Explore the Genome Browsers created by G-OnRamp

Yating Liu

07/2018

# 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/> and log into your account. (Note that this instance will only be available during the G-OnRamp workshop.) Click on “Shared Data” in the top menu bar and select “Histories” from the drop-down menu to see the list of shared Histories. Click on the link of the “Drosophila miranda MSH22 UCSC Genome Browser” History and then click on “Import history” on the top right corner to import the History to your Galaxy.

“**rename the scaffolds**” tool shortens the scaffold names to less than 32 characters. It outputs two datasets: 1) “9: rename the scaffolds on data 1: name mapping”, which contains the original sequence names and the sequence names after renaming; 2) “8: rename the scaffolds on data 1: renamed\_reference”, which is the renamed reference sequences we use for the following analysis (Homology, RNA-Seq, Repeat finder, and *etc.*).

**Q1. Which dataset in the History contains the sequences of the *D. miranda* whole genome assembly? Download the dataset “9: rename the scaffolds on data 1: name mapping”. How many scaffolds are in the *D.* *miranda* whole genome assembly? What are the original sequence names and what are their corresponding sequence names after renaming?**

“**faToTwoBit**” tool converts a DNA sequence file in FASTA format into a [twoBit Sequence Archive](https://genome.ucsc.edu/goldenpath/help/twoBit.html). “**twoBitInfo**” tool reports the length of each scaffold and the gap locations stored in a twoBit Sequence Archive.

**Q2. Which scaffold is the largest in the *D. miranda* assembly? How long is the largest scaffold?** (Hint: run the **twoBitInfo** tool on the dataset “15: faToTwoBit on data 8”. Select the “Sequence lengths” option for the “Type of output file” field)

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

(Hint: You can run the **Group** tool on the dataset “27: twoBitInfo on data 15: Gaps”. Select “Column: 1” for “Group by column” field to group the gaps by their scaffolds. Then click on “Insert Operation” to insert an operation. Choose “Count” for the Type and “Column: 1” for the On column to count the number of gaps in each scaffold)

# 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?**

**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>.)

**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.)

**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?**

**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?**

**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?**

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

**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*?** (Hint: refer to FlyBase for information of genes *Rad23*, *Zip102B*, and *CG32850* in*D. melanogaster*)