi.ac.uk/jdispatcher/sss/ncbiblast>
- ☛ FlyBase BLAST (*D. melanogaster*)
 - ☛ <http://flybase.org/blast/>

20

Accessing *tblastx* at NCBI

21



DEMO: Ortholog assignment for the N-SCAN prediction contig10.001.1

22

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

23

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

24

Nomenclature for *Drosophila* genes

- Case of the initial letter in *Drosophila* gene symbols:
 - 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 (**C**omputed **G**ene)
- Some genes have a different **gene symbol** (e.g., *ey*)
- Suffix after the gene symbol denotes different isoforms
 - mRNA = **-R**; protein = **-P**
 - ey-R4* = Transcript for the A isoform of *ey*
 - ey-PA* = Protein product for the A isoform of *ey*

25

Changes in nomenclature (FlyBase release FB2022_06)

- To avoid confusion (for non-*Drosophila* researchers), FlyBase gene symbols are no longer case-sensitive in *D. melanogaster* annotation release **6.49**
- D. melanogaster* RefSeq gene records at NCBI are derived from an older annotation release (**6.46**)
 - Gene symbols are still case-sensitive

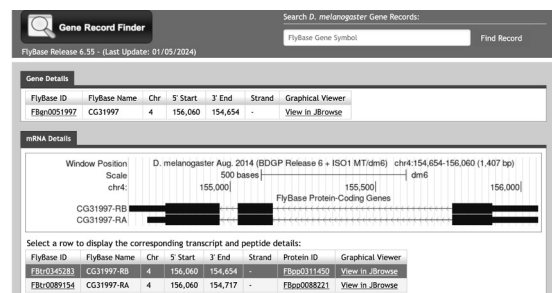
Recommendation: continue to treat gene symbols as case-sensitive until the next NCBI RefSeq update for *D. melanogaster*

26

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

27



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

28

Annotation workflow: Step 3

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

29

BLAST parameters for CDS mapping

- Select the “Align two or more sequences” checkbox
 - ☒ Align two or more sequences
- Settings in the “Algorithm parameters” section
 - Verify the Word size is set to 3
 - Word size: 3
 - Turn off **compositional adjustments**
 - Compositional adjustments: No adjustment
 - Turn off the **low complexity filter**
 - Filter: ☐ Low complexity regions

30

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

31

Strategies for finding small CDS

- ☛ Examine **RNA-Seq** coverage and splice junction predictions
 - ☛ 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 Exons Finder**
 - ☛ Minimize changes in CDS size
 - ☛ Available under “Resources & Tools” on the F Element Project page
- ☛ See the “Annotation Instruction Sheet” for details

32

DEMO: Map CDS 3_10720_1 of *CG31997* against contig10 with *blastx*

33

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

34

Discussions and coffee break



duncan c: <https://flic.kr/p/nSfc14>

35

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

36

Basic biological constraints (involute 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

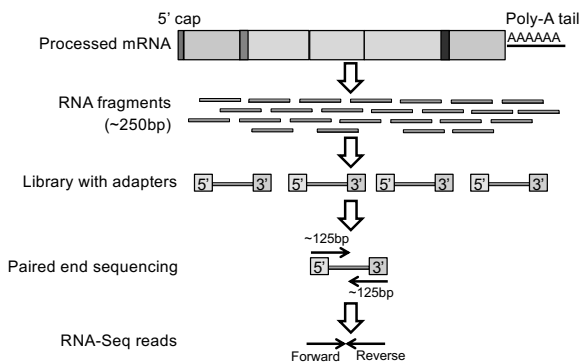
37

Interpreting RNA-Seq data

- ⊗ RNA-Seq evidence tracks:
 - ⊗ RNA-Seq coverage (read depth)
 - ⊗ Splice junction predictions (TopHat, regtools)
 - ⊗ Assembled transcripts (Cufflinks, Oases, StringTie)
- ⊗ 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

