Lab_1_questions

Course: Genomics & Bioinformatics vt21
Author: Mohammed Al-Jaff

Question 1: What did you get? How many lines and words?

Ans:

From the manual instructions of the wc command. The output of wc givs back i) nr of lines ii) nr of words iii) characters and iv) file size in bytes

In the PCA.txt case we get that the file has a total of 2322 lines, 14913 words and 112376 characters (in raw file text content)

Question 2: Write down how many times PCA appears in both the full PCA.txt and the short_pca.txt files.

Ans: in the short file, 'PCA' occurs 16 times and 156 times in the pca.txt file. Note here that we used the command

```
grep -o PCA pca.txt | wc -l
```

with he '-o' flag to count number of occurrences beach without it, grep returns only all the lines that PCA occurs in, not taking into account that PCA could occur more than once in one line.

Question 3: Now that you have some basic UNIX tools at your disposal go and do the hidden word_exercise. Submit the hidden word.

Ans:

1. C
2. O
3. M
4. M
5. A
6. N
7. D
8. _
9. L
10. I
11. N
12. E

Hidden word: COMMAND-LINE

Question 4: use samtools view and head and tail to figure out the first and last position in the file. Also include the exact command you used!

Ans: finding the first line

```
samtools view SGDPDai1.dedup.realn.3mask_recal.chr2.135787850-
135887184.bam | head -1
```

finding the last line

```
    samtools view SGDPDai1.dedup.realn.3mask_recal.chr2.135787850-
135887184.bam | tail -1
```

Question 5: Write down the command you used to extract the name and nucleotide sequence!

Ans: command pipe/chain to get the relevant fields (name and sequence)

```
    samtools view SGDPDai1.dedup.realn.3mask_recal.chr2.135787850-
135887184.bam | head -1 | cut -f1,10
```

Question 6 : You have been given a that has been exported from excel in an odd format (something that is all too common in the life of a bioinformatician). Your task is to transform the file orange.csv into a normally formatted .csv-file. That is the decimal point should be a . and the delimiter (what separates one column from another) should be ,. It also looks like someone has accidentally inserted some letters among the numbers, they also need to be removed.

Ans:

```
    #!/bin/bash

# Basic substituion and transformation flow.
# 1. substitue all commas with a .
# then replace all semmicolons with comma
# replace tabs with comma
# deal with letters at end of and inside numbers.

sed 's/,/./g' orange.csv \
| sed 's/;/,/g' \
| sed 's/   /,/g' \
| sed  -E 's/[a-z]+//g' \
>> orange_juice.txt
```

The final out put resembles the desired criteria:

```
INCOM,PRICE,QUANT,YEAR,ADVERT
0.836,2.62,0.187,1910,0.1383
0.798,2.867,0.218,1911,0.4344
0.802,2.915,0.2,1912,0.3092
0.827,3.4,0.135,1913,0.1806
0.832,2.344,0.259,1914,0.557
0.83,2.534,0.247,1915,0.621
```

```
0.892,3.304,0.227,1916,0.5435
0.925,2.632,0.258,1917,0.6136
0.901,5.842,0.114,1918,0.2204
0.858,4.022,0.227,1919,0.3759
0.816,3.792,0.226,1920,0.3587
0.701,2.217,0.299,1921,0.6427
0.692,3.98,0.197,1922,0.3409
0.814,2.682,0.28,1923,0.7334
0.848,1.924,0.327,1924,0.8155
0.834,3.867,0.249,1925,0.6085
0.854,3.392,0.286,1926,0.7184
0.866,3.379,0.322,1927,0.8453
0.873,5.073,0.259,1928,0.8948
0.917,2.227,0.444,1929,1.3657
0.903,5.012,0.243,1930,0.8096
0.82,2.022,0.417,1931,1.8337
0.726,2.013,0.373,1932,1.2544
0.636,1.584,0.387,1933,1.1665
0.7,2.507,0.341,1934,1.4092
0.745,2.011,0.463,1935,1.9728
0.824,2.623,0.374,1936,1.3354
0.902,3.051,0.37,1937,1.2751
0.857,1.354,0.526,1938,2.2123
0.868,1.307,0.521,1939,1.5325
0.927,1.585,0.52,1940,2.0202
1.031,1.997,0.584,1941,2.4856
1.157,2.362,0.583,1942,2.4783
1.322,3.492,0.592,1943,1.3343
1.363,3.576,0.703,1944,1.7287
1.419,3.559,0.726,1945,1.7502
1.389,3.81,0.657,1946,2.3292
1.278,1.665,0.734,1947,2.1186
1.236,1.301,0.705,1948,1.7493
1.261,1.699,0.634,1949,1.4608
1.28,2.221,0.656,1950,1.4789
1.329,1.847,0.726,1951,1.4918
1.329,1.318,0.743,1952,2.0569
1.37,1.495,0.74,1953,2.0008
1.372,1.674,0.76,1954,1.798
1.403,1.695,0.771,1955,2.193
1.482,2.095,0.767,1956,1.9494
1.501,1.763,0.751,1957,2.0635
1.475,2.507,0.606,1958,1.9575
1.502,2.586,0.711,1959,2.7034
```