Pathways Project: Project Details Tables

This document will show you how to fill out the Project Details Table on the first page of the Annotation Report Form for the Drosophila Pathways Project.

The Project details are important as they inform the researcher looking at your model where it came from, what assembly, etc… Think of it like taking your phone to the store for a software fix – without knowing what version of software, what apps you were using, and what version of apps you were using are, the tech expert can do little to solve your problem. For the same reason, we require you to have this information.

We will briefly go over what each piece of information is, as well as how to find it. Sometimes, there are going to be different ways of finding this piece of information, but we will only be covering one.

Sample

Project Species (e.g., *D. yakuba*)	D. yakuba
NCBI Taxonomy ID (e.g., 7245)	7245
NCBI Assembly ID (e.g., dyak_caf1)	dyak_caf1
Assembly Accession (e.g., GCA_000005975.1)	GCA_000005975.1
Genome Assembly (e.g., May 2011 (WUGSC dyak_caf1/DyakCAF1))	May 2011 (WUGSC dyak_caf1/DyakCAF1) Assembly
Scaffold Name (e.g., chr3R)	chr3R
Scaffold Accession (e.g., CM000160.2)	CM000160.2
Gene ID in target species (e.g., dyak_Rheb)	dyak_Rheb

Gene ID in *D. melanogaster* (e.g. dmel_Rheb)	dmel_Rheb
Accession Number of Ortholog in *D. melanogaster* (e.g., NT_033777)	NT_033777
Chromosome of Ortholog in *D. melanogaster*	3R

Date of Submission (YYYY-MM-DD)	2020-08-12

Table 1: Sample showing the filled out Project Details for *Rheb* in *Drosphila yakuba*

Location Overview

Cell	Where to find Information
Project Species	From your instructor
NCBI Taxonomy ID	From GEP's UCSC Genome Browser
NCBI Assembly ID	From GEP's UCSC Genome Browser
Assembly Accession	From GEP's UCSC Genome Browser
Genome Assembly	From GEP's UCSC Genome Browser
Scaffold Name	From GEP's UCSC Genome Browser
Scaffold Accession	*BLAST* from *DMEL* to target species
Gene ID in target species	Formatted as shown in Table 1
Gene ID in *D. melanogaster* (*DMEL*)	Formatted as shown in Table 1
Accession number of Ortholog in DMEL	From FlyBase
Chromosome of Ortholog in *DMEL*	From FlyBase
Date of Submission	In format (YEAR-MONTH-DAY)

Table 2: Location Overview of Date to be Filled Out

Table 2 shows the broad locations of where each piece of data can be retrieved. Once you have done one or two annotations in this format, you should be comfortable in only using Table 2.

However, for your first round or two, you will need to use the Detailed Instructions shown below.

Detailed Instructions

Project Species
Your faculty mentor/instructor should give you a minimum of two pieces of information when they assign you a model to annotate: (1) the species you are annotating in, and (2) the gene you will be annotation. This is that species. Please note that *D. melanogaster* can never be the project species as this is the reference species, i.e., the species you will compare your model to.

NCBI Taxonomy ID
When you go to GEP's UCSC Genome Browser, click on your species in the "**Represented Species**" section on the left. Then, click on the species name as shown in Figure 1. This takes you to another page, with the Taxonomy ID on the top-left corner of the page, as shown in Figure 1.

NCBI Assembly ID & Assembly Accession
Both of these can be found directory on the GEP's UCSC Genome Browser for your target species, and can be seen in Figure 1.
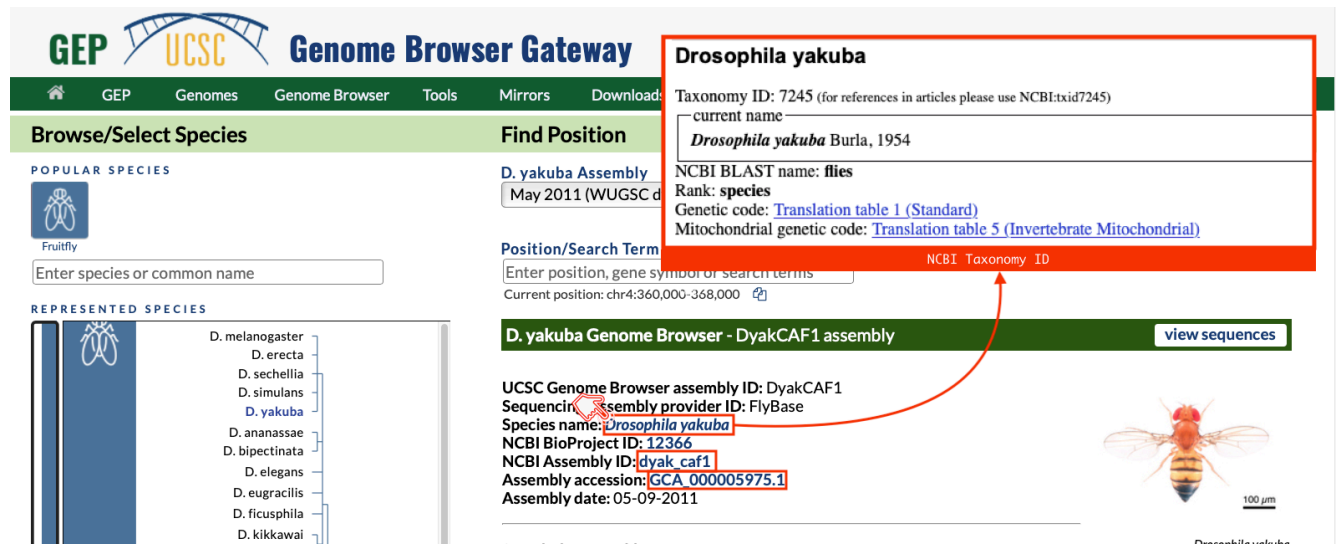
Figure 1: GEP's UCSC Genome Browser for *D. yakuba*