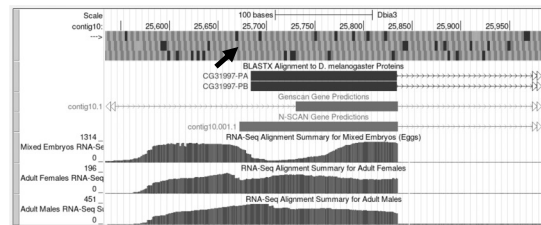
38

Overview of RNA-Seq (Illumina)



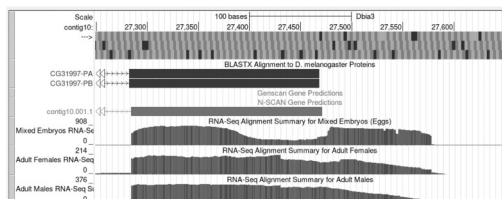
Wang Z *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 10(1):57-63.

39



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

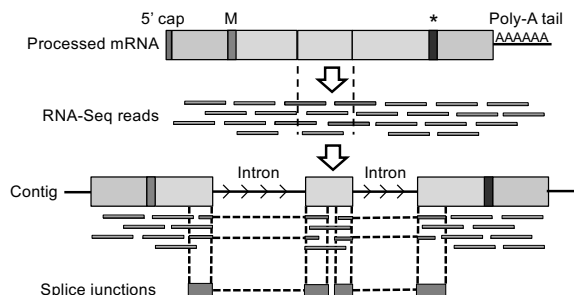
40



EXERCISE: Confirm the placement of the stop codon for CDS 3_10720_1

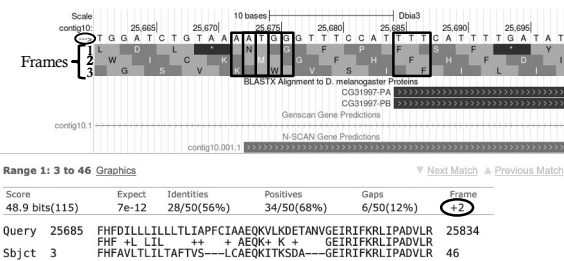
41

Use spliced RNA-Seq reads to identify splice sites



42

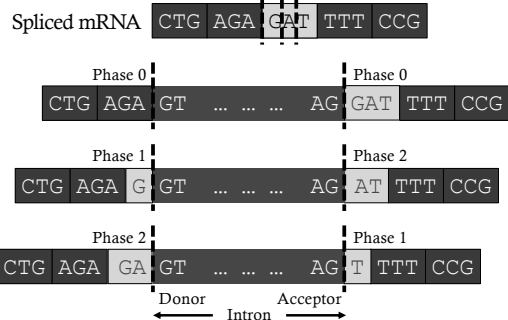
A genomic sequence has 6 different reading frames



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

43

A codon could be derived from nucleotides in adjacent exons



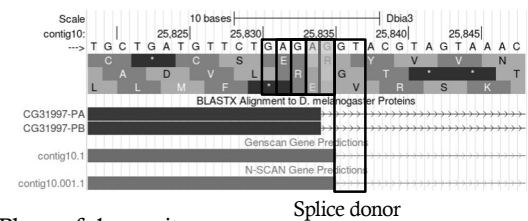
44

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

45

Phase depends on the reading frame



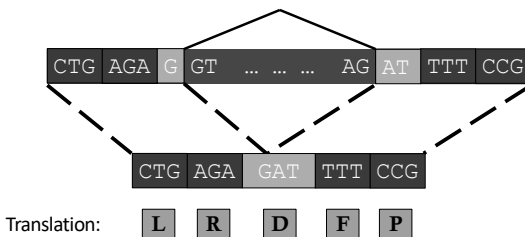
⊛ **Phase of donor site:**

- ⊛ Phase 2 relative to frame +1
- ⊛ Phase 1 relative to frame +2
- ⊛ Phase 0 relative to frame +3

