Hiring tasks 2 and 3 (2023)

Malay (mbasu@kumc.edu)

Table of contents

1 Task 2

E. coli MG1655 is the standard reference strain of *E. coli*. The protein FASTA file for this strain can be downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/NC_000913.faa. Using just `bash` commands can you find out what is the average length of protein in this strain? You may use as many commands as you may wish.

1.1 Data

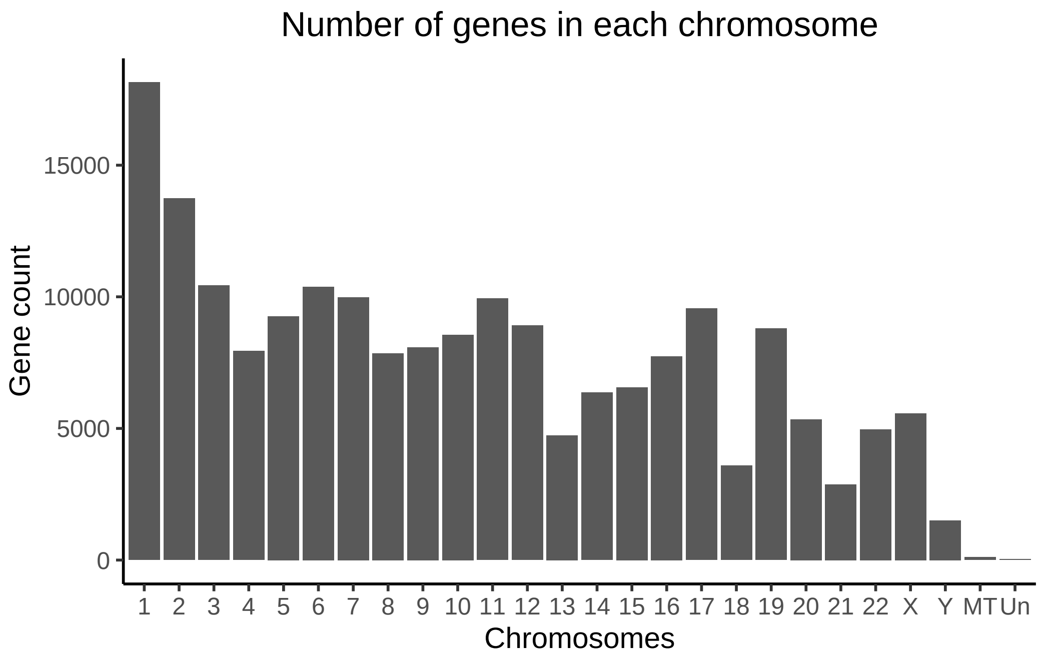
`NC_000913.faa.gz` is a amino acid FASTA file compressed using `gzip`. You download the same file from here: https://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/NC_000913.faa

1.2 Note:

1. You must use only bash commands. No other programming language is allowed. If you are using Mac or Linux, then bash automatically comes along with the machine. For windows you need to install bash. Follow the instructions here: https://superuser.com/questions/608106/how-can-i-use-a-bash-like-shell-on-windows.
2. You can either provide a single command line or a bash script.
3. You may need the following commands in bash to complete this task: `wget, zcat, wc, tr, bc, and grep`. You are not restricted to any of these commands. You can use any or all or any other bash commands in your script or command line.
4. Save your command or bash script as `task2.sh`.
5. Hint: you need to count the number of sequences in the file and the total number of amino acids then divide the latter with the former. The answer is **316**.

2 Task 3

Use `R` and `ggplot2` package draw a a plot of number of genes per chromosome in human genomes. This task requires the data file `Homo_sapiens.gene.info.gz`. You need to use columns 3 and 7 indicating `Symbol` and `chromosome` respectively. You script should create a plot exactly as shown below. Save the plot to PDF file.



2.1 Data

`Homo_sapiens.gene_info.gz` . This is a tab-delimited text file that contains information about all the genes in the human genome. If you are interested in more about this file format check here: https://ftp.ncbi.nih.gov/gene/DATA/README.

2.2 Note:

1. The figure should exactly look like the above figure.
2. There are some data in the `chromosome` column that are ambiguous and look like this: `10|19|3`. You need to discard all rows where the chromosome value contains a `|`.

3 Rules

1. You have 48 hours to finish the task.
2. All program must be a stand-alone script, runnable from the command line. No Jupyter, iPython notebook, or Rmarkdown allowed.
3. Strictly follow the submission instruction. The only way your submission is accepted is through a `GitHub` link. No other way of submission is admissible.

4 How to submit

You should create a new repository on `GitHub`, or use the existing repository that you created for the previous task. Put your program only (not any other file) into this repository and send me a direct link to the file (not the whole repository) to my email address mbasu@kumc.edu. Use reply to the email that I sent initiating this task. The email must reach me within 48 hours of the initial email. So it is better to be safe and avoid submitting your answer at the very last moment. Make sure you received a reply from me acknowledging your submission.

All the best!