



Tip: use the internet

Getting acquainted with the Unix text processors: grep, sed and awk

1. First read and study the associated tutorials:
 1. grep: <https://ryanstutorials.net/linuxtutorial/grep.php>
 2. sed: <https://www.tutorialspoint.com/sed/>
 3. awk: <https://www.tutorialspoint.com/awk/>

2. Perform online challenges involving the above programs

'Hacking challenges'

1. For the next part of today we will further our training in using the command-line Unix applications '**sed**', '**grep**' and '**awk**'.
2. Together with **less**, **more**, **head**, **tail**, **uniq**, **sort** and **cut** they form the heart of many bioinformatics pipelines.
3. Go to <http://www.hackerrank.com> and obtain a user account. Then login and go to this set of challenges:
<https://www.hackerrank.com/domains/shell?filters%5Bsubdomains%5D%5B%5D%5D%3Dgrep-sed-awk>
4. Start with the **grep** challenges, followed by **sed** and finally **awk**.

'Biological' exercises 1 (parsing a GFF FILE)

1. Obtain the *.gff.gz file from https://www.ncbi.nlm.nih.gov/assembly/GCF_000149205.2/
2. We are going to look at the gene predictions for a fungus, *Aspergillus nidulans*, by examining the GFF3 formatted gene prediction file. More, **essential**, information about GFF (*gene feature file*) format here: <http://gmod.org/wiki/GFF3>

Using Unix-tools only, answer the following questions:

1. How many sequences are described in the GFF file for this genome of *A. nidulans*? Does this match with the number of sequences in the associated genomic fasta file:
GCF_000149205.2_ASM14920v2_genomic.fna.gz ? (*find it at the above dir on ncbi*)
2. How many transcripts (**mRNA**) are encoded on sequence "NT_107008.1"?
3. What is the average number of exons per transcript (for all transcripts)?
(*hint: you can use the command line calculator **bc** (add -l for floating point arithmetic)*)
What is their average length?
(*hint: you can use **awk**'s default operators such as **NR** (record number) and **NF** (field number)*)
4. How many single-exons transcripts can be found for this organism?
5. Which transcript encodes the largest protein (number of amino acids)?
And which one the smallest?
6. Calculate the density of transcripts for each sequence (defined as the number of transcripts/megabase).
Which sequence has the highest transcript density? Which one the lowest?

'Biological' exercises 2

Task: Identify differentially abundant transcripts between two conditions, each containing three replicate samples

Data: In triplicate, kallisto quantification outputs (see: <https://pachterlab.github.io/kallisto/>) for fungal rna-seq performed at two conditions, control: 1-3 and treatment 4-6.

Download the dataset from here:

<https://surfdribe.surf.nl/files/index.php/s/I01hsIVE65FBWP7/download> and unpack.

Find a way to combine the information from the six *.tsv files in the six directories and extract transcripts that are on average two-fold higher in C4 than in C1. **For expression values we use the 'tpm' column!.** **How many were there?** **Now do the reverse.**

target_id	length	eff_length	est_counts	tpm
lcl NZ_CP007637.1_cds_WP_010207718.1_1	1518	1219	664	40.1639
lcl NZ_CP007637.1_cds_WP_010207719.1_2	1104	805	1028	94.1605
lcl NZ_CP007637.1_cds_WP_010207720.1_3	1104	805	452	41.4013
lcl NZ_CP007637.1_cds_WP_010207722.1_4	2418	2119	2161	75.1961

Loops

Not covered (today), but incredibly powerful:

for and **while** loops

```
for file in *.gff
do
    echo ${file}
    cat ${file} | grep -v "^#" | grep -P "\tgene\t" -c
    cat ${file} | grep -v "^#" | grep -P "\tgene\t" | \
    while a ; do grep -P "\t$a\t" ${file} | awk .... .... ... ; done
done
```

More:

<https://linuxize.com/post/bash-for-loop/>

<https://linuxize.com/post/bash-while-loop/>