


Genomics Education Partnership
thegep.org



P A T H W A Y S

Pathways Project Annotation Primer

Adapted from Wilson Leung
Katie M. Sandlin
Last Update: 12/31/2025

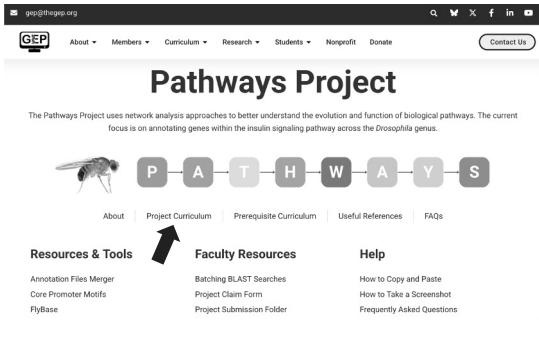
1

Outline

- Overview of comparative genome annotation
- Pathways Project annotation strategy
 - Types of evidence
 - Analysis tools
 - Web databases
- Annotation of a single isoform of *Rheb* in *D. yakuba*

2

Access the Pathways Project Curriculum



gpep@thegep.org

Pathways Project

The Pathways Project uses network analysis approaches to better understand the evolution and function of biological pathways. The current focus is on annotating genes within the insulin signaling pathway across the *Drosophila* genus.

P A T H W A Y S

Resources & Tools: Annotation Files Merger, Core Promoter Motifs, FlyBase

Faculty Resources: Batching BLAST Searches, Project Claim Form, Project Submission Folder

Help: How to Copy and Paste, How to Take a Screenshot, Frequently Asked Questions

3

"Introduction to Pathways Project" Lecture


Project Curriculum

Pilot Project Curriculum | Prerequisite Curriculum

Introduction to Pathways Project Lecture is designed to introduce students to the big picture of the Pathways Project.	Pathways Project: Annotation Walkthrough This walkthrough illustrates how to apply the GEP annotation strategy for the Pathways Project to construct a gene model for the Ras homolog enriched in brain (<i>Rheb</i>) gene in <i>Drosophila yakuba</i> .	Pathways Project: Annotation Workflow The Annotation Workflow is a one page summary of the annotation protocol for the Pathways Project.
Pathways Project: Reference Glossary The Reference Glossary includes definitions for terms that are frequently used in the Pathways Project.	Pathways Project: Annotation Form This "Annotation Form" merged the "Annotation Report" and "Annotation Notebook" into a single document and the latter two items are now archived.	Pathways Project: Annotation Form Exemplar The Annotation Form Exemplar is provided as an example of a completed Annotation Form ready for submission to the GEP's Pathways Project. The optional questions were omitted from the exemplar.

4

Walkthrough illustrating how to construct a gene model for the *Rheb* gene in *Drosophila yakuba*



Pathways Project: Annotation Walkthrough
Katie Sandlin, Wilson Leung, & Laura Reed

Prerequisites

- Understanding Eukaryotic Genes (UEG) Modules 1, 4, 5, & 6
- RNA-Seq Primer
- An Introduction to NCBI BLAST or Sequence Similarity Introduction

Resources & Tools

- All links for this walkthrough, including the recommended prerequisites listed above, can be found on the [Pathways Project](#) page of the GEP website (thegep.org/pathways).
- Annotation Form
- Annotation Form Exemplar
- Annotation Workflow
- Reference Glossary

Available through the "[Pathways Project: Annotation Walkthrough](#)" page on the GEP website

5

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)

