Assignment 8

How to complete this assignment

- Read through the background information
- Some parts of the assignment begin with more background information pertaining to that part and more detailed instructions. **Note that this information will not be available in the Quercus quiz.**
- As you complete the questions, mark your answer to each question.
- Questions will be either:
    - multiple-choice questions that require you to provide either a single answer or to select multiple answers.
    - questions that require a short text answer
- Open the associated assignment quiz on Quercus and enter your answers to each question.
- You may only submit this quiz once, so be sure you answer all questions before submitting the quiz.

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.8/Assignment.8`

Mark breakdown

Part 1 – 5 questions – 5 marks
Part 2 – 5 questions – 5 marks
Part 3 – 9 questions – 10 marks

BACKGROUND

The IMPC is "an international effort by 21 research institutions to identify the function of every protein-coding gene in the mouse genome." (https://www.mousephenotype.org/) To achieve this the IMPC knocks out each mouse gene individually and uses a panel of hundreds of standardized tests to assess mouse phenotypes, this includes, but is not limited to: assessment of behavioural tendencies, x-rays to examine the skeletal system, blood tests, and observations of morphological differences. These can be compared to control mice to determine which phenotypes are affected by the knocked out gene.

In this assignment you will be working with data for mice with the Ap4e1 gene knocked out and associated control mice that was downloaded from https://www.mousephenotype.org/. In humans, Ap4e1 deficiency is associated with a number of phenotypes including seizures, loss of muscle tone, and a number of morphological features.

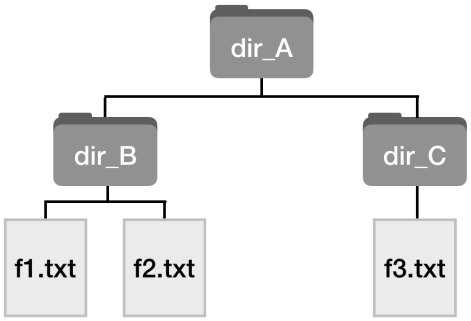
The files `Ap4e1_phenotypes.csv` and `Ap4e1_phenotypes.txt` which use you will use in parts 2 & 3 contain phenotype data for 15 Ap4e1 knockout mutants and 1470 control mice. This is the same data as was used in previous weeks. The columns in these files are as follows:

sampleID	Unique identifier of each mouse
sex	Mouse sex: female or male
sample_type	Indicates whether each mouse is a mutant or control
weight	Mouse weight in grams
hemoglobin	Hemoglobin content in g/dL
cornea_morphology	Morphology of the cornea, one of: normal, right eye abnormal, left eye abnormal, or both eyes abnormal
grip_strength	The force of strength exerted by the mouse in grams
heart_rate_bpm	The heart rate of in beats per minute (BPM)

PART 1: Managing Files & Directories (5 marks)

Set the PWD to: `/home/jovyan/Week.8/Assignment.8/part_1`

In this directory you will see a directory called `dir_A`. The structure of this directory is depicted below.



Answer the following questions pertaining to the files and directories in `dir_A`. Each question will begin by telling you the present working directory (PWD) you should be in to answer the question. Pay close attention to this as this will affect your answer. For example, compare the following two example questions:

Question X
(1 mark)
PWD: `dir_C`
Write a command to change the directory to `dir_B` using the RELATIVE path.

Answer: `cd ../dir_B`

Question Y
(1 mark)
PWD: `dir_A`
Write a command to change the directory to `dir_B` using the RELATIVE path.

Answer: `cd dir_B`

Although the question is the same, the answer differs based on the PWD.

Question 1
(1 mark)
PWD: `dir_A`
Write a command to copy `f1.txt` into `dir_C` using RELATIVE paths.

Question 2
(1 mark)
PWD: `dir_B`
Write a command to list the files in `dir_C` using the RELATIVE path.

Question 3
(1 mark)
PWD: `dir_C`
If you run the following commands:

```
mkdir dir_D
mv f1.txt dir_D/f4.txt
```

Which of the following will be FALSE?

a. `dir_D` will contain one file called `f1.txt`
b. `dir_D` will contain one file called `f4.txt`
c. `dir_C` will not contain a file called `f1.txt`
d. the result of the `ls` command run in `dir_C` will return 2 items

Question 4
(1 mark)
PWD: `dir_C`
Which of the following describes what the command `ls ../dir_B/*2*` does?

a. Outputs all the files in `dir_B` that start with the number 2
b. Outputs all the files in `dir_B` that end with the number 2
c. Outputs all the files in `dir_B` that have the number 2 anywhere in the name
d. Outputs all the files in `dir_B` that either start or end with the number 2

Question 5
(1 mark)
PWD: `dir_C`
Which command would you run to remove the directory `dir_D` and its contents (use the RELATIVE path)?

