

BIOL199 BDB
Fall 2022

Lab 7: Beginning the Pathways Project, Annotation Walkthrough Parts A-C

Adapted by Melinda A. Yang, from "Pathways Project: Annotation Walkthrough" by GEP members Katie Sandlin, Wilson Leung, and Laura Reed, and from "Pathways Project Annotation Notebook" by GEP members Katie Sandlin and Alexa Sawa

Table of Contents

Introduction	1
A. Examine our target gene and its genomic neighborhood surrounding target gene in <i>D. melanogaster</i>	1
B. Identify genomic location of ortholog in target species	4
C. Examine genomic neighborhood of putative ortholog in target species	8

Worksheet link (you will receive a hard copy at start of lab)

[Appendix Link \(you will receive a hard copy at start of lab\)](#)

(Learning) Objectives:

- Use the Genome Browser and the FlyBase website to learn about your target gene in *D. melanogaster*
- Use *BLAST* to help you to construct a hypothesis on the location of the putative ortholog in your target species for your target gene
- Analyze multiple lines of evidence to support your hypothesis on the location of the putative ortholog in your target species for your target gene

Pre-lab Assignments:

1. Read Appendix 1-3 in depth, and you may find it useful to additionally read Appendix 4-6. I encourage reviewing the protocol as well.
2. Complete the pre-lab quiz before lab at http://tiny.cc/pathwaysA-C_prelabquiz

Introduction

The Pathways Project is focused on annotating genes found in well characterized signaling and metabolic pathways across the *Drosophila* genus. The current focus is on the insulin signaling pathway which is well conserved across animals and critical to growth and metabolic homeostasis. The long-term goal of the Pathways Project is to analyze how the regulatory regions of genes evolve in the context of their positions within a network.

The images accompanying the steps for annotation of the coding regions for the Ras homolog enriched in brain (*Rheb*) gene (**target gene**) in *Drosophila yakuba* (**target species**). **You will work through the instructions, but focus on annotation of your gene and species. This means your images will differ from the ones in the example. You may find it useful to repeat the instructions for *Rheb* in *D. yakuba* if you find yourself getting confused.**

It is important to note that the *Rheb* gene in *D. yakuba* is relatively straightforward in comparison to what your own project might be. Some steps might appear to be excessive, but keep in mind that your gene/species may be much more complex, so it is important to follow this protocol as we have tried to equip you with most of what you will need should you encounter complexities.

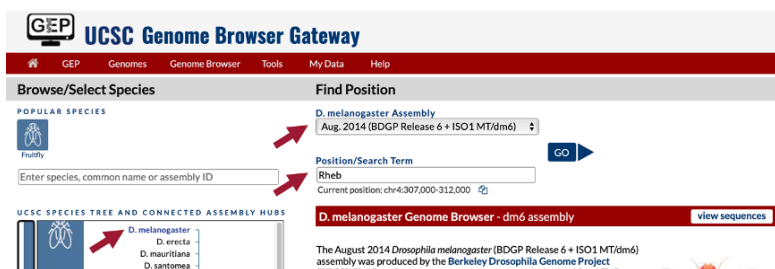
In Parts A-C, you will learn about your target gene in D. melanogaster (Part A), form a hypothesis on where the gene is in your target species (Part B), and use homology-based arguments (i.e. synteny) to make the case that the region you specified is where your target gene is located in your target species (Part C).

A. Examine our target gene and its genomic neighborhood in *D. melanogaster*

The goal of this section is to understand the context of the target gene for the model organism *D. melanogaster*.

Instructions

1. **Make sure you know your target gene and target species.** Before beginning any analysis, let's learn a little bit about our target gene. Navigate to FlyBase (<https://flybase.org/>) and enter the gene symbol in the "Jump to Gene" text box in the top right-hand corner of the FlyBase home page. Find the 'Gene Summary' and read the description of your target gene's function.
2. Navigate to the GEP UCSC Genome Browser Gateway Page (http://tiny.cc/GEP_BrowserGateway).
 - Click on "D. melanogaster" in the "REPRESENTED SPECIES" table.
 - Ensure "Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)" under the "D. melanogaster Assembly" field is selected.
 - Enter your gene symbol under the "Position/Search Term" field.
 - Click on the "Go" button.