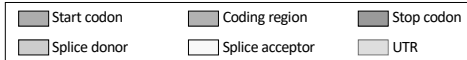
6

What is gene annotation?

```

AAACAACAATCATAAATAGAGGAAGTTTCGGAAATACGATAAGTGAATATCGTTCT
TAAAAAAGAGCAAGAAGCTTTAACCATTGAAAACAAGATTATCCAAATAGCCGTAAGA
GTTTCATTTAATGACAAATGACCAATGGCCGCAAAAGTCGATGAAGGACTAGTCGGAACCTGGA
AATAGGAATGCGCCAAAAGCTAGTGCAGCTAAACATCAATTGAAAACAAGTTGTACATC
GATGCGCGGAGGCGCTTTTCTCTCAGATGGCTGGGGATGCCAGCACGTTAATCAGGAT
ACCAATTGAGGAGCTGGCCACGCTCACCTAGAGCCGGCCAATAAGGACCCATCGGGGGG
GCCGCTTATGTGGAAGCCAAACATTAAACCATAGGCAACCGATTGTGGGAATCGAATT
TAACAAAACGGCGGTGAGCCACCGCTCAACAAGTCCAAAAGCCATCTGGGGGCATACG
CCCTCATCAAAATTTGGGCGGAACTTGGGGCGAGGACGATGATGGCCCGATAGCACCCAG
CGTTTGGACGGGTGAGTCATTCCACATATGCCAACGCTCTGGTGTGACAGTCGGTGCCA
TAGCGCTGGCCGTTGGCCCGCTGCTGGTCCCTAATGGGGACAGGCTGTGCTGTTGG
TGTTGGAGTCGGAGTTGCCTTAACTCGACTGGAAAATAACAATGCGCCGGCAACAGGAG
CCCTGCCTGCCGTGGCTCGTCCGAAATGTGGGGACATCCTCAGATTGCTCACAATC
ATCGCCGGGAATGNTAANGAATTAATCAAAATTTGGCCGACATAATGNCAGATTGAGA
ACGATTTAACAAAAATGGTCGGCCCGCTTGTAGTGCAACAGGGTCAAATATCGCAAGCT
CAAAATTTGGCCCAAGCGGTGTGGTTCCGTATCCGGTAATGTCGGGGCACAATGGGGA
GCCACACAGGCCCGCTTGGGGCCCAAGGTATTTCCAAGCAATCACTGGATGGGAGGA

```



7

Evidence-based annotation

- Human-curated analysis
 - 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

8

Collect, analyze, and synthesize

- **Collect:**
 - GEP UCSC 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

9

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"

10

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

11

Annotation Goals

- Identify the ortholog of a gene of interest from an **informant** genome (*D. melanogaster*) in the **target** genome (e.g., *D. yakuba*)
- For **ALL** unique isoforms, identify and map the locations of all coding exons (**CDS**) in the target genome

12

Control Display of Evidence Tracks

- 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 only have a subset of these display modes

19

GEP UCSC Genome Browser

Each evidence track in the Genome Browser has an associated description page that contains a discussion of the track, the methods used to create the track, and, in some cases, filter and configuration options to fine-tune the information displayed in the track (e.g., RNA-Seq tracks). To view the description page, click on the gray mini-button to the left of a displayed track (left) or on the label for the track in the display controls section (right).

- Click or drag in the base position track to zoom in
- Drag tracks left or right to a new position
- Drag gray mini-button up or down to reorder tracks
- Type "T" for keyboard shortcuts

20

Annotation Workflow

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

Repeat steps 5-7 for each **unique** isoform

21

Download the Annotation Workflow from the GEP website

Pathways Project Workflow

Project Curriculum

Pathways Project: Annotation Walkthrough

Pathways Project: Annotation Form Exemplar

Pathways Project: Annotation Form

Pathways Project: Annotation Form Exemplar

22

Annotation Workflow (1)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

23

Step 1

- **Genomic neighborhood:** A local region of the genome containing a small number of genes (5-15) or other features (e.g., repeats).
 - The genomic neighborhood, as related to the Pathways Project, is the **region containing the target gene and its neighboring two closest upstream genes and two closest downstream genes.**
- Target gene: *Rheb*
- Target species: *D. yakuba*

