**Assignment 1 – Introduction to Linux (worth 10% of MB6301)**

**Course co-ordinator: Dr Marcus Claesson (**[**m.claesson@ucc.ie**](mailto:m.claesson@ucc.ie)**)**

**Demonstrator: Ross Holohan (**[**ross.holohan@ucc.ie**](mailto:ross.holohan@ucc.ie)**)**

**Student: James Henry Hehir (120224791)**

**For each question, please record your answer in a Word document, and also what you typed on the command-line to get this answer. When finished, upload the document with your answers to blackboard under Assignments and MB6301 Assignment 1.**

**Deadline: End of Sunday the 28th Feb**

**If you cannot meet this deadline you will have to fill in and submit the Late Submission Form (see Assignment folder on blackboard) and provide any supporting documentation (e.g. medical cert).**

When no specific question is asked, perform the tasks and paste in the syntax including the full prompt, as well as the output that is not directed into a file, so that it is clear where you are located when doing the tasks.

To figure out how the command should be used, use ‘man’, ‘info’.

Remember, an important part of Bioinformatics is the ability to find information on how to solve a problem online.

Q1. Download the Homo sapiens refseq gene file from NCBI to home directory. **(0.5)**

([ftp://ftp.ncbi.nlm.nih.gov/refseq/H\_sapiens/mRNA\_Prot/human.1.protein.faa.gz](ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz))

**jimmy@jimmy-VirtualBox[jimmy] wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H\_sapiens/mRNA\_Prot/human.1.protein.faa.gz**

**--2021-02-28 13:25:54-- ftp://ftp.ncbi.nlm.nih.gov/refseq/H\_sapiens/mRNA\_Prot/human.1.protein.faa.gz**

**=> ‘human.1.protein.faa.gz’**

**Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 165.112.9.230, 165.112.9.229, 2607:f220:41e:250::11, ...**

**Connecting to ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|165.112.9.230|:21... connected.**

**Logging in as anonymous ... Logged in!**

**==> SYST ... done. ==> PWD ... done.**

**==> TYPE I ... done. ==> CWD (1) /refseq/H\_sapiens/mRNA\_Prot ... done.**

**==> SIZE human.1.protein.faa.gz ... 1264214**

**==> PASV ... done. ==> RETR human.1.protein.faa.gz ... done.**

**Length: 1264214 (1.2M) (unauthoritative)**

**100%[======================================>] 1,264,214 1.52MB/s in 0.8s**

**2021-02-28 13:25:57 (1.52 MB/s) - ‘human.1.protein.faa.gz’ saved [1264214]**

**jimmy@jimmy-VirtualBox[jimmy] ls**

**bioinf\_files Downloads Music Public**

**Desktop filenames.txt Pictures Templates**

**Documents human.1.protein.faa.gz practical\_1 Videos**

Q2. Make Homo\_sapiens directory and copy the fasta file into directory and unzip.**(1)**

**jimmy@jimmy-VirtualBox[jimmy] mkdir Homo\_sapiens**

**jimmy@jimmy-VirtualBox[jimmy] sudo cp human.1.protein.faa.gz ./Homo\_sapiens**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] gzip -dk human.1.protein.faa.gz**

Q3. Count the number of sequences in the fasta file.**(0.5)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] grep -c ">" human.1.protein.faa**

**9971**

Q4. Add your student ID number, in square brackets, to the end of all header lines and copy the first three proteins into the answer sheet. **(1)**

**sed -i '/^>/ s/$/[120224791]/g' human.1.protein.faa**

Q5. Export all description lines containing the word “olfactory” to a text document and count them. **(0.75)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] grep olfactory human.1.protein.faa > olfactory\_count.txt**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] wc -l olfactory\_count.txt**

**46 olfactory\_count.txt**

Q6. Do the same thing for the description lines containing the word “transport” and count them.**(0.75)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] grep transport human.1.protein.faa > transport\_count.txt**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] wc -l transport\_count.txt [ 2:17PM]**

**127 transport\_count.txt**

Q7. Concatenate the olfactory and transport output files redirect the output to a txt file. **(0.5)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] cat olfactory\_count.txt transport\_count.txt > overall\_count.txt**

Q8. Remove the “>” sign form the start (left) of each line, direct into a new file and copy the first 5 lines and the last 5 lines into the answer sheet. **(1.5)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] sed 's/^.\{,1\}//' overall\_count.txt > overall\_count1.txt**

**NP\_001372981.1 olfactory receptor 2T7 [Homo sapiens][120224791]**

**NP\_001013374.2 olfactory receptor 8U8 [Homo sapiens][120224791]**

**NP\_001013376.2 olfactory receptor 9G9 [Homo sapiens][120224791]**

**NP\_001372980.1 olfactory receptor 2T7 [Homo sapiens][120224791]**

**NP\_001335218.1 olfactory receptor 10AC1 [Homo sapiens][120224791]**

**XP\_006718218.1 sodium-coupled monocarboxylate transporter 2 isoform X1 [Homo sapiens][120224791]**

**XP\_016872733.1 sodium-coupled monocarboxylate transporter 2 isoform X2 [Homo sapiens][120224791]**

**XP\_011518222.1 sodium-coupled monocarboxylate transporter 2 isoform X2 [Homo sapiens][120224791]**

**XP\_006718219.1 sodium-coupled monocarboxylate transporter 2 isoform X3 [Homo sapiens][120224791]**

**XP\_011518223.1 sodium-coupled monocarboxylate transporter 2 isoform X4 [Homo sapiens][120224791]**

Q9. Use two cut commands on the same command-line, one piped after the other, to isolate only the name of the protein on each line, i.e. the name between the first space the first “[“, from the output in step 8. Note: as there may be different number of spaces you may have to use a certain range **(1.5)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] cat overall\_count1.txt | cut -f1 -d "[" | cut --complement -f1 -d " " > protein\_names.txt**

Q10. Sort the output from step 9 in reverse order and count the unique lines. Paste in the first 5 lines and the last 5 lines in the answer sheet. **(2)**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] tac protein\_names.txt > protein\_names\_reversed.txt**

**jimmy@jimmy-VirtualBox[Homo\_sapiens] wc -l protein\_names\_reversed.txt**

**173 protein\_names\_reversed.txt**