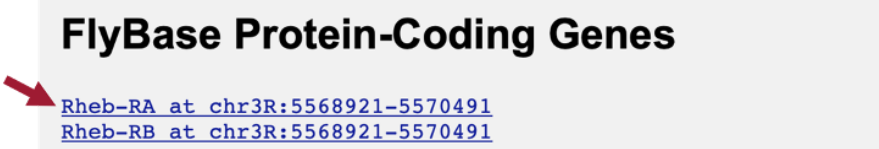


Comments

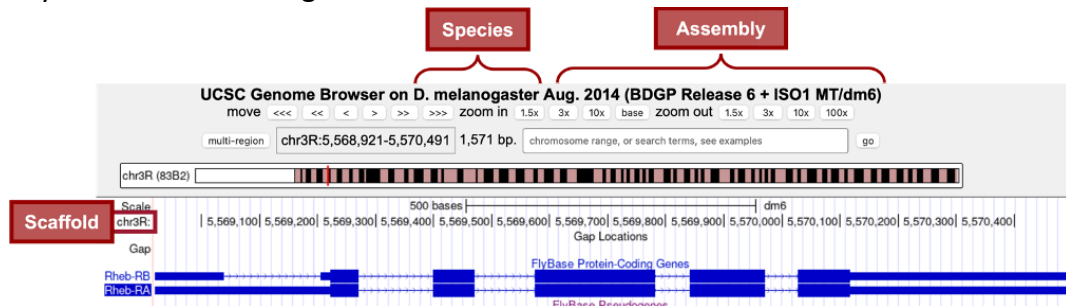
Later on, we will do more online research to find more about your gene.

Answer questions A1-A2

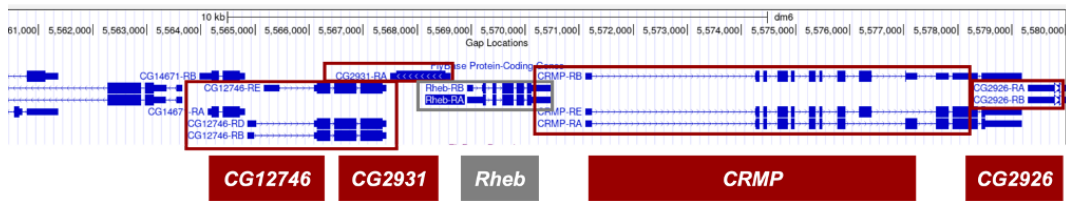
3. You may get something like the following page if there are multiple matches. If so, under “FlyBase Protein-Coding Genes,” click on the first option (it should have the gene name in the line).



4. Because the Genome Browser remembers our previous track display settings, click on “default tracks” in the display configuration buttons below the Genome Browser image. This should only show the gene structure models from the “FlyBase Protein-Coding Genes” track.



5. Zoom out until you can see two genes on either side of your gene.



The genomic neighborhood of *Rheb* includes *CG12746*, *CG2931*, *CRMP*, and *CG2926* (red boxes).

+ strand: the coding strand for genes transcribed from left to right

- strand: the coding strand for genes transcribed from right to left

Upstream: located on the 5' side of the coding strand of the target gene

Downstream: located on the 3' side of the coding strand of the target gene

Upstream and downstream are relative to your gene's orientation.

DISCUSS and CHECK-IN: For *Rheb*, which is transcribed from left to right in this image (see arrow), which of the four genes listed are upstream, and which are downstream? For each neighboring gene, how many isoforms does it have (put next to each gene symbol)?

Upstream genes: _____ Downstream genes: _____

6. Click ‘default tracks’ and then adjust tracks so you set a comparative genomics track (e.g., Drosophila Conservation (28 Species)) to “pack”. Then, click on “refresh”

Note that *Rheb* has two isoforms in *D. melanogaster*, spanning the same coordinates. It doesn't matter which isoform you select here, but if the **coordinates of the isoforms in your gene are different, you should choose the longest isoform for this step.**

In the “FlyBase Protein-Coding Genes” track, we can see the gene structure for the two isoforms of the *Rheb* gene (i.e., *Rheb-RA* and *Rheb-RB*). Note that *Rheb* is on the “chr3R” scaffold.

Answer question A3

We are now viewing the genomic neighborhood of the *Rheb* gene in *D. melanogaster* (i.e., region of the chr3R scaffold containing the *Rheb* gene and its two closest upstream and two closest downstream genes).

Remember a scaffold is a larger contiguous DNA sequence that are composed of multiple contigs (which are in turn composed of multiple reads). They usually represent a large section of a chromosome.