24

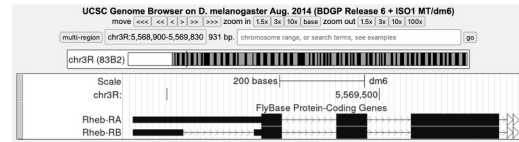
Protocol for examining the genomic neighborhood surrounding target gene (*Rheb*) in *D. melanogaster*

- Use [Genome Browser](#) to view target gene in the *D. melanogaster* Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) assembly
1. Identify orientation (+/- DNA strand) of target gene and its nearest genomic neighbors in *D. melanogaster*
 2. Once you identify the two neighboring genes on each side of your target gene, the arrows on the target gene will determine upstream (before, 5') or downstream (after, 3')

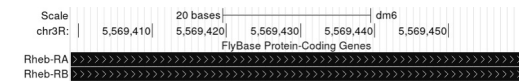
25

Determine the orientation of a feature in the Genome Browser

- Use the direction of the arrows within introns

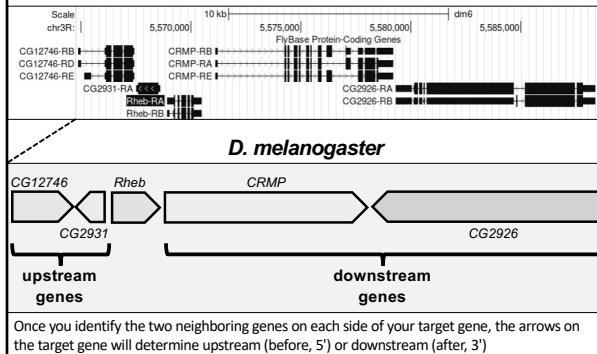


- If the viewing region for the feature does not contain introns, use the direction of the arrows within exons



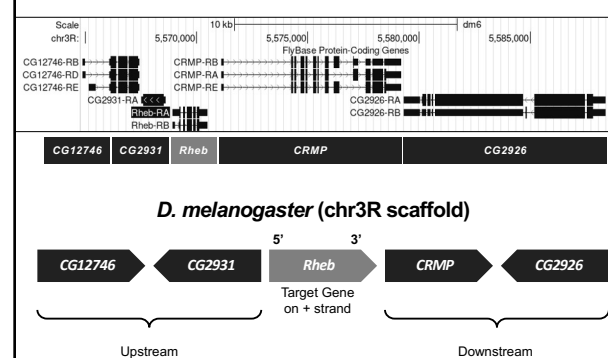
26

Region surrounding the *D. melanogaster* *Rheb* gene



27

Step 1: Examine genomic neighborhood surrounding target gene (*Rheb*) in *D. melanogaster*



28

Annotation Workflow (2)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

29

Step 2: Identify genomic location of ortholog in target species (*D. yakuba*)

- Use the [Genome Browser](#) to retrieve protein sequence of target gene from *D. melanogaster*

- Click on an isoform in 'FlyBase Protein-Coding Genes' track; then click on 'Translated Protein from predicted mRNA'

1. Use 'Genome BLAST' link of target species genome to run *tblastn* search of *D. melanogaster* protein

- **Query:** *D. melanogaster* protein
- **Database:** target species' whole genome assembly

2. Identify best **collinear** set of protein alignments in target species-coordinates, scaffold name & accession

30

Detect sequence similarity with BLAST

- BLAST = **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Why is BLAST popular?
 - Provide statistical significance for each match
 - Good balance between sensitivity and speed
- Identify **local** regions of similarity

31

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:

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

Arrows indicate the BLAST program translates the nucleotide sequence *before* performing the search.