PART 2: Working with Files (5 marks)

Set the PWD to: `/home/jovyan/Week.8/Assignment.8/part_2`

In this directory you will see a file called `Ap4e1_phenotypes.csv`. The data in this file is described in the background section.

Question 6
(1 mark)
Use a command to view `Ap4e1_phenotypes.csv`. How many columns does the file have?

Question 7
(1 mark)
Write a command to get the last 8 bytes of `Ap4e1_phenotypes.csv`. What is the output?

Question 8
(1 mark)
Create a new file called `Ap4e1_phenotypes.txt` that contains all the same information as `Ap4e1_phenotypes.csv` but with tabs instead of commas. What command did you use?

Question 9
(1 mark)
Create a new file called `Ap4e1_sample_weights.txt` that contains only the `sampleID` and `weight` columns from the file `Ap4e1_phenotypes.txt`.
What command did you use?

Question 10
(1 mark)
Consider this command:

```
head -n 4 Ap4e1_phenotypes.txt | cut -c 30-50 | sed 's/a/A/g'
```

Which of the following describes what the above command does?

a. Outputs the first 4 lines of the file, removes the 30th to the 50th character from each line, and replaces all the lower case "a"s with upper case "A"s.
b. Outputs the first 4 lines of the file, removes all the columns that do not contain 30-50 characters from each line, and replaces all the lower case "a"s with upper case "A"s.
c. Outputs the first 4 lines of the file, from the 30th to the 50th character, and replaces all the lower case "a"s with upper case "A"s.
d. Outputs the first 4 lines of the file, removes all the characters up to character 30 from each line, and replaces all the lower case "a"s with upper case "A"s.

PART 3: Ap4e1 Data Analysis (10 marks)

Set the PWD to: `/home/jovyan/Week.8/Assignment.8/part_3`

In this directory you will see a file called `Ap4e1_phenotypes.txt`. This file should be the same as the file you created in part 2. The data in this file is described in the background section.

For some of these questions you will be required to complete a command that is written. Only include the portion of the command that is missing in your answer. For example:

Question Z
(1 mark)
Complete this command so that it will output the 4th and 6th column of last 8 lines of the file and redirect the output to a file called `fileZ.txt`.

```
tail -n 8 Ap4e1_phenotypes.txt | _________
```

Write the rest of the command (do not include what is already there!) in the text box.
The full command should be:

```
tail -n 8 Ap4e1_phenotypes.txt | cut -f 4,6 > fileZ.txt
```

The answer that you submit should be:

```
cut -f 4,6 > fileZ.txt
```

Hint: Remember that `tail -n +2 Ap4e1_phenotypes.txt` will return all but the first (header) line of the file.

Question 11
(1 mark)
Output a count of the total number of mice in the file.

```
tail -n +2 Ap4e1_phenotypes.txt | _________
```

Fill in the command below to match the command you used (do not include what is already there!).

Question 12
(1 mark)
Output the number of male and female mice in the dataset.
Fill in the command below to match the command you used (do not include what is already there!).

```
tail -n +2 Ap4e1_phenotypes.txt | cut -f 2 | _________
```

Question 13 (SELECT ALL THAT APPLY)
(2 marks)
Consider this regular expression:

```
[a-zA-Z]*_X_[0-9]?
```

Which of the following DO NOT match the regular expression?
(**Note:** The number of marks for this question does not necessarily reflect the number of options that should be selected.)

a. sample_X_2
b. Mouse_X_94
c. TissueSample_X_
d. _X_3
e. HEK293_X_4
f. ESC_X_
g. _X_56

Question 14
(1 mark)
Sample IDs in this file have the following format:
3 capital letters or numbers, hyphen, 5 capital letters or numbers
Ex: 81Z-CB321
Which of the following regular expressions would match sample IDs in this file?

a. `[a-z0-9]{3}-[a-z0-9]{5}`
b. `[A-Z0-9]{3}-[A-Z0-9]{5}`
c. `[A-Z0-9]?-[A-Z0-9]?`
d. `[a-z0-9]?-[a-z0-9]?`

Question 15
(1 mark)
Write a command that will output all the lines in `Ap4e1_phenotypes.txt` that DO NOT contain the word “abnormal”, along with the line numbers of the output lines.

Question 16
(1 mark)
Write a command to output the number of mice with each unique type of cornea morphology. How many mice have “both eyes abnormal”?
(**Hint:** don’t forget to use the sort command!)

Question 17
(1 mark)
Write command to find the numbers you will need to determine the percentage of mutant mice with normal cornea morphology.
What percentage of mutant mice have normal cornea morphology?

a. 30%
b. 66%
c. 50%
d. 10%

Question 18
(1 mark)
Write a command to sort `Ap4e1_phenotypes.txt` by the weight column in descending order (highest number at the top). What is the sample ID of the heaviest mouse in the dataset?

Question 19
(1 mark)
Write a command to find the sample ID of the heaviest mutant mouse in the dataset. What is the sample ID of this mouse?