7. Take a screenshot of your genome browser with the genomic neighborhood. To do so, follow these directions:

- Right click somewhere inside the Genome Browser image
- Select 'View image'
- A new tab should open with the image (if not, make sure your browser has pop-ups enabled)
- Right-click on the image in the new browser tab and click 'Save Image As...'
- Save the file into a folder on your computer for the Pathways Project, preferably as: <TeamName>_<GeneName>_dmel_GenomicNeighborhood.png

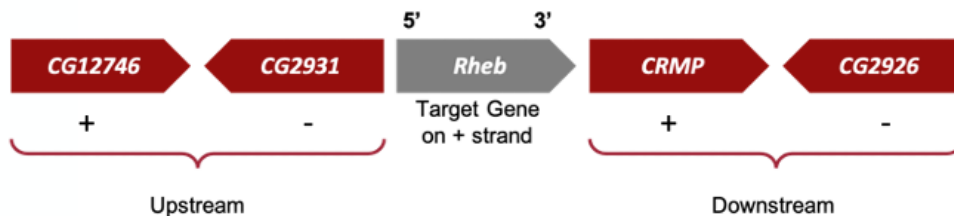
CHECK-IN: Check with us that you saved the image correctly before uploading as per question A4.

8. Reset your tracks to 'default tracks'. We aren't concerned with the comparative genomics track for now, but if you develop a report to submit to the GEP, this image needs to include the comparative genomics track for further review.

9. Draw a sketch of the genomic neighborhood of your gene in *D. melanogaster*. To do so, complete the following:

- Zoom into your gene to examine the direction of the arrows.
- If pointing to the right, your sketch will require an arrow pointing to the right labeled with your gene name.
- Zoom out and repeat for each of the genes in the neighboring region.
- The closest genes are the first ones whose CDS (i.e. thicker rectangles) you hit moving from your target gene right or left.

D. melanogaster (chr3R scaffold)



10. If your target gene has multiple isoforms in *D. melanogaster*, determine the longest isoform for your *tblastn* search in the next section. We will use the longest isoform with the longest protein-coding region. This is because we will compare the translated protein to the genomic assembly of your target species. A larger combination of letters to search makes it more likely the returned hit is not due to chance.

Note that we are NOT taking a screenshot in a typical way. This is because the image produced by clicking 'view image' will be standardized unlike what someone might take with a typical screenshot.

Filename for example would be:
TeamAwesome_Rheb_dmel_GenomicNeighborhood.png

Answer question A4

Since the direction of the arrows within their introns point to the right, *Rheb*, *CG12746*, and *CRMP* are on the positive strand. *CG2931* and *CG2926* are on the negative strand since the direction of the arrows within their introns point to the left.

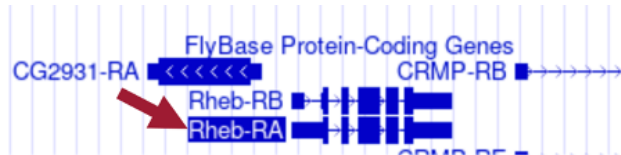
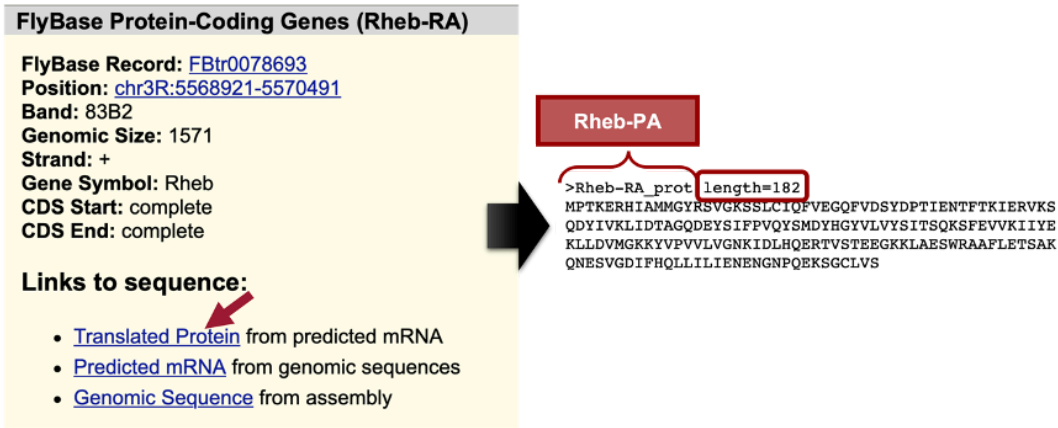
Each protein-coding gene annotated by FlyBase in *D. melanogaster* has an annotation symbol that begins with the prefix "CG" (i.e. Computer Gene). Unless genes are characterized experimentally and formally named, they are referred to by this symbol.

Answer question A5

Answer question A6

B. Identify genomic location of ortholog in target species

Now that we’ve examined the genomic neighborhood of *Rheb* in *D. melanogaster* (dmel), we need to identify the location of your gene in your target species.