32

Genome BLAST links for the Pathways Project

- Limit BLAST searches to the **specific genome assembly** that is part of the Pathways Project

Pathways Project Genome Assemblies

Last Update: 2026-01-02

Species	Genome Browsers	NCBI BLAST Link
D. melanogaster	Aug. 2014 (BDGP Release 6 + ISO1 MT6m)	Genome BLAST
D. mauritiana	Apr. 2020 (UC Irvine ASM438214v1/DmaurRefSeq1)	Genome BLAST
D. sechellia	Feb. 2020 (UC Irvine ASM438219v1/DsecRefSeq1)	Genome BLAST
D. simulans	Oct. 2021 (Princeton Prin_Dsim_3.1/OsimRefSeq3)	Genome BLAST
D. yakuba	Aug. 2021 (Princeton Prin_Dyak_Tai18E2_2.1/DyakRefSeq3)	Genome BLAST
D. santomea	Sep. 2021 (Princeton Prin_Dsan_1.1/DsanRefSeq2)	Genome BLAST

33

Identify genomic location of ortholog in target species (*D. yakuba*)

34

Expect **multiple alignment blocks** when aligning a protein with multiple CDS's against a scaffold

35

Use the **Hit Table** to determine the best collinear set of *tblastn* alignments

Range	<i>D. melanogaster</i> (Query) Start	<i>D. melanogaster</i> (Query) End	Target Species (Subject) Start	Target Species (Subject) End	E-Value	Identities (%)	Subject Frame
1	6	44	11,148,568	11,148,431	0.002	46	-3
2	18	108	11,418,759	11,419,094	5e-06	28	+3
3	1	20	19,150,809	19,150,868	2e-78	90	+3
4	16	45	19,150,981	19,151,070	2e-78	83	+1
5	40	109	19,151,150	19,151,359	2e-78	97	+2
6	111	153	19,151,422	19,151,550	2e-78	93	+1
7	153	182	19,151,610	19,151,699	2e-78	93	+3

best collinear set of alignments to Rheb-PA

36

Genome databases use **different names** (sequence IDs) to refer to the **same sequence**

- INSDC (GenBank, EBI, DDBJ): CM028602.2
- NCBI RefSeq: NC_052530.2
- UCSC Genome Browser: NC_052530

NCBI RefSeq records are derived from, but might not be identical to, the source INSDC records

37

GEP UCSC Genome Browser supports INSDC and RefSeq accession numbers

RefSeq accession

INSDC accession

38

Annotation Workflow (3)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

39

Step 3: Examine genomic neighborhood of putative ortholog in target species (*D. yakuba*)

1. Use target species' **Genome Browser** to navigate to the collinear set of alignments identified in Part 2
2. Use **synteny to verify ortholog assignment**
 - o Run *blastp* search for target gene AND its two upstream and two downstream genes on either side
 - **Query:** accession of each protein in target species
 - Click on isoform in 'NCBI RefSeq Genes' track;
 - Click on the accession number link next to the 'GenBank Record' label to view the RefSeq mRNA record in a new window;
 - Scroll down to the 'CDS' section and copy accession number for the translated protein sequence labeled 'protein_id'
 - **Database:** refseq_protein | Organism: *Drosophila melanogaster* (taxid: 7227)

40

Use the NCBI RefSeq Genes predictions to gather **additional evidence** for the ortholog assignment

- Each bioinformatics tool has different strengths and weaknesses
- Trade-offs:
 - o Sensitivity versus specificity
 - o Required computational resources versus accuracy
- Program parameters often determined by heuristics
 - o Optimized for the whole genome assembly

Need to evaluate and reconcile potentially contradictory results on the genome browser

41

Retrieve the protein sequence for the NCBI RefSeq Gene prediction (1)

Click on the accession number link to view the GenBank Record for the RefSeq mRNA

Scroll to the bottom of the GenBank Record window to view sequence of translated protein from predicted mRNA

42

Retrieve the protein sequence for the NCBI RefSeq Gene prediction (2)

PREDICTED: *Drosophila yakuba* GTP-binding protein Rheb homolog (LOC6537476), mRNA

