Bioinformatics Homework #1 UNIX

# Extract information from a genome annotation file (.gtf).

**Due September 3rd**

## Prepare data

1. Go to ~/Documents/2018-Bioinfo/.
2. Create a sub-directory called homework-1
3. Obtain the gtf file by issuing the command

* wget ftp://ftp.ensembl.org/pub/release-93/gtf/homo\_sapiens/Homo\_sapiens.GRCh38.93.gtf.gz  
  gunzip Homo\_sapiens.GRCh38.93.gtf.gz # will take ~10 sec

1. Use the shell script to answer the following questions

Here is the [explanation](ftp://ftp.ensembl.org/pub/release-93/gtf/homo_sapiens/README) of the gtf file.

## Questions

1. (2 pts) How large is the gtf file you just obtained (hint: check out the -h flag in the manual for ls)  
   1.1 GB or 1144423988 bytes (both are acceptable)
2. (3 pts) What command would you use to view the first 5 lines of the file, and what did you learn about this annotation file (use natural language to explain what you understand from those 5 lines).  
   (1 pts) $ head -n 5 Homo.gtf  
   (1 pts) The genome build is GRCh38.p12, the genome was released 2013 Dec, and was last updated 2018 Jan. The accession is NCBI:GCA\_000001405.27 (anything along this line will suffice).
3. (2 pts) How many lines in this file are comments (start with a “#”)? Write down the answer and the command you used to obtain it.  
   $ grep “^#” Homo.gtf | wc -l (must have the caret sign. Double quotes are optional. The -l flag is also optional)
4. (3 pts) What command would you use to view the last 5 lines of the file? How can you highlight the word “gene\_name” in those lines and what gene name did you see?  
   $ tail -n 5 Homo.gtf | grep “gene\_name” (grep --color=auto is also correct.)
5. (2 pts) What command can you use to view the content of the file with control (namely, you can scroll up and down), and how can you make each row be displayed in one screen line, with the overflow cutoff rather than wrapped around?  
   $ less -S Homo.gtf
6. (2 pts) How can you use grep to “remove” the comment lines in order to view the actual annotations with the command above? Write down the command (hint: use pipe)  
   $ grep -v “^#” Homo.gtf | less -S
7. (2 pts) How many genes are there in this build of the human genome? (hint: the third column of the file indicates the type of the feature. Use grep to select only those that match “gene”)? Write down the answer and the command you used to obtain it.  
   $ grep -v “^#” Homo.gtf | cut -f 3 | grep “gene” | wc -l

* Now let’s pick a gene and learn a bit more about it. Please refer to the table below to determine which gene you will be looking up

1. (4 pts) Write down the command you will use to do the following: 1) select all lines that contain the gene name; 2) then select those lines from the last step that contain the word “protein\_coding”; 3) finally write the output to a file named “my\_gene.txt” and 4) report the total number of lines in this file. (hint: use grep twice).  
   58395  
   $ grep -v “^#” Homo.gtf | cut -f 3 | grep “gene” | wc -l  
    *note: I didn’t realize that this file contains multiple sources of gene annotations, some of which are duplicates. To get the answer, one should restrict the search to those lines where the second column is “ensembl\_havana”. This is beyond what we taught in the class, and the point of this question is to test the students’ ability to use the cut and grep commands. So the 58k genes answer, while wrong, is accepted. The command to use to get the actual answer is as follows:*$ awk ‘($2==”ensemble\_havana”)’ Homo.gtf | cut -f 3 | grep “gene” | wc -l
2. (4 pts) Looking only at those lines where the second column says “ensembl\_havana”, answer the following questions
   1. (1 pts) Which chromosome is this gene on? (Note where do you get this information)
   2. (1 pts) How long is this gene?
   3. (1 pts) How many exons are there?
   4. (1 pts) How many coding exons there are
   5. (bonus 2 pts) How many amino acids does this gene have? (explain how you calculated this) *note: the bonus points will be added to the total score until it reaches 24*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Last name | Gene name | Chromosome | Length | # exons | # amino acids |
| A-E | APOE | 19 |  |  |  |
| F-J | FOXO1 |  |  |  |  |
| K-Z | SOD1 |  |  |  |  |