Introduction to Linux and text processing

Color codes

Text in Blue – Refer to "Info" section
Text background in Turquoise – Commands
Text background in Grey – Command output
Text in Pink – Points to remember
Text background in Red – Alert/Warning

1 Introduction

Linux is an open-source operating system (OS) developed based on the kernel created by Linus Benedict Torvalds. In the last two decades, Linux has gained much popularity and now is being used on many platforms. Nowadays, most of the high- end servers to mobile phones (Android OS or iOS) run on different variants of Linux.

Linux computers/servers are installed for multi-user usage. In this course, we will work on a virtual machine (VM) running Ubuntu desktop edition. Most of the commands specified in this manual can be used in any other distribution (i.e., CentOS, Debian, etc.) of Linux operating system.

Resource: How to install Linux?

To install Ubuntu on personal computers, please follow the instructions in the following link.
https://tutorials.ubuntu.com/tutorial/tutorial-install-ubuntu-desktop#0

1.1 The Terminal

We use terminal (AKA command line interface) to interact with the operating system. The terminal by default runs one of the "shells". Shell is a program that sits between the user and the kernel and translates user commands (text) into machine code. The advantages of using command line are greater control and flexibility over the system or software and multiple commands can be saved in a file and executed as a program.

The most common shells are:

Bourne Shell
Bourne Again Shell – BASH (variant is Z Shell)
C Shell (variant is T Shell)
K Shell

Among these Bourne Again Shell (BASH) is the most popular one. This is the default shell on the system, and we will be using it throughout this course.

Info

Linux	Unix derivative, most popular variant of Unix
OS	Software that commands the hardware and make the computer work

VM	A resource that can run an OS independent of the host OS (e.g., Ubuntu VM on a Mac OS)
Ubuntu	Free Linux distribution (distro) based on Debian (an oldest OS based on Linux kernel)
Kernel	Core interface between a computer's hardware and its processes, manages available resources

1.2 Connecting to Linux Server

To connect to a remote server, use 'ssh' command.

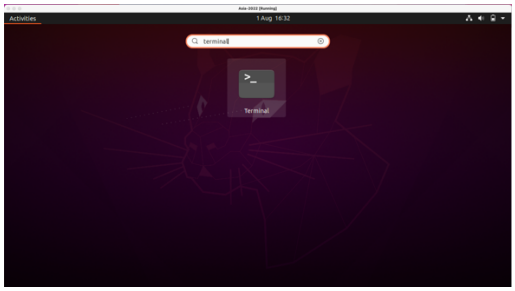
ssh user@servername
ssh user@ipaddress

There are many other commands to connect to the server (telnet, rsh etc...) which are less secure and outdated.

In this course, we will use a virtual machine running Ubuntu desktop edition. We can directly access the OS using the terminal without the 'ssh' command.

2 Linux command structure

When you open a terminal in Linux/Ubuntu (terminal icon on the left sidebar looks like the below image; use search bar by typing the word "terminal" if necessary), you will see a command prompt, ready to take commands. The default location on the terminal is your "home directory". It is represented with ~ (tilde) symbol.



Open a terminal window, copy the command below and paste it into your command line to make sure that we are all working in the same directory.

```
cd ViralBioinfAsia2022/course_data/Introduction_to_Linux_Unix_Text_processing
```

All Linux commands are single words (can be alpha-numeric), with optional parameters followed by arguments. For historical reasons, some of the early commands are only two letter long and case sensitive. Most of the command options (also called flags) are single letters. They should be specified after the command before specifying any input.

```
ls -l
ls -l Exercises
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
```

```
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

"ls" is the command to list the contents of the directory, "-l" is the option for long listing and "Exercises" is the input, which is optional in this case. Without the input, "ls" shows the contents of the current directory (Type "ls -l").

3 First Commands

Directories are the Unix equivalent of folders on a PC or a Mac. They are organized in a hierarchy, so directories can have sub-directories and so on. Directories, like folders, are useful to keep your data files organized. The location or directory you are currently in, is called the current working directory. The location or "full pathname" of the file SARS-CoV-2.fa in the Introduction_to_Linux_Unix_Text_processing directory can be expressed as:

/home/manager/course_data/Introduction_to_Linux_Unix_Text_processing/SARS-CoV-2.fa

Tab completion

Typing out longer file names can be boring, and you are likely to make typos that will, at best, make your command fail with a strange error and at worst, overwrite some of your carefully crafted analysis. Tab completion is a trick that normally reduces this risk significantly. Instead of typing out "ls Interesting_stuff/", try typing "ls Int" and press the Tab button (instead of Enter). The rest of the folder/file names that begin with "Int" should be listed. If you have two folders/files with similar names (e.g., my_awesome_scripts/ and my_awesome_results/) then you might need to give your terminal a bit of a hand to work out which one you want. In this case if you type "ls –l m", when you press Tab the terminal would read "ls –l my_awesome_". You could then type "s" followed by another press of Tab button and it would figure out that you meant "my_awesome_scripts/".

Points to remember:

- *Linux commands are case sensitive and are always single words*
- *Options follow the command - and they start with a single hyphen (-) and a character or a double hyphen (- -) and a word*
- *Single character options can be combined*
- *Argument can be one or more inputs*
- *You can write more than one command separating with a semicolon; You can use the "tab" keystroke to autofill the command.*

Info

ssh	Program for logging in to a remote machine specified with a host name
left sidebar	Vertical task management panel
PC	A personal computer
Mac	A Macintosh computer

Important commands

(a) ls
Lists information about the files/directories. Default is the current directory. Sorts entries alphabetically.

Commonly used options:
-l long list
-a show all files (including hidden files)
-t sort based on last modified time

Type:
ls -l

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

Information (from left to right):

- File permissions
- Number of links
- Owner name
- Group name
- Number of bytes
- Abbreviated month, last modified date and time
- File/Directory name

(b) pwd (print working directory)
Returns the path of the current working directory to the standard output.

Type:
pwd

