Mars 2019

Modify and extract information from large text files with sed & awk

Exercices

Pre-requisite: knowledge of UNIX environment

Data in : /home/formation/save/tp_sed_awk

For exercices, connect to **« `genologin` »** by using « `putty` » (from windows machine) or « ssh » (from linux machine).

Once connected on genologin, move to the work working directory and create a subdirectory named : **tp_sed_awk**

Link data files with a symbolic link

- ln -s /home/formation/save/tp_sed_awk/f1_R1.fq f1_R1.fq
- ln -s /home/formation/save/tp_sed_awk/f1_R2.fq f1_R2.fq
- ln -s /home/formation/save/tp_sed_awk/ncRNA_Tbarophilus.fasta
- ln -s /home/formation/save/tp_sed_awk/hg19_exons.bed

TP1 : Regular expressions

Exercice 1

What will be the result of the command « ls abc[13] ». Explain.

1. abc1 abc3
2. abc1 abc2 abc3 abc13
3. abc1 abc13 abc2
4. abc1 abc2 abc3
5. abc abc1 abc13 abc3

Exercice 2

What is the command to list all the lines of the `file` file which begin with the string « $US » ? Explain.

1. grep ^$US file
2. grep '^$US' file
3. grep ^$US* file
4. grep '^$US*' file

Exercice 3

Using the `touch` command, create the files : `f2_R1.fq`, `f2_R2.fq`, `f3_R1.fq`, `f3_R2.fq`, `f4_R1.fq`, `f4_R2.fq`, `wt_f1_R1.fq`, `wt_f1_R2.fq`, `wt_f2_R1.fq`, `wt_f2_R2.fq`, `WT_f3_R1.fq`, `wt_f3_R2.fq`, `wt_f4_R1.fq`, `wt_f4_R2.fq`, `wT_f5_R1`

Q1 - List the files whose name :

Filter 1	Begin with « wt » or « WT » or « Wt » or « wT »
Filter 2	Begin with « wt » or « WT » or « Wt » or « wT » Ends with « fq »
Filter 3	Begin with « wt » or « WT » or « Wt » or « wT » Ends with « fq » « f3 » is in the name

TP2 : Using grep and sed

Exercice 1

- Q1 - By using the `« wget »` command, load the fruitfly chromosome file from Ensembl and unzip it. File location is : ftp://ftp.ensembl.org/pub/release-90/fasta/drosophila_melanogaster/dna/Drosophila_melanogaster.BDGP6.dna.toplevel.fa.gz
- Q2 - Count the number of lines
- Q3 - Count the number of sequences
- Q4 - Using sed, remove chromosome 2L sequence (with header line) from file and create a new file named `Drosophila_melanogaster.BDGP6.dna.toplevel.No2L.fa`
- Q5 - Using sed, extract chromosome 2R sequence (without header line) from the Drosophila_melanogaster.BDGP6.dna.toplevel.fa file and put it in R2.fasta
- Q6 - Using sed, add ">2R extracted_chromosome" as the new header of 2R.fasta
- Q7 - Using sed, rename chromosomes by adding « chr » before the number.

Exercice 2

- Using sed, convert the `f1_R1.fq` file to fasta. Rename `f1_R1.fa` the newfile.

Exercice 3

Consider the fasta file named « `ncRNA_Tbarophilus.fasta` »

By using `grep` and `sed` build a file which, from the analysis of the header, extract the name of the gene (without space) and its genomic positions.

Example for the three first headers :

```
>ENA|CP002372.1:102131..102217:tRNA|CP002372.1:102131..102217:tRNA.1 Thermococcus barophilus MP tRNA-Ser
GCCGGGATCGCCTAGCCTGGGATGGCGCGGGCCTTGAGAGCCCGTGGGCGTTTGCCCGCC
GGGGTTCAAATCCCCGTCCCGGCGCCA
>ENA|CP002372.1:104485..104572:tRNA|CP002372.1:104485..104572:tRNA.1 Thermococcus barophilus MP tRNA-Leu
GCGGGGGTTGCCGAGCCTGGTCAAAGGCGCGGGATTCAGGGTCCCGTCCCGCAGGGGTTC
CGGGGTTCAAATCCCCGCCCCCGCACCA
>ENA|CP002372.1:1050973..1051060:tRNA|CP002372.1:1050973..1051060:tRNA.1 Thermococcus barophilus MP tRNA-Leu
GCGGGGGTTGCCGAGCCTGGTCAAAGGCGCGGGATTGAGGGTCCCGTCCCGTAGGGGTTC
CGGGGTTCAAATCCCCGCCCCCGCACCA
```

Will give

```
tRNA-Ser    102131      102217
tRNA-Leu    104485      104572
tRNA-Leu    1050973     1051060
```

Exercice 4

- Q1 : Consider paired-end files f1_R1.fq and f1_R1.fq. Count the number of lines in both files. Is it correct as expected for paired-end files ? Delete blank lines in each file. Count again the number of lines in both files. Is it correct ?

TP3 : Using grep, sed and awk

Exercice 1

Consider the file `hg19_exons.bed`.

Q1 - By using awk, generate a new file named new_hg19_exons.bed with fields 1, 2, 3 et 6 separated by a space.

Q2 - By using awk, generate a new file named newtab_hg19_exons.bed with fields 1, 2, 3 et 6 separated by a tabulation.

Q3 - By using awk, generate a new file named new_chr1_hg19_exons.bed with fields 1, 2, 3 et 6 separated by a space only for chromosome 1 (chr1).

Q4 – By using awk , generate a new file (new1.bed) containing a 5th field giving the length of exon.

Q5 - By using awk , generate a new file (new2.bed) containing only exons of size > 100nt.

Exercice 2

You want to prepare a file command to align a set of fastq files from the cluster by using the STAR alignment software. For each file, you have to write the lines :

module load bioinfo/STAR-2.6.0c ; STAR --genomeDir star-index --readFilesIn *file_name*.fq --outFileNamePrefix *file_name*

By using `awk` build the file including all commands for the fastq files (ending with « `.fq` ») in your directory.

Exercice 3

- Print the total number of reads in file `f1_R1.fq`, the total number of unique reads, the percentage of unique reads, the most abundant sequence, its frequency, and its percentage of total: