Explore the Genome Browsers created by G-OnRamp

(Answer key)

Yating Liu

06/2018

# 1. Introduction

This exercise will explore the *Drosophila miranda* UCSC Assembly Hub created by G-OnRamp in order to illustrate how you can use the Genome Browsers to address interesting biological questions.

# 2. Use Galaxy to explore the genome assembly

Open a new web browser window and navigate to the G-OnRamp server at <http://cloud5.galaxyproject.org/>. (Note that this instance will only be available during the G-OnRamp workshop.) Log into your account and then import the “Drosophila miranda MSH22 UCSC Genome Browser” History from “Shared Data”.

**Q1. Which dataset in the History contains the sequences of the *D. miranda* whole genome assembly? How many scaffolds are in the *D.* *miranda* whole genome assembly?**

The dataset “3: GCF\_000269505.1\_DroMir\_2.2\_genomic.fna” contains the genomic sequences for the *D. miranda* assembly. There are six scaffolds in the *D. miranda* assembly.

**Q2. Which scaffold is the largest in the *D. miranda* assembly? How long is the largest scaffold?**

The largest scaffold is scaffold\_3. It consists of 33,007,066 bases.

Solution 1: run the **twoBitInfo** tool on the dataset “15: faToTwoBit on data 8”. Select the “Sequence lengths” option for the “Type of output file” field.

Solution 2: Click on “Display at Track Hub UCSC main” link, then click on “view sequences” button on UCSC Genome Browser Gateway page (Figure 1, Figure 2).

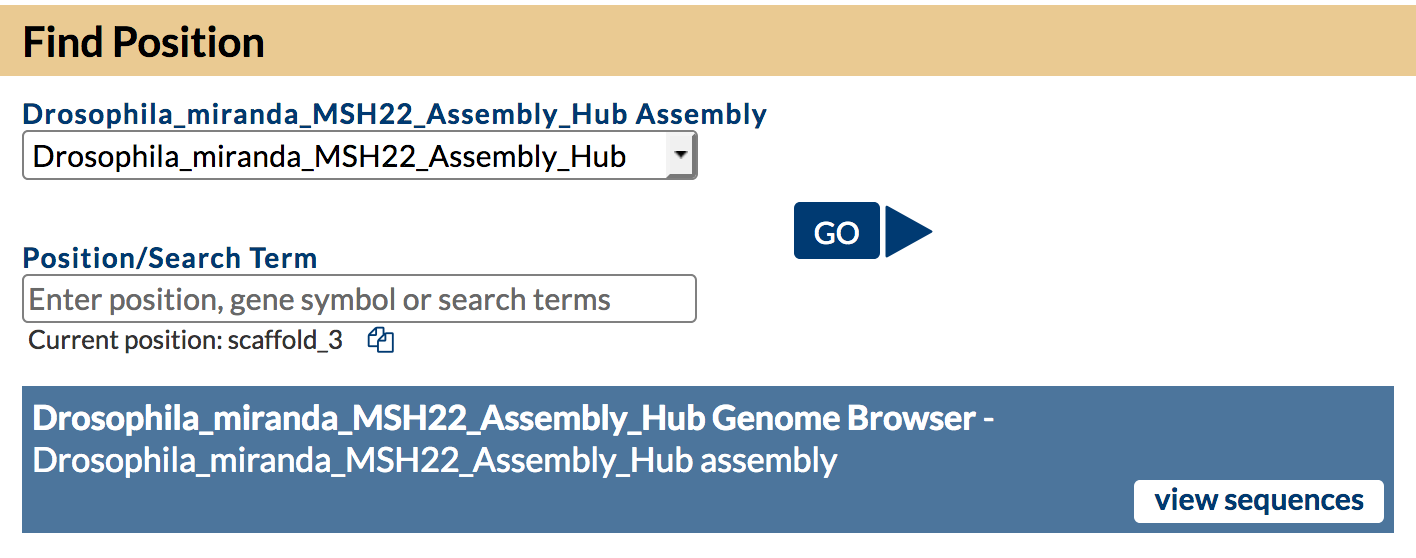


Figure 1: Click on "view sequences" button to get the information of each sequence

