Comparative Gene Annotation with Galaxy / G-OnRamp Answer Key

Wilson Leung

07/2018

# Exercise: characterize the revised *D. miranda* gene model using blastp

Perform a NCBI blastp search of the predicted protein sequence Dmir\_novel\_gene1\_mRNA1 against the refseq\_protein database. Compare the blastp results with the results obtained previously using scaffold\_6.g49.t1 as the query, and answer the following questions:

1. Which protein sequence in the refseq\_protein database is the best match to Dmir\_novel\_gene1\_mRNA1?

PREDICTED: tetratricopeptide repeat protein 39B-like isoform X2 [*Drosophila miranda*] (XP\_017155608.1)

1. Are the best matches in the blastp results for Dmir\_novel\_gene1\_mRNA1 similar to those for scaffold\_6.g49.t1?

Yes. The best matches are still to the tetratricopeptide repeat protein 39B-like genes.

1. How do the E-values and alignment scores of the RefSeq protein matches differ between the two blastp searches?

The alignments have lower E-values and higher alignment scores.

1. Where is the DUF3808 conserved domain in the Dmir\_novel\_gene1\_mRNA1 sequence? How does it compare to the location and extent of the DUF3808 conserved domain in scaffold\_6.g49.t1?

The DUF3808 domain is located at 53–516 of the Dmir\_novel\_gene1\_mRNA1 sequence. The alignment covers almost the entire conserved domain (3–467) out of 467 aa.

1. Based on the multiple sequence alignment of the DUF3808 conserved domain, how many positions within the domain have an information content of at least 4.0 bits?

9 positions