

Assignment #1 - Plant DNA C-Values

The purpose of this assignment is to practice using functions from the dplyr package. The data set is from the Plant C-values Database (*MD Bennett and IJ Leitch, release 6.0, December 2012, data.kew.org/cvalues*). A C-value is the amount of unreplicated DNA in a gamete, regardless of the ploidy of the organism. **dplyr functions** should be used to answer each of the following questions; please include a sentence stating your answer in addition to showing your code:

- a) Read in the C-values dataset. How many rows and columns are in the dataset? What are the data types for each column? Be explicit. (2 marks)
- b) How many unique species of Arabidopsis are in this dataset? (1 mark)
- c) What are the minimum and maximum C-values for Angiosperms? (2 marks)
- d) What was the greatest number of C-values derived from the same paper? (1 mark)
- e) What year were the C-values for the species *apetalum* and *kraussiana* published? (2 marks)
- f) Which plant group has the largest mean C-value? Does it also have the largest standard deviation? (2 marks)
- g) What is the median value for ploidy in angiosperms? (1 mark)
- h) Make a new column called 'total_dna' that multiplies the C-value by ploidy level. Assume a ploidy level of 1 if ploidy data is missing. Write to a csv file the columns 'Genus', 'Species', 'Ploidy level' and 'total_dna'. (3 marks)

1 mark will be given according to the following rubric:

- 1.0 - code is well-documented and concise
- 0.5 - code is either well-documented or concise, but not both
- 0 - no attempt was made to document code, extra variables are created, code is difficult to read

Total: 15 marks

Submission: Each student will upload a .R file and a .csv file to Quercus. Please include your first and last name, the date of submission, and the assignment number.

Due date: 11:59pm October 2nd, 2018