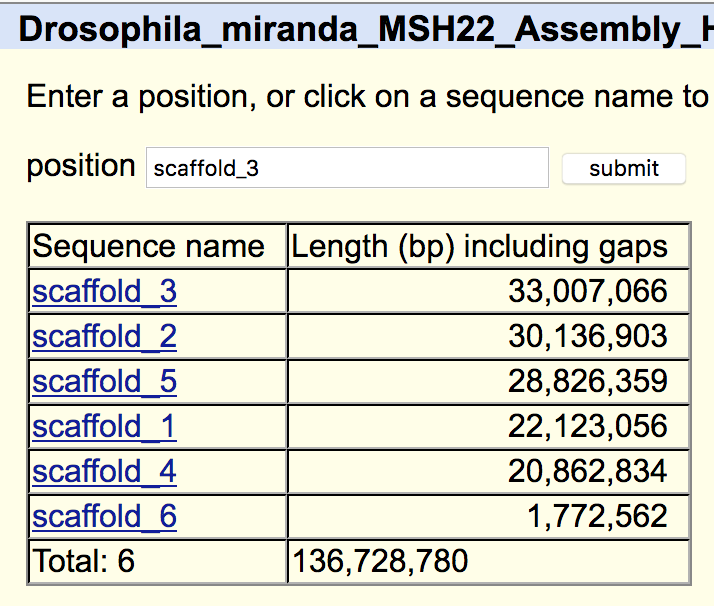


Figure 2: The table for each sequence, including the total number of sequences, sequence name, length and link for each sequence

**Q3. How many gaps are in the assembly? How many gaps are in each scaffold?**

[Hint: You can use the **Group** tool to group the gap items by column (*e.g.*, scaffold)]

There are 7,148 gaps in the assembly.

The number of gaps on each scaffold are as follow:

|  |  |
| --- | --- |
| **Scaffold** | **# Gaps** |
| scaffold\_1 | 1253 |
| scaffold\_2 | 1583 |
| scaffold\_3 | 1313 |
| scaffold\_4 | 1118 |
| scaffold\_5 | 1643 |
| scaffold\_6 | 238 |

# 3. Use the UCSC Assembly Hub to explore a genomic region

Open the *D. miranda* UCSC Genome Browser Assembly Hub. Navigate to position scaffold\_6:789,800-806,800. Scroll down to the track configuration section and hide the HISAT RNA-Seq alignment tracks (i.e., “SRR364798 HISAT S” and “SRR364800 HISAT S”) in order to simplify the display.

**Q4. How many Augustus gene predictions are in this region? How many of these Augustus gene predictions overlap with features in either the BLAST or BLAT alignment tracks?**

There are three Augustus gene predictions (scaffold\_6.g14314.t1, scaffold\_6.g14315.t1, and scaffold\_6.g14316.t1) in this region. Only the Augustus gene prediction scaffold\_6.g14316.t1 overlaps with the BLAST and BLAT alignments.

**Q5. For the Augustus gene predictions that do not overlap with features in the TBLASTN or BLAT alignment tracks, perform a NCBI BLASTP search of the translated protein sequence against the “Reference proteins (refseq\_proteins)” database. Based on the BLASTP search results, do these predictions correspond to protein-coding genes in *Drosophila*?**

(Hint: To obtain the protein sequence of a predicted gene, click on the feature in the Genome Browser image, and then click on the “Translated Protein” link. The NCBI BLAST web server is available at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.)

Most of the significant matches to the Augustus predictions scaffold\_6.g14314.t1 and scaffold\_6.g14315.t1 are to retrotransposons. The scaffold\_6.g14314.t1 prediction contains two conserved domains (R1-I-EN) and (RT\_nLTR\_like) that are often found in LINE retrotransposons (Figure 3). The scaffold\_6.g14315.t1 prediction also contains a RT\_nLTR\_like conserved domain, and it shows significant matches to Jockey retrotransposons in many different insects (Figure 4). Hence these gene predictions are unlikely to correspond to protein-coding genes in *D. miranda*.

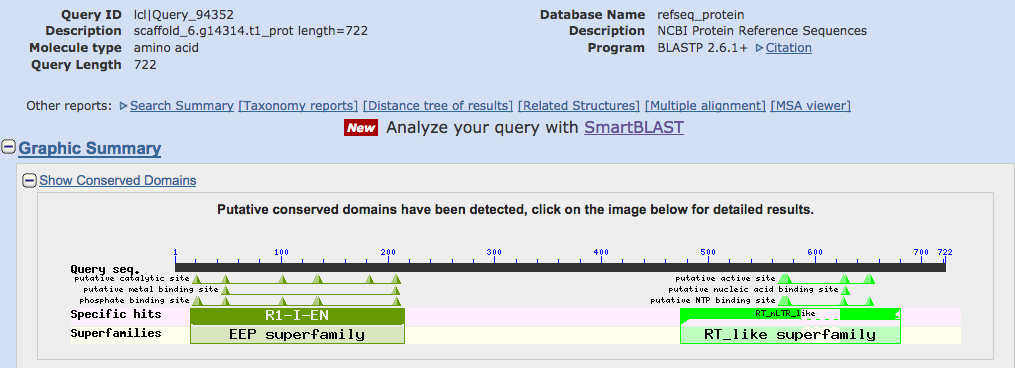


Figure 3. The Augustus gene prediction scaffold\_6.g14314.t1 contains the R1-I-EN (endonuclease domain encoded by R1 and I-element retrotransposons) and the RT\_nLTR\_like (non-LTR reverse transcriptase) conserved domains.

