8.3 File Analysis – Tutorial

At the end of this tutorial you should be able to:

- Write regular expressions
- Search for regular expressions in files
- Sort files based on specific columns
- Count occurrences of values in files
- Extract information from files by piping together commands

How to complete this tutorial

- Go through each question in order and complete any tasks that are described in the question.
- 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 quiz on Quercus and enter your answers to each question to verify that you completed the tutorial questions correctly.
- Alternatively, open the Quercus quiz when you start the tutorial and verify your answers as you complete the tutorial. **Note that there may be some information that is in this file that is not in the Quercus quiz!**
- The answers will be released at the end of the week.

Before you begin:

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.8/8.3.Files/Tutorial.8.3`
- Verify that the directory contains a file called `clinvar_data.txt`
- View the `clinvar_data.txt` file to familiarize yourself with the contents

Data Sources:

ClinVar Data was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/clinvar/)

8.3.1: Regular Expressions

Question 1
Write a command to output all the lines in `clinvar_data.txt` that contain the word "Pathogenic". What command did you use?

Question 2
Write a command to output all the lines in `clinvar_data.txt` that contain a variant that is "Pathogenic" or "Likely_pathogenic". Which regular expression would work in this command?

a. `"pathogenic+"`
b. `"[Pathogenic,Likely_pathogenic]"`
c. `"[Pp]athogenic"`
d. `"Pp*athogenic"`

Question 3
Write a command to output all the lines in `clinvar_data.txt` that contain a variant that is NOT "Pathogenic" or "Likely_pathogenic". What command did you use?

Question 4
Which of the following regular expressions would output all the lines in `clinvar_data.txt` that contain two or more consecutive capital Ps?

a. `"PP+"`
b. `"PP?"`
c. `"PP*"`
d. `"PP"`

Question 5
Write a command to output all the lines in `clinvar_data.txt` that contain a variant with the clinical significance "Other". Use the necessary option to output the line numbers along with the lines.
What command did you use?

Question 6
Genes that start with the letters "SLC" are from the solute carrier family of membrane transport proteins. Use a command to determine which disease contains variants in SLC proteins. Which disease did you identify?

a. Glaucoma
b. Type 2 Diabetes
c. Pheynlketonuria
d. Cystic Fibrosis

8.3.2: Counting Occurrences

Question 7

Create a file called `clinvar_glaucoma_data.txt`. It should contain the header line from `clinvar_data.txt` and all the lines from `clinvar_data.txt` that contain the word "glaucoma".
Sort `clinvar_glaucoma_data.txt` by chromosome (ascending) using a general numeric sort and output the first line.
What command did you use?

Question 8

You will find that the first line in the output from the previous question is the header row of the file. To ignore the header row when sorting, we can use the `tail` command with the option `-n +2` which returns every line from line 2 to the end. Pipe the output of `tail -n +2 clinvar_glaucoma_data.txt` to your `sort` command and your command for selecting the first line. What is the first line now?

a. `chromosome position disease gene clinical.significance`
b. `10 31520308 glaucoma ZEB1 Pathogenic`
c. `19 38519292 glaucoma RYR1 Uncertain`
d. `1 171635499 glaucoma MYOC Uncertain`

Question 9

To better understand why the `-g` option is important when sorting numbers, sort `clinvar_glaucoma_data.txt` again. Sort the file by chromosome (ascending), after removing the header row, and output the first line. This time do NOT use a numeric sort (don't use the `-g` option). What is the first line now?

a. `chromosome position disease gene clinical.significance`
b. `10 31520308 glaucoma ZEB1 Pathogenic`
c. `19 38519292 glaucoma RYR1 Uncertain`
d. `1 171635499 glaucoma MYOC Uncertain`

Question 10

How many lines are there in `clinvar_data.txt`? What command did you use?

Question 11
How many unique genes are there in `clinvar_data.txt`?
(**Hint:** Use `tail -n +2` so you don't count the header line, and don't forget to use the `sort` command before the `uniq` command.)

a. 43
b. 2192
c. 20
d. 44

Question 12
How many variants are there in the file for each disease?
Fill in the command below to match the command you used (do not include what is already there!).

```
cut -f 3 clinvar_data.txt | _________
```

Question 13
How many mutations are there in the gene "AKT2" for patients with Type 2 Diabetes?

a. 15
b. 24
c. 10
d. 34

Question 14
How many "Benign" or "Likely_benign" are variants are in the DMD gene?

a. 193
b. 691
c. 884
d. 2680