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

Please download this homework file and provide the answers below the questions. For unix commands, start with a “$ “. Submit the finished document to ICON.

## 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

## Questions

1. (2 pts) How large is the gtf file you just obtained (hint: check out the manual of ls for a flag that makes ls print file sizes in human readable format)
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 learned about this genome annotation file from those 5 lines).
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.
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?
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 cut off rather than wrapped around?
6. (2 pts) How can you use grep to “remove” the comment lines and view the actual annotations with control (like above)? Write down the command (hint: use pipe)
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.

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 select all lines that contain the gene name; then select those lines from the last step that contain the word “protein\_coding”; finally write the output to a file named “my\_gene.txt”, and report the total number of lines in this file. (hint: use grep twice).
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?   
      *note: the bonus points will be added to the total score until it reaches 24*

|  |  |
| --- | --- |
| **Last name starts with** | **Gene name to look up** |
| A-E | APOE |
| H-J | FOXO1 |
| K-T | SOD1 |