Instructions	Comments
<p>1. In the “FlyBase Protein-Coding Genes” track in your genome browser, click on the gene symbol/isoform name on the left of the gene model to view additional details about your gene.</p>  <p>Click on the gene symbol/isoform name on the left of the gene model in the FlyBase Protein-Coding Genes track, in this case the “Rheb-RA”</p>	
<p>2. Under the “Links to sequence” heading, click on the “Translated Protein” link</p> 	<p>After clicking ‘Translated Protein’, you should see a string of letters much like the image on the right. This is a FASTA file showing the sequence of amino acids (aa) for the translated protein of your gene in Dmel. There are 182 aa in the translated protein of Rheb-RA in Dmel.</p> <p>Answer question B1.</p>
<p>3. Copy the entire sequence (including the header) so we can use it in our <i>tblastn</i> search.</p>	<p>Perhaps good to put the sequence in your lab notebook!</p>
<p>4. Open a new tab and navigate to the Pathways Project Genome Assemblies page (http://tiny.cc/genome_assemblies). Using links for the Pathways Project Genome Assemblies ensures you all always navigate to the correct genome assembly database when performing BLAST searches.</p>	<p>The reason we are using this link is because the National Center for Biotechnology Information (NCBI) periodically updates the genome assemblies used for BLAST searches, which can cause a host of issues in a class mid-semester.</p>

5. Click on the “NCBI *BLAST* Link” for your target species. By clicking the link for your target species, you are selecting the **entire** genome of your target species as your database to search.

GEP Pathways Project Genome Assemblies		
Last Update: 2022-06-28		
Show/Hide Columns	Clear Filters	Export to CSV
Species	Genome Browsers	NCBI BLAST Link
<input type="text" value="Search Species"/>	<input type="text" value="Search Genome Browsers Column"/>	<input type="text" value="Search NCBI BLAST"/>
<i>D. melanogaster</i>	Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)	Genome BLAST
<i>D. erecta</i>	Oct. 2018 (AGI DereRS2/DereRefSeq1)	Genome BLAST
<i>D. sechellia</i>	Feb. 2020 (UC Irvine ASM438219v1/DsecRefSeq1)	Genome BLAST
<i>D. simulans</i>	Oct. 2021 (Princeton Prin_Dsim_3.1/DsimRefSeq3)	Genome BLAST
<i>D. yakuba</i>	Aug. 2021 (Princeton Prin_Dyak_Tai18E2_2.1/DyakRefSeq3)	Genome BLAST
<i>D. ananassae</i>	Sep. 2021 (University of Maryland, ASM1763931v2/DanaRefSeq2)	Genome BLAST

We need to *BLAST* our *Rheb* protein sequence from *D. melanogaster* against the entire genome of *D. yakuba* to narrow down the possible regions where *Rheb* could be in *D. yakuba*.

Answer question B2.

6. Configure BLAST to use the protein sequence for the gene in *D. melanogaster* to do a BLAST search to find the corresponding DNA region in your target species.

- Make sure the “*tblastn*” tab is selected
- Paste the protein sequence for your gene that we copied from step B3 into the “Enter Query Sequence” text box. The “Job Title” text box should then automatically populate.
- Select the check box next to “Show results in a new window”.
- Click on the “BLAST” button.

The screenshot shows the NCBI BLAST tblastn interface. A red box highlights the 'tblastn' tab, with a note: 'tblastn searches translated nucleotide database using a protein query'. Another red box highlights the 'Enter Query Sequence' field, which contains the protein sequence for *D. melanogaster* Rheb-PA, with a note: '*D. melanogaster* Rheb-PA (protein query)'. A third red box highlights the 'Database' dropdown, which is set to 'Prin_Dyak_Tai18E2_2.1 RefSeq assembly [GCF_016746365.2]', with a note: '(nucleotide database containing the entire *D. yakuba* genome assembly)'. The 'Job Title' field is populated with 'Rheb-RA_prot length=182'. The 'BLAST' button is visible at the bottom.

Note that ‘-RA’ refers to mRNA, whereas ‘-PA’ refers to the protein sequence. Isoforms will always have the suffix and are not italicized, whereas gene symbols are italicized and have no suffix. (e.g. *Rheb* is the gene, *Rheb-RA* is an mRNA isoform, *Rheb-PA* is a protein isoform).

For *tblastn*, the *BLAST* search will translate the entire genome of *D. yakuba* before searching for a match to the *Rheb-PA* sequence from *D. melanogaster*, and compare the translated amino acids for all reading frames to the amino acid sequence for *D. melanogaster*.

DISCUSS and CHECK-IN: Explain what your query sequence is and what your database/subject sequence are for your *tblastn* analysis.

Query sequence: _____ Database/subject sequence: _____

7. Take a moment to review [Appendix 1](#) about BLAST results. Indicate here whether you may have found a putative ortholog of your gene in your target species, and if so, what is the accession number. Indicate your reasoning for what lines of evidence support your hypothesis.