```
/home/manager/course_data/Introduction_to_Linux_Unix_Text_processing
```

(c) cd
Change current working directory to the specified directory.

Type:
cd Exercises
pwd

```
/home/manager/course_data/Introduction_to_Linux_Unix_Text_processing/Exercises
```

We are now in the directory "Exercises". Typing the command "cd .." changes it to the parent directory from which the previous command was typed in. Typing "cd" will change the current directory to the home directory.

```
cd ..
cd
cd ViralBioinfAsia2022/course_data/Introduction_to_Linux_Unix_Text_processing
```

(d) mkdir
This command creates a directory in the current working directory if no directory exists with the specified name.

Type:

```
mkdir Practice
ls -l
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
drwxrwxr-x 2 manager manager  4096 Aug 12 16:10 Practice
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

(e) rmdir
This command is used to remove directories.

Type:

```
rmdir Practice
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

(f) touch
It is file's timestamp changing command. However, it can be used to create an empty file. This command is generally used to check if there is write permission for the current user.

Type:

```
touch temp-file
ls -l
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
```

```
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager     0 Aug 12 16:12 temp-file
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

(g) rm
rm is used for removing files and directories.

Type:
rm temp-file
ls -l

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

To remove directories use "-r" option. Please remember once a file or directory is deleted, it will not go to "Recycle bin" in Linux and there is no way you can recover it.

(h) cp
Copies the content of the source file/directory to the target file/directory. To copy directories, use "-r" option.

Type:
touch temp1
cp temp1 temp2
ls -l

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
-rw-rw-r-- 1 manager manager     0 Aug 12 16:14 temp1
-rw-rw-r-- 1 manager manager     0 Aug 12 16:14 temp2
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

(i) mv
To move/rename a file or a directory.

Type:

```
mkdir temp
mv temp1 temp/.
mv temp2 temp3
ls -l
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
drwxrwxr-x 2 manager manager  4096 Aug 12 16:15 temp
-rw-rw-r-- 1 manager manager     0 Aug 12 16:14 temp3
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

The second command moves the "temp1" file into the directory "temp". The "." (dot) at the end of the command retains the name of the file, whereas the third command renames the file "temp2" to "temp3".

(j) ln
Link command is used to make links to files/directories.

Type:
```
ln -s temp/temp1 .
ls -l
```

```
drwxrwxr-x 2 manager manager  4096 Jul 25 16:20 Exercises
-rw-rw-r-- 1 manager manager 11158 Jul 25 16:41 human_viruses.txt
-rw-rw-r-- 1 manager manager  2049 Jul 25 16:20 outbreak.csv
-rw-rw-r-- 1 manager manager  1134 Jul 25 16:20 readME.md
-rw-rw-r-- 1 manager manager   675 Jul 25 16:20 SARS-2-variants.tsv
-rw-rw-r-- 1 manager manager 30428 Jul 25 16:20 SARS-CoV-2.fa
-rw-rw-r-- 1 manager manager 78471 Jul 25 16:20 SARS-CoV-2.gb
drwxrwxr-x 2 manager manager  4096 Aug 12 16:15 temp
lrwxrwxrwx 1 manager manager    10 Aug 12 16:16 temp1 -> temp/temp1
-rw-rw-r-- 1 manager manager     0 Aug 12 16:14 temp3
-rw-rw-r-- 1 manager manager 13372 Jul 25 16:20 viruses.csv
```

We encourage you to create links rather than copying data in order to save space.

4 File viewers

(a) cat
The concatenate command combines files (sequentially) and prints on the screen (standard output).

Type:
```
cat SARS-CoV-2.fa
```

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete
genome
```

```
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
……
```

(b) more/less
These commands are used for viewing the content of the files; faster with large input files than text editors; not the entire file is read at the beginning.

