

Command line practical

More advanced shell

To begin:

Open Terminal in your VM. Navigate to your folder containing the data used for the previous practical.

Hint: Previous Practical Step 2

1. Using Core utilities

A. grep: globally search a regular expression and print

- `man grep` Everything you need to know.
- `grep 'EAS' sequences.fa`
- `grep 'Contig' sequences.fa`
- `grep 'contig' sequences.fa`
- `grep -i`: grep, case insensitive `grep -i 'contig' sequences.fa`
- `grep -r`: recursive grep `grep -r 'Contig' ./`
- You may also combine options, e.g.: `grep -ir 'contig' ./`
- `grep -A`: show lines after the grep hit `grep -i -A 1 'contig' sequences.fa` Useful if the sequences are in one line.
 - Also try:
 - `grep -B 10 '<searchterm>' <filename>` (10 lines Before)
 - `grep -C 5 '<searchterm>' <filename>` (5 lines of Context)
 - Combinations are possible: `grep -A 1 -B 2 '<searchterm>' <filename>` (1 line After, 2 lines Before the hit)

B. Output redirection: `>` (write to file)

- `grep 'Contig' sequences.fa > selection.fa` Creates a file, if it does not exist. Writes the output into the file. *Caution: existing content will be overwritten. Do not write to the same file that you read from.*
- `grep 'Contig' sequences.fa >> selection.fa` Creates a file, if it does not exist. Appends the output to the end of the file. Existing content is not overwritten. Rather, the file grows. (You should still not write to the same file that you read from.)

C. sed: stream editor

- Add taxa name to end of all header lines:

```
sed 's/>.*/& Escherichia coli/' sequences.fa > outfile_sed.fa
```

D. awk: 'awk' stands for the names of its authors "Aho, Weinberger, and Kernighan"

- Clean up a fasta file so only first column of the header is outputted:

```
awk '{print $1}' outfile_sed.fa > output_awk.fa
```

2. Working with FASTA files

A. wc: count number of lines, words, characters

- Mnemonic: wc → word count. `-l` (for counting lines) is the most commonly used option. More information: `man wc`.

- **Exercise: sequence count**

- Find a way to output the number of Sequences in a FASTA file. You may use 'temporary' files. Tip: individual sequences are identified by their headers. Header lines in a FASTA file begin with '>'.

- **Possible solutions: sequence count** `man grep` `grep -c '>' sequences.fa`

```
grep '>' sequences.fa > tempfile.txt wc -l tempfile.txt
```

B. Output redirection: | (pipe)

- `ls -l | wc -l` 'Pipes' the output of `ls -l` into the input of `wc -l`. Prints out the number of items in this directory. No temporary file necessary.
- `grep '>' sequences.fa | less` 'Pipes' the output of grep into the input of less. Allows to comfortably read the list of headers in a FASTA file. No temporary file necessary.

- Displayed using less: Search using `/` Quit with `q`
- **Exercise: filter;** Output the first five headers in a FASTA file
 - Possible solution: `grep '>' sequences.fa | head -n 5`
- **Exercise: one specific line;** Output line number 19 of a given file.
 - Possible solution: `head -n 19 sequences.fa | tail -n 1`
- **Exercise: collecting sequences;** Save the first three and the last three headers from a FASTA file in a new file. This cannot be solved (easily) with only one command.
 - Possible solution: `grep '>' sequences.fa | head -n 3 > newfile.fa` and `grep '>' sequences.fa | tail -n 3 >> newfile.fa`

3.Managing Files

A. tar: tape archive

```
tar -cvf new_tar_archive.tar <list>
```

- Mnemonic: `tar` → tape archive
- `c` : create new archive
- `v` : verbose output
- `f` : file to use; can take files and directories as arguments
- **Exercise: create a tar archive;**
 - `tar -cvf sequences.tar sequences.fa`

Show the content of a tar archive: `tar -tvf new_tar_archive.tar`

Unpack a tar archive: `tar -xvf tar_archive.tar`

- **Exercise: unpack a tar archive;**
 - `tar -xvf sequences.tar`

B. gzip: compress files

- **Exercise: compress fasta file;**
 - `ls -lh sequences.fa`

- `gzip sequences.fa`
- `ls -lh sequences.fa.gz`

Compression reduces file size. Useful for copying over network, Mail attachments, storing on removable media, ...

- **Uncompress:** `gzip -d sequences.fa.gz` Mnemonic: `d` → decompress

UCLA Microbial Metagenomics Workshop 2015

Modified from Hydra Workshop 2015 - Day 1