8. Take a screenshot of the “Descriptions” panel of your *tblastn* results of the amino acid sequence of the *D. melanogaster* protein coding isoform for your target gene. Name the image file *GeneName_TargetSpecies_tblastn.png* and save it into your Pathways directory.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Drosophila yakuba strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	137	735	100%	2e-78	97.14%	30730773	NC_052530.2
<input checked="" type="checkbox"/> Drosophila yakuba strain Tai18E2 chromosome 3L, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	136	454	96%	7e-37	43.75%	25180761	NC_052529.2
<input checked="" type="checkbox"/> Drosophila yakuba strain Tai18E2 chromosome X, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	85.1	571	91%	8e-19	35.57%	24674056	NC_052528.2
<input checked="" type="checkbox"/> Drosophila yakuba strain Tai18E2 chromosome 2L, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	83.2	324	90%	3e-18	33.33%	31052931	NC_052527.2
<input checked="" type="checkbox"/> Drosophila yakuba strain Tai18E2 chromosome 2R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	77.4	275	80%	3e-16	30.40%	23815334	NC_052528.2

This is a typical screenshot. Model it after Figure 1 in the link from the previous step.

For more on accession numbers, see [Appendix 2](#). Note that your accession number should be in your assembly's 'view sequences' list on Genome Browser.

Answer questions B3-B4.

9. Under the “Descriptions” tab, uncheck the box next to “select all” to hide alignments from accessions we are not interested in (i.e. less good matches)

10. Click on the putative ortholog link in the “Description” column to navigate to the alignment.

For *Rheb* in *D. yakuba*, you'd click on “*Drosophila yakuba* strain Tai18E2 chromosome 3R, ...”

11. In the blue toolbar for the *BLAST* hit, **select the “Subject start position” option** from the drop-down menu of the “Sort by” field to order the matches based on the start coordinates, in ascending order.

Alignment view: Pairwise

1 sequences selected

Download GenBank Graphics Sort by: **Subject start position**

Drosophila yakuba strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun sequence

Sequence ID: **NC_052530.2** Length: 30730773

Range 1: 19151150 to 19151359

Score: 137 bits (345) 2e-78 Expect: 68/70 (97%) Method: Compositional matrix adjust. Identities: 68/70 (97%) Positives: 68/70 (97%) Gaps: 0/70 (0%) Frame: +2

Features: gtp-binding protein *rheb* homolog

Query: 40 NTFTKIERVKSQDIYVKLIDTAGQDEYSIFPVQYSDYHGVLVYSITSQKSFVVKLIY 99

Sbjct: 7623630 AQTWDTAGQERYRAITSAYYRGAVGALLVYDIKHLTYENVERWLRLRDHADQNIY-IM 7623886

Score: 60.5 bits (145) Expect: 2e-10 Method: Compositional matrix adjust. Identities: 39/109 (36%) Positives: 57/109 (52%) Gaps: 1/109 (0%) Frame: +3

Features: ras-related protein *rab-11a*

Query: 55 VKLIDTAGQDEYSIFPVQYSDYHGVLVYSITSQKSFVVKLIY 163

Sbjct: 7623630 AQTWDTAGQERYRAITSAYYRGAVGALLVYDIKHLTYENVERWLRLRDHADQNIY-IM 7623886

Query: 115 LVGNKIDLHQERTVSTEEGKLAESWRAAFLETSKQNSVGDIFHOLL 163

Sbjct: 7623807 LVGNKIDLHQERTVSTEEGKLAESWRAAFLETSKQNSVGDIFHOLL 7623953

The query coordinates (55 – 163) and subject coordinates (7,623,630 – 7,623,953) are shown in red and blue, respectively, while the E-value and Percent Identity are in green.

12. Fill in Table 1 with the appropriate results for each range using the information from your *tblastn* search.

D. yakuba shows 12 ranges listed in ascending order. Each range corresponds to a contiguous portion of the subject sequence (i.e., *D. yakuba* genome, chr 3R) that shows significant sequence similarity (i.e., low E values) with a portion of the query sequence (i.e., *D. melanogaster* Rheb-PA amino acid chain).

Answer question B5.

13. Go over [Appendix 3](#), and use the information there and from Table 1 in your worksheet to hypothesize the orthologous region of your target gene.

14. If you are not sure you have found the best match, or that other hits are potentially also good alignments, see [Appendix 4](#) for details on investigating the other *tblastn* alignments to *D. yakuba* chromosome 3R scaffold.

Answer questions B6-B8.

At this point, you should have a hypothesis for where the ortholog of your gene is in your target species, and several lines of evidence supporting the region you have chosen. However, similarity, lack of other good matches, and similar gene structure are not enough. Part C describes how you can use similarity across the entire genomic neighborhood to further support your hypothesis (or refute your hypothesis!).

DISCUSS and CHECK-IN: Make sure your group is in agreement on your hypothesis (worksheet question B7), and check with your instructor or TA to make sure the hypothesis is appropriate before moving on.

C. Examine genomic neighborhood of putative ortholog in target species

In Part A, we sketched the genomic neighborhood of *Rheb* in *D. melanogaster*. Here we will examine the genomic neighborhood of *Rheb* in *D. yakuba* and then compare the order and orientation of these genes to what we found in *D. melanogaster*.

Based on parsimony, the genes surrounding *Rheb* in *D. yakuba* should be identical or very similar in function to the genes in the genomic neighborhood of *Rheb* in *D. melanogaster*. Additionally, the neighboring genes should also match in orientation (look at the direction of transcription of the neighboring genes). Observing matching genes of matching orientation is called synteny, and it is one of the strongest pieces of evidence for our hypothesis in Part B that we have found the orthologous gene in our target species. For example, since *Rheb* and *CG2926* are on the positive and negative strand, respectively, in *D. melanogaster*, these two genes should also be on different strands in *D. yakuba*.

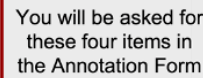
Instructions

1. Navigate to the GEP UCSC Genome Browser Gateway Page (http://tiny.cc/GEP_BrowserGateway).

- Click on your target species in the “REPRESENTED SPECIES” table.
- Ensure that you have an assembly field related to your target species, and then confirm that the assembly version you copied into the worksheet for B2 is selected.
- Enter the accession number under the “Position/Search Term” field and the approximate coordinates you recorded to examine the region you found in the *tblastn* search (from worksheet, B6).
- Click on the “Go” button.

Comments

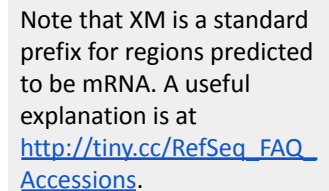
e.g. For the “D. yakuba Assembly” field, I would confirm that “Aug. 2021 (Princeton Prin_Dyak_Tai18E2_2.1/DyakRefSeq3)” is selected.



3. Zoom out 3x.

DISCUSS and CHECK-IN: Why do we need to go through and do manual annotation? Why can't we just use the positions provided in one or more of the gene model predictions?

XM_039375862 is the accession number for the *D. yakuba* RefSeq mRNA transcript that is aligned to this region of the *D. yakuba* chr3R scaffold by BLAT.



The genomic neighborhood includes the nearest two upstream genes and nearest two downstream genes.

If you do not see a comparative genomics track, check with your instructor or TA.

8. Take a screenshot of your genome browser with the genomic neighborhood of your target gene for your target species. To do so, follow these directions:

- Right click somewhere inside the Genome Browser image
- Select 'View image'
- A new tab should open with the image (if not, make sure your browser has pop-ups enabled)
- Right-click on the image in the new browser tab and click 'Save Image As...'
- Save the file into a folder for the Pathways Project on your computer, preferably as:
 <TeamName>_<GeneName>_<TargetSpeciesSymbol>_GenomicNeighborhood.png

<Target Species Symbol> for *D. yakuba* would be dyak, so example file would be: TeamAwesome_Rheb_dyak_GenomicNeighborhood.png

Answer question C2.

9. Reset your tracks to 'default tracks'. We aren't concerned with the comparative genomics track for now, but if you develop a report to submit to the GEP, this image needs to include the comparative genomics track for further review.

Interlude: In the next steps, you will use synteny to gather additional evidence for the ortholog assignment. In practice, this means comparing each predicted gene (target and those in neighboring region) in the “BLAT Alignments of NCBI RefSeq Genes” track to the *D. melanogaster* genome, and checking if the best hit is a match to the corresponding neighboring gene determined in Part A. Figures show results for the putative ortholog of the *Rheb* gene in *D. yakuba*.

Instructions

10. In the Genome Browser image, click on the label for the BLAT alignment that is likely your putative ortholog.



11. Make sure your cursor is inside the white box. From there, scroll to the bottom of the GenBank Record window to the “translation” sequence within the “CDS” section.