46

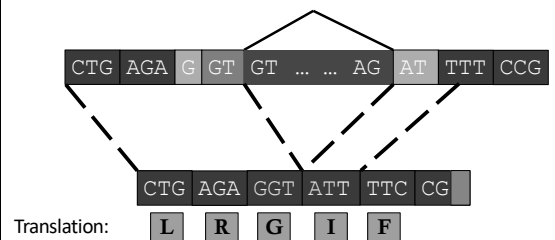
Phase of the donor and acceptor sites must be compatible

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



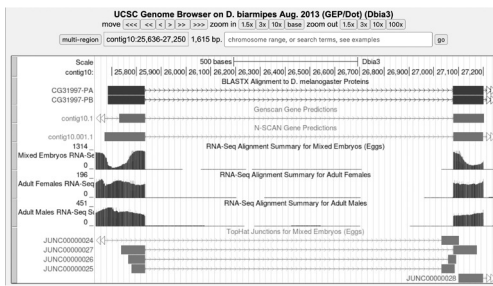
47

Incompatible donor and acceptor phases result in a frame shift



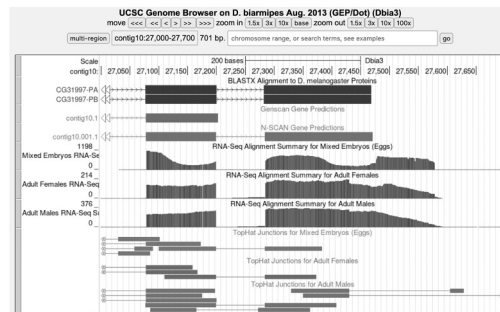
⊛ Phase 0 donor is incompatible with phase 2 acceptor

48



DEMO: Use RNA-Seq to annotate the intron between CDS 1_10720_0 and 2_10720_2 of the *CG31997* ortholog

49



EXERCISE: Determine the coordinates for CDS 2_10720_2 and 3_10720_1 of the *CG31997* ortholog

50

Annotation workflow: Step 5

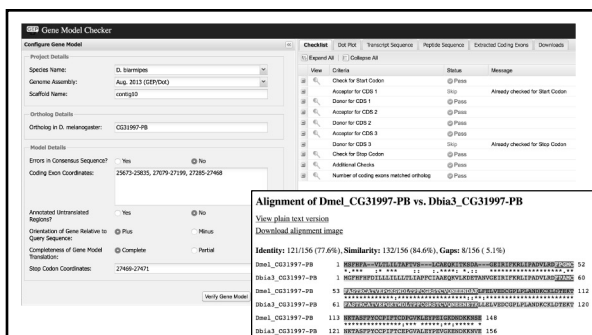
1. Identify the likely *D. melanogaster* ortholog
2. Observe the gene structure of the ortholog
3. Map each CDS to the project sequence
4. Determine the exact coordinates of each CDS
5. **Verify the model** using the Gene Model Checker
6. Repeat steps 2-5 for each additional isoform

51

Verify the final gene model using the Gene Model Checker

- Gene model should satisfy biological constraints
- **Explain errors or warnings** in the Annotation Report for the F Element project
- 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 require for project submission

52



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

53

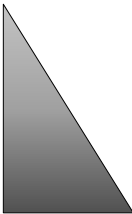
Annotation workflow: Step 6

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

54

Next step: practice annotation

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

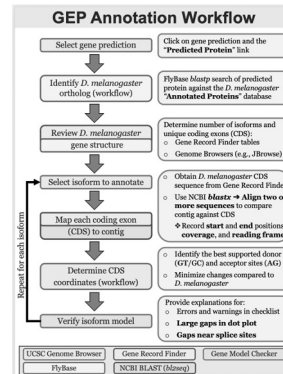


Difficulty

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

55

Questions?



56