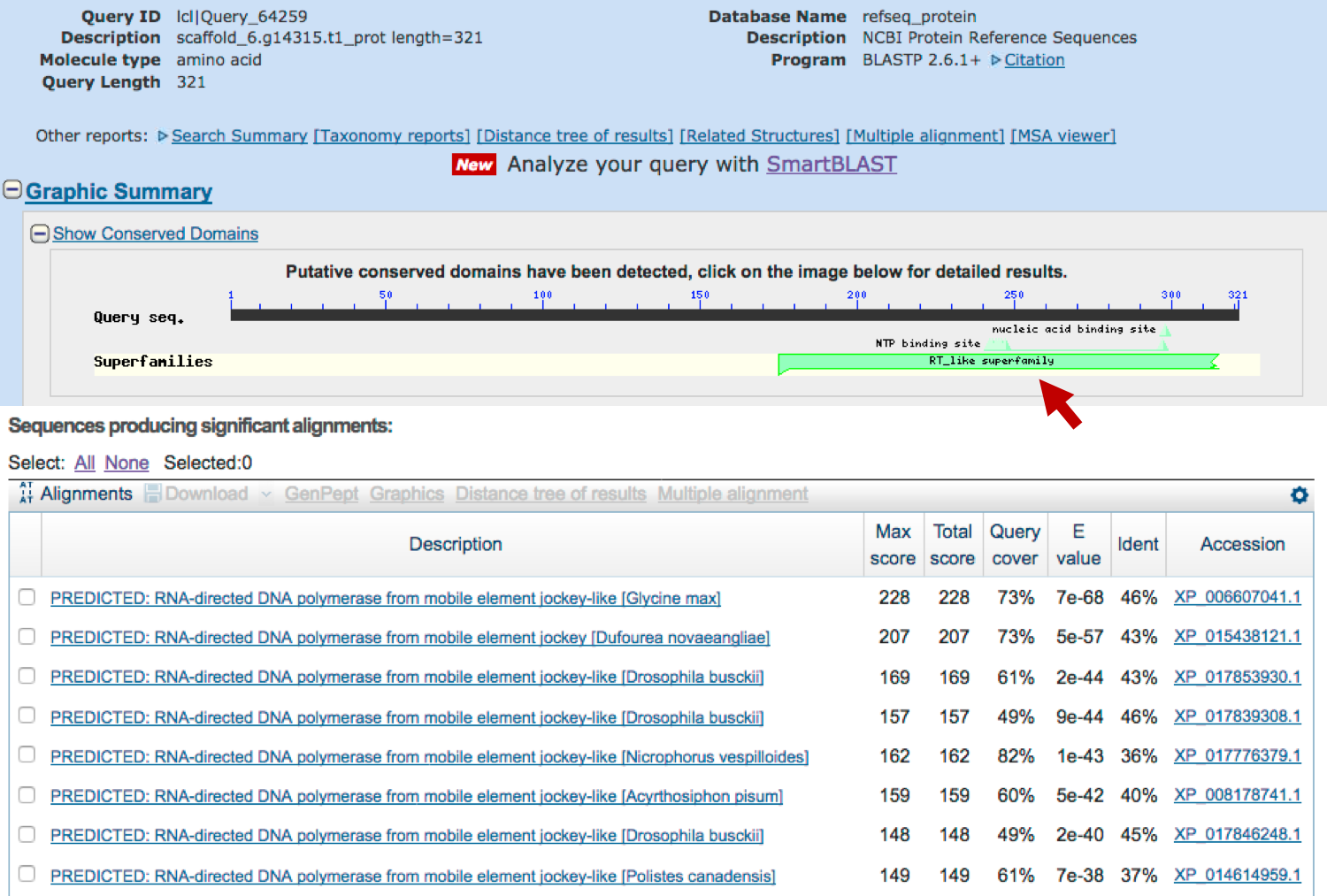


Figure 4. A BLASTP search of the Augustus gene prediction scaffold\_6.g14315.t1 against the Reference proteins database shows that the end of the predicted protein contains a RT\_nLTR\_like conserved domain (red arrow), and the prediction shows significant similarity to the Jockey LINE retrotransposons in other insects.

Change the display mode of the “BLAT Alignment” track to “pack”. When you click on a feature in the “BLAT Alignment” track, a details page will appear where you can view the GenBank record and the transcript alignment.