NCBI RefSeq Genes (XM_039375862)

GenBank Record: [XM_039375862](#)

gene prediction method: omomom. Supporting evidence includes similarity to: 11 Proteins, and 100% coverage of the annotated genomic feature by RNAseq alignments, including 116 samples with support for all annotated introns

[CDS](#)

/db_xref="GeneID:6537476"
 300..848
 /gene="LOC6537476"
 /codon_start=1
 /product="GTP-binding protein Rheb homolog"
 /protein_id="XP_039231796.1"
 /db_xref="GeneID:6537476"

Sequence of translated protein from predicted mRNA

[polyA_site](#)

ORIGIN
 1 ttgcacatct tcgacagcag cactgactca acttgagaat tactgttttc ttttagagga

Comments

For *D. yakuba* and the putative *Rheb* ortholog, that is XM_039375862

The Genbank record is an entry in the database Genbank, which is an annotated collection of all publicly available DNA sequences. There are links to much additional information tied to this sequence here.

The red arrow indicates the computationally predicted protein sequence for the putative ortholog. Because this is a predicted *protein* sequence, the accession says 'XP' instead of 'XM'.

12. Copy the accession number for the translated protein sequence (red arrow from figure in previous step). You will need this for a BLAST search in step C14.

13. Take the time now to input the XM and XP accessions into the appropriate rows in Table 2 of your worksheet. You can also copy the strand information from the information in the yellow region below the GenBank record.

Remember strand information corresponds to the direction of transcription.

Answer question C3b in Table 2 for the 'Target Gene' column.

14. Navigate to NCBI BLAST (<http://tiny.cc/blastpage>) and click on the "Protein BLAST" button to do a *blastp* search.

DISCUSS: As a group, check that everyone understands how a *blastp* search differs from a *tblastn* or *blastx* search.

15. In the input page, do the following:

- Paste the accession number for the translated protein sequence you copied into the "Enter Query Sequence" text box.
- Check the "Job Title" text box automatically populated.
- Under "Choose Search Set," select "Reference proteins (refseq_protein)" as the "Database" to search.
- In the "Organism" text box, enter "**Drosophila melanogaster (taxid:7227)**".
- Select the check box next to "Show results in a new window".
- Click on the "BLAST" button.

For *D. yakuba*, the translated protein sequence corresponding to XM_039375862 is XP_039231796.

Standard Protein BLAST

BLAST programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

XP_039231796

Or, upload file [Browse...](#) No file selected.

Job Title

XP_039231796:GTP-binding protein Rheb homolog...

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Reference proteins (refseq_protein) [?](#)

Organism

Drosophila melanogaster (taxid:7227) [Add organism](#)

Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be

Exclude

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WPI) ☐ [?](#)

Optional

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated B

Choose a BLAST algorithm [?](#)

BLAST [Search database refseq_protein using Blastp \(protein-protein BLAST\)](#)

☒ Show results in a new window

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

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

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)

New col Descri
Click 'Sele Columns'.

16. Examine the best hits and determine if there is an isoform with the highest similarity to your predicted protein sequence.

Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download			New Select columns					
Show			100					
<input checked="" type="checkbox"/> select all 53 sequences selected								
GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer								
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Ras homolog enriched in brain, isoform A [Drosophila melanogaster]	Drosophila melanogaster	364	364	100%	6e-131	97.25%	182	NP_730950.2
<input checked="" type="checkbox"/> Rap1 GTPase, isoform B [Drosophila melanogaster]	Drosophila melanogaster	136	136	87%	5e-41	43.75%	184	NP_001189023.1
<input checked="" type="checkbox"/> Ras-associated protein 2-like, isoform A [Drosophila melanogaster]	Drosophila melanogaster	116	116	90%	5e-33	34.55%	182	NP_477402.1
<input checked="" type="checkbox"/> Ras oncogene at 85D [Drosophila melanogaster]	Drosophila melanogaster	110	110	85%	9e-31	38.06%	189	NP_476899.1
<input checked="" type="checkbox"/> Ras-like protein A, isoform B [Drosophila melanogaster]	Drosophila melanogaster	109	109	85%	3e-30	34.62%	197	NP_726881.1
<input checked="" type="checkbox"/> Ras-like protein A, isoform C [Drosophila melanogaster]	Drosophila melanogaster	108	108	85%	6e-30	34.62%	201	NP_525063.1
<input checked="" type="checkbox"/> Ras-related protein interacting with calmodulin, isoform B [Drosophila melanogaster]	Drosophila melanogaster	107	107	85%	6e-29	33.97%	264	NP_001163165.1
<input checked="" type="checkbox"/> Ras oncogene at 64B [Drosophila melanogaster]	Drosophila melanogaster	101	101	81%	3e-27	33.56%	192	NP_523917.2

The best *blastp* match to the putative ortholog of Rheb-PA in *D. yakuba* is “Ras homolog enriched in brain, isoform A [Drosophila melanogaster]” (Accession: NP_730950.2) with an E-value of 6e-131 and a sequence identity of 97.25%.