NCBI Reference Sequence: XM_039231796.2
FASTA GenBank

```

... ..
CDS             300..848
               /gene="LOC6537476"
               /codon_start=1
               /product="GTP-binding protein Rheb homolog"
               /protein_id="XP_039231796.1"
               /db_xref="GeneID:6537476"
               /translation="MPTKERLIAMGVRSGKSSLCIQVEGQFVDSYDPTIENTFTK
               IERVKSDQYIVKLIDTAGQDEYISIFPQVYSMDYHGIVLYVYSITSQKSFVWKIYYEKL
               LDVNGKYYVIVLVKNTLDLQPERTYSTEEGKLAESWRAAFLETSAKQNESVGDIFW
               QLTLILENENPQKSSCLVSV"
               /polyA_site
               1226
               /gene="LOC6537476"
               /experiment="COORDINATES: polyA evidence [ECO:0006239]"
ORIGIN
1   ttgcatact tcgacagcag cactgacta actgagaat tactgtttc ttttagaga
61   ttctgaaat aataatgaa ttaaatcgg cggaaagcc ctatctgga tagcaacaag
121  tatcttcag aagaabaata aatcagaag caatattca cacataaca gttatagag
181  caaagcatt cccacacaag gcacaaaaa gtcgaagga cgcagcaac aaaaaagtc
    
```

Sequence of translated protein from predicted mRNA

43

Configure the NCBI *blastp* search of XP_039231796

Standard Protein BLAST

Enter Query Sequence: XP_039231796
protein sequence of the putative ortholog to Rheb-PA in *D. yakuba*

Job Title: XP_039231796-GTP-binding protein Rheb homolog

Choose Search Set: Reference proteins (refseq_protein)

Organism: *Drosophila melanogaster* (taxid:7222)

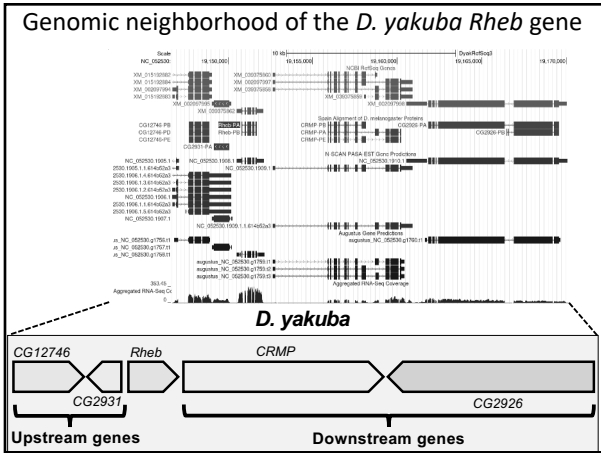
Program Selection: blastp (protein-protein BLAST)

BLAST

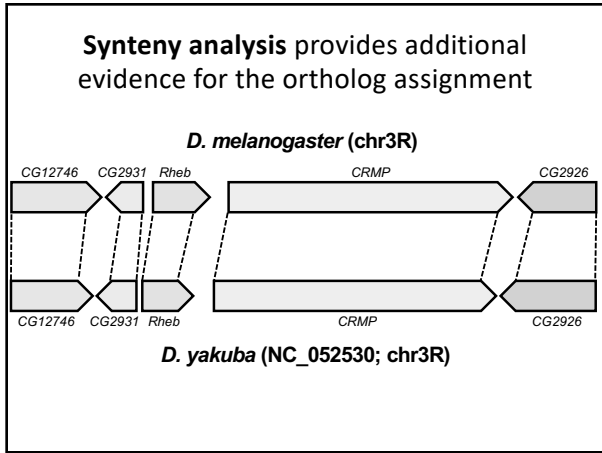
Query: sequence we are looking to match
- predicted protein sequence of the putative ortholog to Rheb-PA in *D. yakuba*

Database (Subject): collection of sequences we are searching for matches
- *D. melanogaster* reference protein database (refseq_protein)

44



45



46

Annotation Workflow (4)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

47

Step 4: Determine structure of target gene (*Rheb*) in *D. melanogaster*

- Use the **Gene Record Finder** to identify the gene structure of the target gene (*Rheb*) in *D. melanogaster*