Genome Assembly

On the Pathways Project Genome Assemblies page, look for your species; the genome assembly is the text within the genome browser. For example, the Genome Assembly we use for *D. yakuba* is the "May 2011 (WUGSC dyak_caf1/DyakCAF1)" assembly.

Scaffold Name

When you go to your gene in GEP's UCSC Genome Browser within your target species as shown in Pathways Annotation Walkthrough, you should be greeted with something similar to Figure 2. On the top-left, you should see your scaffold name, in this case, **chr3R**.
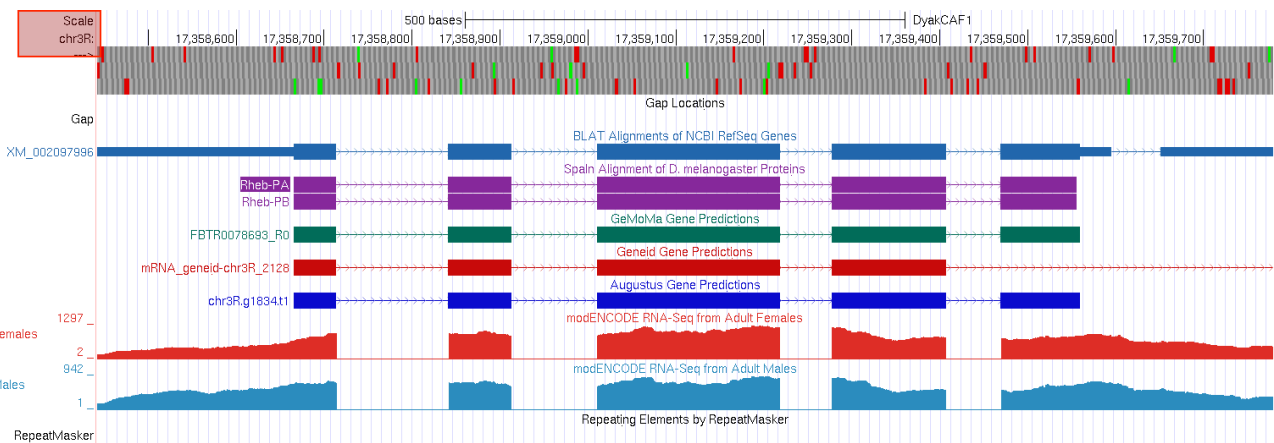
Figure 2: GEP's USCS Genome Browser for *Rheb* in *D. yakuba*

Scaffold Accession

The Scaffold Accession of chr3R can be retrieved through *tblastn*. Go to the Pathways Project Genome Assemblies page and click on the "**Genome BLAST**" link to your corresponding species. Then, when you run *tblastn* on the protein sequence of your gene of interest, the scaffold accession should be present on the right of your hit (Figure 3), in our case, being **CM_000160.2**.

Figure 3: Description pane by running *tblastn* of *Rheb* in our *D. yakuba* assembly

Gene ID in Target Species and *D. melanogaster*

This is going to always be in the format **species_gene**. Short forms of species can be seen in Table 3. Please note also that the names of genes are case sensitive, so for example, *tor* and *Tor* are two different genes.

D. melanogaster	dmel	D. elegans	dele	D. suzukii	dsuz	D. busckii	dbus
D. erecta	dere	D. eugracilis	deug	D. takahashii	dtak	D. hydei	dhyd
D. sechellia	dsec	D. ficusphila	dfic	D. obscura	dobs	D. arizonae	dari
D. simulans	dsim	D. kikkawai	dkik	D. miranda	dmir	D. mojavensis	dmoj
D. yakuba	dyak	D. serrata	dser	D. persimilis	dper	D. navojoa	dnav
D. ananassae	dana	D. rhopaloa	drho	D. pseudoobscura	dpse	D. virilis	dvir
D. bipectinata	dbip	D. biarmipes	dbia	D. willistoni	dwil	D. grimshawi	dgri

Table 3: Species Short Forms

Accession Number and Chromosome of Ortholog in *D. melanogaster*

Both of these can be found on FlyBase when you search for your gene. Search for your gene in the **Jump to Gene** (**J2G**) search bar on the top right (remember that gene names are case-sensitive). Near the top of the page, within the "Genomic Location" section, you can find the Chromosome and it's accession number in *D. melanogaster* as shown in Figure 4.

Figure 4: Chromosome and Its Accession in *D. melanogaster*