Exercise02: Bash introduction #2

The goal of this exercise is to improve skills working with basic bash commands and awk in UNIX environment.

> Commands to learn awk, uniq, sort, file redirection with '>'

To submit your result, follow these steps:

- Step 1. Clone this template repository to your working directory and execute "setup.sh"
- Step 2. Fill in the command used in the command0X.sh in the "command" directory. The commands should generate the result of step 3. The result can either be printed to the terminal or written to a file.
- Step 3. Save the result to ./result/result0X_X.txt or ./result/result0X_X.csv for each command.
- Step 4. Add edited files to git and commit

  ```
  git add .
  git commit -m "COMMIT MESSAGE"
  ```

- Step 5. Submit your answers by pushing the cloned repository.

  ```
  git push origin master
  ```

command01.sh

1. Download the GTF file of Drosophila melanogaster and save it as **d_melanogaster.genes.gtf.gz** in the "data" directory. (No result file)

   Link: ftp://ftp.ensembl.org/pub/release-103/gtf/drosophila_melanogaster/Drosophila_melanogaster.BDGP6.32.103.gtf.gz

2. Extract the gzip-compressed GTF file. (No result file)

command02.sh

The first few lines of GTF file begin with "#". These lines are called header lines.

1. Use a command to extract only the header line from a GTF file and store the result to **result02_1.txt** in the "result" directory.

2. Count the number of lines in the GTF file except for header lines and save the number to **result02_2.txt**.

command03.sh

1. Extract unique chromosome names in "d_melanogaster.genes.gtf" and save it to **result03_1.txt**.

> You can find the structure of GTF file from this link or from the lecture slides.

2. Count genes in each chromosome and find **the chromosomes which have 100 or more genes**. Sort chromosome names alphabetically and write the **chromosome names as column 1** and the **count of genes as column 2** to **result03_2.csv**.

    > Columns in CSV file, which means Comma-Separated Values, should be separated with comma, ",". When counting the number of genes in a chromosome, count the lines of which the feature type is "gene".

command04.sh

1. Extract the distinct genomic feature types (e.g., gene, exon, transcript ...) from the GTF file. Sort the values alphabetically and save them to **result04_1.txt**.

2. Find the line in which the feature type is "gene" and the gene name is "Raf". Save the line to **result04_2.txt**.

3. The "Raf" gene has multiple transcripts. Find all transcripts and store the attribute "transcript_name" (e.g., transcript_name "Raf-RE";) to **result04_3.txt**.

    > You can use "tr" command with "-d" option for removing unwanted characters including double quotes or semicolons.

    ```
    echo '"""HELLO WORLD!!"""' | tr -d '"' # Result: HELLO WORLD!!
    # Multiple characters can be added with "|".
    echo 'gene_name "Raf";' | tr -d '"|;' # Result: gene_name Raf
    ```

4. Count the number of exons of each transcript from "Raf" gene and save the count to **result04_4.csv** (remember CSV files are comma separated). Write the **transcript names (value of transcript_name) as column 1** and the **count of exons as column 2** like this:

    ```
    Raf-XX,5
    Raf-AA,3
    ...
    ```

5. Calculate the total exon length of each transcript from "Raf" gene and save the result to **result04_5.csv**. Write the **transcript names as column 1** and the **length of exons as column 2** like this:

    ```
    Raf-XX,3300
    Raf-AA,2500
    ...
    ```

> The position of GTF is 1-based, which means the 100nt-length region from 1st position to 100th position in chromosome 1 is represented as "chr1 1 100". Please consider this when calculating the length from position indices.

command05.sh

1. Download *E. coli* GTF file and unzip it as "e_coli.genes.gtf" in data directory. (No result file)

   Link: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gtf.gz

2. Count the genes in the positive (+) and negative (-) strand of *E. coli* genome and save the counts to **result05_2.csv**.

3. Compare the distinct genomic feature types of d_melanogaster.genes.gtf and e_coli.genes.gtf. Find the features which only exist in *D. melanogaster* and not in *E. coli*. Sort the values alphabetically and write them to **result05_3.txt**.

If you can't execute a shell file due to "Permission denied" error, please try this command.

```
chmod +x ./<SOME_SHELL_FILENAME>.sh
./<SOME_SHELL_FILENAME>.sh
```