48

Retrieve the *Rheb* record for *D. melanogaster* using the Gene Record Finder

- Type ***Rheb*** into the search box, then press [Enter]

49

The A and B isoforms of *Rheb* use the same set of CDS's in *D. melanogaster*

Transcript Details	Polypeptide Details					
Options: Export All Unique CDS to FASTA Export All CDS for Selected Isoform to FASTA Download CDS Workbook						
CDS usage map:						
Isoform	1_9812_0 2_9812_2 3_9812_2 4_9812_1 5_9812_0					
Rheb-PA	1 2 3 4 5					
Rheb-PB	1 2 3 4 5					
Isoforms with unique coding exons:						
Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences					
Rheb-PA	Rheb-PB					
Select a row to display the corresponding CDS sequence:						
FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)	
1_9812_0	5,569,223	5,569,271	+	0	16	
2_9812_2	5,569,400	5,569,471	+	2	23	
3_9812_2	5,569,575	5,569,782	+	2	68	
4_9812_1	5,569,842	5,569,971	+	1	43	
5_9812_0	5,570,028	5,570,117	+	0	30	

50

Annotation Workflow (5)

- Examine genomic neighborhood surrounding target gene in *D. melanogaster*
- Identify genomic location of ortholog in target species
- Examine genomic neighborhood of putative ortholog in target species
- Determine structure of target gene in *D. melanogaster*
- Determine approximate location of coding exons in target species
- Refine coordinates of coding exons
- Verify and submit gene model

51

Step 5: Determine approximate location of coding exons in target species

- Obtain the protein sequence of each CDS of the target gene in *D. melanogaster* using the Gene Record Finder (in Polypeptide Details tab, click on each CDS and then copy/paste the entire sequence from the pop-up window)
- Run *tblastn* search (**check the box to align two or more sequences**) for each CDS of the target gene
 - Query:** each individual CDS of target gene in *D. melanogaster*
 - Database:** accession number of scaffold containing ortholog in target species (identified in Part 2) and narrow region of scaffold to search via "Subject subrange"
 - Algorithm parameters:** "Compositional adjustments" - No adjustment and uncheck box for filtering "Low complexity regions"

52

Detect conserved *D. melanogaster* CDSs in the target genome with *tblastn*

- Coding sequences evolve slowly
- Exon structure changes **very** slowly
- Initial hypothesis:
 - Gene structure is conserved between *D. melanogaster* and the ortholog in the target species
- Map each CDS separately to the target genome

53

Retrieve the *D. melanogaster* CDS sequence from the Gene Record Finder

- In the **Polypeptide Details** section of the Gene Record Finder, select a row in the CDS table

54

Compare the *D. melanogaster* CDS 1_9812_0 against the *D. yakuba* scaffold NC_052530

Align Sequences Translated BLAST: tblastn

Enter Query Sequence
 Enter accession number(s), g(i)s, or FASTA sequence(s) Clear
 >Rheb:1_9812_0
 MPTKERHIAMMGYRSV CDS-1 (Rheb:1_9812_0)
 from *D. melanogaster*
 Query subrange From To

Or, upload file No file selected.
 Job Title Rheb:1_9812_0
 Enter a descriptive title for your BLAST search

Align two or more sequences
 Enter Subject Sequence
 Enter accession number(s), g(i)s, or FASTA sequence(s) Clear Subject subrange
 NC_052530 D. yakuba NC_052530 scaffold
 From 19051000 To 19252000
 Limit the search area of the NC_052530 scaffold to the region surrounding our putative ortholog

Don't include commas in the "Subject subrange" or BLAST will search outside of the region.