17. Take a moment now to take a screenshot of your entire *blastp* search (i.e. Descriptions and your search settings above Descriptions. Copy the screenshot into your lab notebook, with the label for the neighboring gene near your screenshot. I suggest putting a label of “Pathways C17 *blastp* search comparing putative ortholog of <insert target gene symbol> in <target species> to *D. melanogaster* reference proteins”

18. When you are certain of the best *blastp* hit, input the needed information into the appropriate rows of the table in Table 2A of your worksheet.

19. Repeat steps 10-18, but for each of the two neighboring genes on both sides of your putative target ortholog (choosing the one with the longest CDS sections if >1 isoform). Also repeat steps if you have a nested gene (i.e. a gene that overlaps with your target gene).

DISCUSS: Nested genes aren't common, but if you have one, make sure to let your instructor or TA know!

Target Species	NCBI RefSeq Gene (mRNA) Accession	XM_015192882	XM_002097995	XM_039375862	XM_002097997	XM_002097998
	NCBI RefSeq Protein Accession	XP_015048368	XP_002098031	XP_039231796	XP_002098033	XP_002098034
	Strand (+/-)	+	-	+	+	-
Best blastp Result	Accession	NP_649551	NP_649552	NP_730950	NP_730954	NP_649554
	<i>D. melanogaster</i> Gene Symbol	CG12746	CG2931	Rheb	CRMP	CG2926
	E-Value	0.0	0.0	6e-131	0.0	0.0
	Percent Identity	84.30%	96.72%	97.25%	99.66%	89.28%

To find the gene symbol corresponding to the description of the gene, consider searching the term in FlyBase (<https://flybase.org/>).

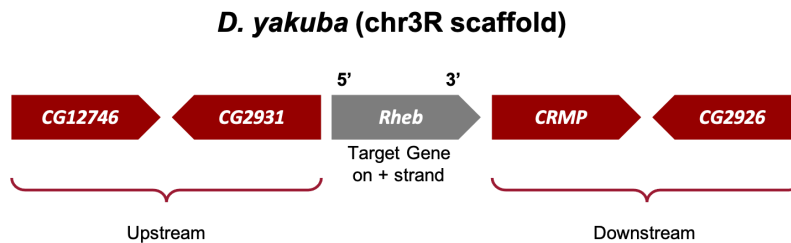
A search of ‘Ras homolog enriched in brain’ indicates that this refers to the *Rheb* gene. Using the principle of parsimony, the *blastp* search result supports the hypothesis that XP_039231796.1 is the putative ortholog of the Rheb protein in *D. yakuba*.

Since BLAST searches take time, this will help us troubleshoot faster later on if we have any problems with our chosen best hit.

Answer question C3c in Table 2 for the ‘Target Gene’ column.

Answer questions C3b and C3c in Table 2 for all other columns (‘put n/a as appropriate for that gene if it does not exist’)

20. Draw a sketch of the genomic neighborhood for your target species. You may want to confirm strand information/orientation of each gene by zooming into each gene and checking the direction of transcription.



Note that if you have found a best hit, you can draw the sketch with the gene symbol of the best hit, rather than the XM or XP accession.

Answer question C4.

21. Return to your sketch of the genomic region in *D. melanogaster* from question A5 on your worksheet, and use the information there to fill in the two remaining rows of Table 2 in your worksheet. Then, use that information and your information from the best *blastp* result to determine if your genes in the two species are orthologs.

- Put 'yes' if the gene symbol for *D. melanogaster* in Table 2B matches the *D. melanogaster* gene symbol matched for the predicted transcript of your target species in Table 2A
- Put 'no' if that column did not match.

Answer question C3a and C3d for all columns.

22. Read [Appendix 6](#), and use the logic for *Rheb* and *D. yakuba* (and for *D. pseudoobscura*) to help you address whether you have located the correct genomic neighborhood in your target species and therefore can annotate the ortholog to the *D. melanogaster* gene for next week's lab.

Answer question C5.

DISCUSS and CHECK-IN: Which genes do NOT show results supporting a hypothesis where the chosen region in *D. pseudoobscura* is orthologous to *Rheb* in *D. melanogaster*, because synteny is not observed? (Note you should specify the gene(s) in *D. pseudoobscura*, and note whether they are upstream or downstream genes.)

At this point, you should now have strong support for whether you have found an orthologous region to your target gene in your target species. The evidence you found here will be an important part of your Results section. Next time, you will complete the remainder of this lab, which is to take your orthologous region and annotate the coding sequences, i.e. determine the positions of the start and stop codons, as well as the positions indicating the transitions from exons to introns and vice versa. The last step will be to analyze figures comparing your annotated target gene in your target species to the target gene in *D. melanogaster* and describe the differences and similarities in the two orthologous genes.