Homework-2 Deepika Wali

1. **Create a new directory in your home directory called homework-01. Navigate into that directory. Show your work by providing the command you used.**

1.1 Created new directory “homework-01” using mkdir command. For navigating into homework-01, I used cd command.

**Output**:

walid@20cb4a715e4b:~$ mkdir homework-01

walid@20cb4a715e4b:~$ ls

Desktop Downloads Music Pictures Projects data homework-01

walid@20cb4a715e4b:~$ cd homework-01

walid@20cb4a715e4b:~/homework-01$

1. **Using a UNIX pipe, write a command that counts how many directories are in the /home directory on your linux container on iapetus. Show your work by providing the command you used.**

2.1 Directing in my /home directory and then to count how many directories are there. I will use ls command to first list it and then pipe it through wc (word count) which also counts the character.

**Output**:

walid@20cb4a715e4b:/home$ ls | wc -l

1

walid@20cb4a715e4b:/home$ ls

walid

1. **Unix computers have a convenient file located at /usr/share/dict/words that stores all (or at least the majority) of words in the English language. What is the 55,000th word in that file? Show your work by providing the command you used.**

3.1

**Output**:

walid@20cb4a715e4b:/$ awk 'NR==55000' /usr/share/dict/words

hunk's

1. **Using a Unix command, how many visible and invisible characters are in the 55,000th word? Show your work by providing the command you used.**

**Output**:

There are total 6 visible characters and including invisible charater (\n) there are 7 characters. These can be seen by following commands

walid@20cb4a715e4b:~$ awk 'NR==55000' /usr/share/dict/words | wc -c

7

walid@20cb4a715e4b:~$ awk 'NR==55000' /usr/share/dict/words| cat -A

hunk's$

walid@20cb4a715e4b:~$ awk 'NR==55000' /usr/share/dict/words

hunk's

1. **How many GTF data lines are in this file? Note that the first few lines in the file beginning with “#” are so-called “header” lines describing thing like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.**

**Show your work by providing the command you used.**

* 1. There are2575494 gtf data lines after excluding the headers.

**Output**:

walid@20cb4a715e4b:~$ grep -v "#" human.genes.gtf | wc -l

2575494

1. **How many GTF data lines are in this file for protein coding genes? Note that entries (lines in the file for protein coding genes will contain the following text: gene\_biotype “protein\_coding”**

**Use the string above with a command you have learned to find such lines.**

**Hint: the UNIX pipe may come in handy here…**

**Show your work by providing the single command you used.**

* 1. The answer is 2352541

We used -w option along with \ to search for exact word “protein\_coding” along with the quotation mark. The result of which is passed through pipe to wc-l to show the number of lines.

**Output**:

walid@20cb4a715e4b:~$ grep -w \"protein\_coding\" human.genes.gtf | wc -l

2352541

1. **How many GTF data lines are in this file for exons from protein coding genes?**

**Show your work by providing the single command you used.**

* 1. The answer is 1911094

**Output:**

walid@20cb4a715e4b:~$ grep -w \"protein\_coding\" human.genes.gtf | grep "exon" | wc -l

1911094

1. **How many**[**CDS**](https://www.biostars.org/p/65162/)**exons (“CDS” in column 3) from protein coding genes exist on per each chromosome (column 1 in the GTF file)? That is, the count per chromsome.**

**Show your work by providing the command or commands you used.**

**8.1**

**Output:**

walid@20cb4a715e4b:~$ awk '$3 == "CDS"' human.genes.gtf | grep protein\_coding | cut -f1| sort| uniq -c| less -S | head

64952 1

26698 10

42284 11

42166 12

10435 13

24171 14

26046 15

32587 16

44869 17

11850 18

(Showing top 10 records)

1. **Explain how you might design an analysis of this file that would reveal how many distinct protein coding genes there are in the human genome.**

**Hint: you may not have learned all of the command you might need - the point is to think about what you could do with the commands you know of and what limitations would have to be addressed.**

**9.1**

**Output:** The answer is 19986

walid@20cb4a715e4b:~$ grep protein\_coding human.genes.gtf | cut -f9| cut -f1 -d ';'| sort | uniq -c | wc -l

19986

**Explanation on the 9the answer**: We first filtered the protein\_coding records from the file using the grep command. From the derived set, we then cut the 9th column which holds the gene data. We again cut the first column from the column 9 and get the gene\_ids. The best way to find the unique protein coding genes would be to find the unique gene\_ids in the derived set of data.

We then use sort and uniq -c to count the unique such records. Finally, wc -l helps to find the count of unique gene\_ids for the human genome.

The answer we finally got is 19,986 such protein coding genes from the data.

**Bonus**

1. **Using grep and the Unix words file from above, write a command that returns only five-letter words. Hint: it is okay to do some googling.**

**10.1**

Output: Keeping the apostrophe (Assuming ‘ is included in 5 letters)

walid@20cb4a715e4b:~$ cat /usr/share/dict/words | egrep '^.....$' | wc -l

6815

**10.2**

Output: Without apostrophe in the word (Assuming ‘ is not included in 5 letters)

walid@20cb4a715e4b:~$ grep -v "'" /usr/share/dict/words | egrep '^.....$' | wc -l

6138