55

Customize algorithm parameters to increase the sensitivity of the tblastn search

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with a sign

Algorithm parameters

General Parameters
 Max target sequences
 Expect threshold
 Word size
 Max matches in a query range

Scoring Parameters
 Matrix
 Gap Costs
 Compositional adjustments No adjustment Turn off compositional adjustments

Filters and Masking
 Filter Low complexity regions Turn off the low complexity filter
 Mask Mask for lookup table only Mask lower case letters

56

Record the results of the tblastn alignment to CDS 1_9812_0

Descriptions Graphic Summary **Alignments**

Alignment view
 1 sequences selected

Download GenBank Graphics
 Drosophila yakuba strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2.2.1, whole genome shotgun
 Sequence ID: NC_052530.2 Length: 30730773 Number of Matches: 1
 Range 1: 19150809 to 19150856 GenBank Graphics Next Match Previous Match
 Score 34.3 bits (77) Expect 1e-06 Identities 15/16 (94%) Positives 16/16 (100%) Gaps 0/16 (0%) Frame +3
 Query 1 MPTKERHIAMMGYRSV 16
 19150809 MPTKERHIAMMGYRSV 19150856
 Sbjct 19150809 MPTKERHIAMMGYRSV 19150856
 Query Descr Rheb:1_9812_0
 Query Length 16

Query: sequence we are looking to match
 • each individual CDS of Rheb in *D. melanogaster*
 Database (Subject): collection of sequences we are searching for matches
 • BLAST will translate NC_052530 scaffold sequence of *D. yakuba*

57

Summary of the tblastn results for Rheb CDS's against the D. yakuba scaffold NC_052530

CDS	FlyBase ID	Query Length Size (aa)	D. melanogaster (Query) Start	D. melanogaster (Query) End	Target Species (Subject) Start	Target Species (Subject) End	Subject Frame
1	1_9812_0	16	1	16	19,150,809	19,150,856	+3
2	2_9812_2	23	1	23	19,150,987	19,151,055	+1
3	3_9812_2	68	1	68	19,151,156	19,151,359	+2
4	4_9812_1	43	1	43	19,151,422	19,151,550	+1
5	5_9812_0	30	1	30	19,151,613	19,151,702	+3

- Query: CDS's from the *D. melanogaster* Rheb gene
- Subject: *D. yakuba* scaffold NC_052530 (subrange: 19051000-19252000)

58

Annotation Workflow (6)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

59

Step 6: Refine coordinates of coding exons

Use the Genome Browser to:

1. Verify start and stop codon coordinates identified in Part 5
2. Determine phases of donor and acceptor splice sites using Frame identified in tblastn CDS-by-CDS search in Part 5
 - o donor is typically a "GT" and acceptor is typically an "AG" (Georgia Tech is in Atlanta Georgia)
3. Use splice junction predictions to verify coordinates of each intron

60

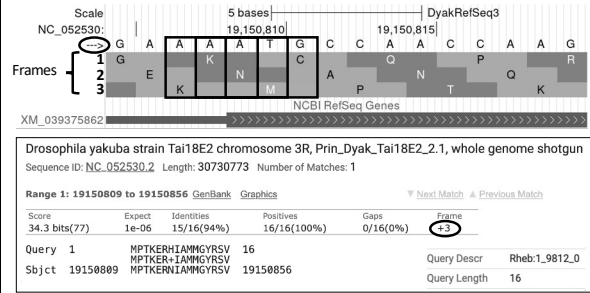
The proposed gene model should satisfy **basic biological constraints***

- 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

61

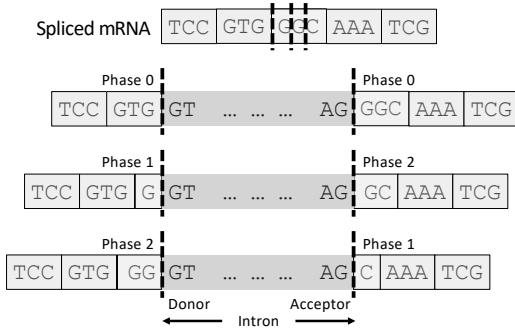
A genomic sequence has 6 different reading frames