**Q6. Change the display mode of the BLAT alignment track to “full”. Click on the BLAT alignment to the transcript NM\_143661. What is the name of this transcript? How long is this transcript? Scroll down to the “FEATURES” section of the GenBank report for NM\_143661. Where is the coding region within this transcript?** (Hint: see the CDS section of the FEATURES table.)

The name of this transcript is “*Drosophila melanogaster* *Rad23*, transcript variant A (*Rad23*), mRNA”. This transcript has a total length of 1489 bp. The coding region is located at positions 132–1376 of the transcript (Figure 5).

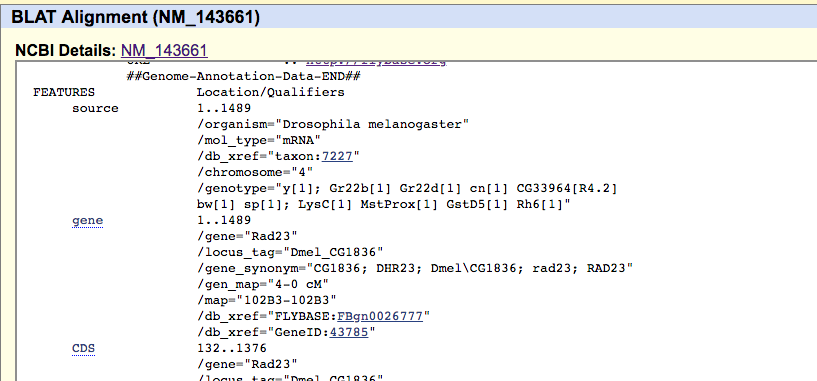


Figure 5: Scroll down to the “FEATURES” section of the GenBank report for NM\_143661 to get the CDS region

**Q7. Scroll down to the “Genomic Alignments” section of the BLAT Alignment details page. What is the percent identity between the NM\_143661 and scaffold\_6? What is the orientation of the transcript alignment? Does the alignment include the entire length of the NM\_143661 transcript?**

The alignment between the NM\_143661 and scaffold\_6 has 78.2% percent identity. The transcript and scaffold\_6 are in the same orientation (i.e., plus strand). The alignment only covers part of the transcript (positions 198–1397).

**Q8. Click on the alignment statistics link (next to the “browser” link) under the “Genomic Alignments” section to view the transcript alignment. What do the capital blue and red letters within the cDNA and genomic sequences symbolize?**

The capital blue letters correspond to matching bases between the coding regions of the NM\_143661 transcript and the scaffold\_6 sequence. The capital red letters correspond to matching bases between the untranslated regions of the NM\_143661 transcript and the scaffold\_6 sequence.

**Q9. Go back to the Genome Browser view of the region at scaffold\_6:789,800-806,800. There is another set of BLAT and TBLASTN alignments at 804,000–805,500 that is also supported by the RNA-Seq data from the virgin males (SRR364798) and virgin females (SRR264800) samples. Based on the RNA-Seq read coverage tracks (“SRR364798 Sequence Coverage” and “SRR364800 Sequence Coverage”) and the regtools splice junction tracks (“SRR364798 Splice Junctions” and “SRR364800 Splice Junctions”), how many introns does this feature have?**

The feature in this region likely has two introns (three transcribed exons).

**Q10. Using the BLAT alignments and the procedure described above, characterize the feature located at 804,000-805,500 of scaffold\_6.**

The GenBank records for the BLAT alignments NM\_205872, NM\_001144383, NM\_143662, and NM\_001297784 show that they correspond to the B, D, E and F isoforms of the *Zip102B* gene in *D. melanogaster*, respectively.

In *D. melanogaster*, the gene *CG32850* is located next to *Zip102B*. To determine the location of the putative ortholog of *CG32850* in the *D. miranda* assembly, we can enter the accession numbers of the transcripts of *CG32850* into the “enter position or search terms” text box of the Genome Browser.

**Q11. The RefSeq accession number for the A isoform of *CG32850* is NM\_166753, and the RefSeq accession number for the B isoform of *CG32850* is NM\_001272124. Can you find the BLAT alignments to these transcripts in the *D. miranda* assembly? Are the genes *Rad23*, *Zip102B*, and *CG32850* syntenic between *D. melanogaster* and *D. miranda*?**

There are no BLAT alignments to the A isoform of *CG32850*. The BLAT alignment to the B isoform of *CG32850* is located at 809,689-816,907, which is next to the *Zip102B* ortholog in *D. miranda*. Comparison of the gene order and the relative orientations of *Rad23*, *Zip102B*, and *CG32850* shows that this region is syntenic between *D. melanogaster* and *D. miranda*.