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

62

A codon could be derived from nucleotides in adjacent exons



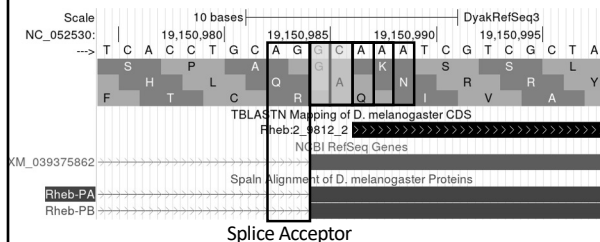
63

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

64

Phase depends on the reading frame

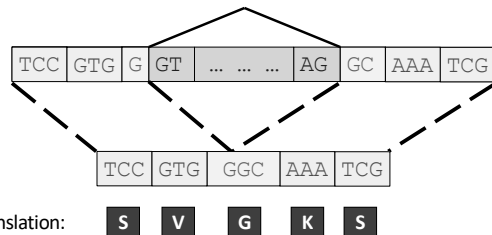


- Phase of acceptor site:
 - Phase **2** relative to frame +1
 - Phase **0** relative to frame +2
 - Phase **1** relative to frame +3

65

Phase of donor and acceptor sites must be compatible

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



66

Incompatible donor and acceptor phases results in a frame shift

Translation: **S V G A N**

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

67

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

68

Overview of RNA-Seq (Illumina)

Wang Z et al. Nat Rev Genet. 2009 10(1):57-63.

69

Use spliced RNA-Seq reads to identify splice sites

70

RNA-Seq evidence tracks on the GEP UCSC Genome Browser

71

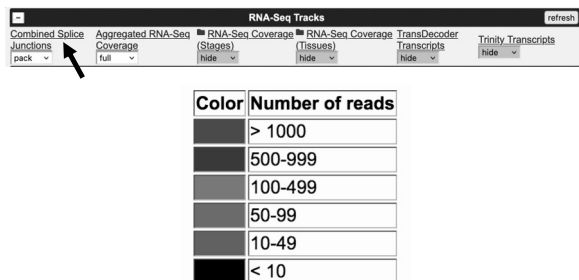
Splice junction score corresponds to the number of spliced RNA-Seq reads supporting the predicted intron

Item: JUNC00113129
 Score: 8
 Position: NC_052530:19151014-19151152
 Genomic Size: 139
 Strand: +

Item: JUNC00113128
 Score: 5875
 Position: NC_052530:19150962-19151252
 Genomic Size: 291
 Strand: +

72

Color of the splice junctions correspond to the number of reads supporting the junction



73

Gene model for Rheb-PA
in *D. yakuba*

CDS #	FlyBase ID	Frame	Splice Acceptor Phase	Start Coordinates	End Coordinates	Splice Donor Phase
1	1_9812_0	+3		19,150,809	19,150,857	1
2	2_9812_2	+1	2	19,150,985	19,151,056	1
3	3_9812_2	+2	2	19,151,154	19,151,361	2
4	4_9812_1	+1	1	19,151,421	19,151,550	0
5	5_9812_0	+3	0	19,151,613	19,151,699	

Stop codon: 19,151,700-19,151,702

74

Annotation Workflow (7)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

75

Step 7: Verify and submit gene model

- Use the [Gene Model Checker](#) to verify that your proposed gene model satisfies the basic biological constraints (e.g., begins with a start codon, has compatible splice sites, and ends with a stop codon)

76

Verify the final gene model using the Gene Model Checker

- Examine the checklist and explain any errors or warnings in the Pathways Project 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 Annotation Report
 - Look for large vertical and horizontal gaps
 - See the “[Quick Check of Student Annotations](#)” document on the GEP web site

77

Annotation Workflow (8)

1. Examine genomic neighborhood surrounding target gene in *D. melanogaster*
2. Identify genomic location of ortholog in target species
3. Examine genomic neighborhood of putative ortholog in target species
4. Determine structure of target gene in *D. melanogaster*
5. Determine approximate location of coding exons in target species
6. Refine coordinates of coding exons
7. Verify and submit gene model

Repeat steps 5-7 for each **unique** isoform

78