Type:
more SARS-CoV-2.fa

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete
genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
……
```

Press "Enter" to view lines further and "q" to quit the program

(c) head/tail
These commands show first/last 10 lines (default) respectively from a file.

Type:
head SARS-CoV-2.gb

```
LOCUS       NC_045512              29903 bp ss-RNA     linear   VRL 18-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
```

5 File editors

There are many non-graphical text editors like ed, emacs, vi and nano available on most Linux distributions. Some of them are very sophisticated (e.g., vi) and for advanced users.

Nano (earlier called pico) is like any graphical editor without a mouse. All commands are executed using the keyboard, using the <CTRL> key modifier. It can be used to edit virtually any kind of text file from the command line. Nano without a filename gives you a standard (blank) nano window.

At the bottom of the screen, there are commands with a symbol in front. The symbol tells that you need to hold down the Control (Ctrl) key, and then press the corresponding letter of the command you wish to use.

Ctrl+X will exit nano and return you to the command line.

Nano Quick Reference

Ctrl+X: Exit the editor. If you've edited text without saving, you'll be prompted as to whether you really want to exit.

Ctrl+O: Write (output) the current contents of the text buffer to a file. A filename prompt will appear; press Ctrl+T to open the file navigator shown above.

Ctrl+R: Read a text file into the current editing session. At the filename prompt, hit Ctrl+T: for the file navigator.

Ctrl+K: Cut a line into the clipboard. You can press this repeatedly to copy multiple lines, which are then stored as one chunk.

Ctrl+J: Justify (fill out) a paragraph of text. By default, this reflows text to match the width of the editing window.

Ctrl+U: Uncut text, or rather, paste it from the clipboard. Note that after a Justify operation, this turns into unjustify.

Ctrl+T: Check spelling.

Ctrl+W: Find a word or phrase. At the prompt, use the cursor keys to go through previous search terms, or hit Ctrl+R to move into replace mode. Alternatively, you can hit Ctrl+T to go to a specific line.

Ctrl+C: Show current line number and file information.

Ctrl+G: Get help; this provides information on navigating through files and common keyboard commands

Getting help in Linux

All Linux commands has manual pages. To access them, use "man" or "info" command. The manual page gives a detailed explanation of the command, all available options and sometimes, also provides examples. For example, to view the manual page for "ls" command

Type:
man ls

LS(1) User Commands LS(1)

NAME
ls - list directory contents

SYNOPSIS
ls [OPTION]... [FILE]...

DESCRIPTION

List information about the FILEs (the current directory by default). Sort entries alphabetically if none of -cftuvSUX nor --sort is specified.

Mandatory arguments to long options are mandatory for short options too.

-a, --all
do not ignore entries starting with .

-A, --almost-all
do not list implied . and ..

--author
with -l, print the author of each file

Please explore manual pages of all the above commands for available options.

6 Commands for text processing

(a) cut
The cut command is a command line utility to cut a section from a file. Please see "man cut" for available options.

To cut a section of file use "-c" (characters)

Type:
cut -c1-10 SARS-CoV-2.fa

```
>NC_045512
ATTAAAGGTT
CGAACTTTAA
TAATTACTGT
TTGCAGCCGA
CCTGGTTTCA
GTGGCTTTGG
CTTAGTAGAA
GCTCGAACTG
GTAGTGGTGA
……….
```

The option "-c1-10" will output first 10 characters from the input file.

Few options:
-c: cut based on character position
-d: cut based on delimiter
-f: field number

We have a file named "human_viruses.txt" with all the virus names, genbank ids and genome length. These fields are separated by "|" symbol.

Type:
head human_viruses.txt

```
gi|116326654|ref|NC_008523.1| Tomato leaf curl Karnataka virus-associated DNA beta
DNA-A, complete genome |1358
gi|116326656|ref|NC_008524.1| Tomato leaf curl virus-Pune-associated DNA beta DNA-
A, complete genome |1380
gi|116536742|ref|NC_008517.1| Tomato leaf curl Pune virus, complete genome |2756
gi|1189489108|ref|NC_034480.1| Bwamba virus strain M459 segment S nucleoprotein
and nonstructural protein NSs genes, complete cds |1096
gi|1189489111|ref|NC_034481.1| Nyando virus strain MP401 segment S nucleoprotein
and nonstructural protein NSs genes, complete cds |902
gi|1189489154|ref|NC_034501.1| Kaeng Khoi virus strain PSC-19 segment S
nucleoprotein and nonstructural protein NSs genes, complete cds |975
gi|1215835591|ref|NC_035211.1| Human fecal virus Jorvi4, complete genome |2343
gi|1215835602|ref|NC_035213.1| Human fecal virus Jorvi3, complete genome |1942
gi|13277517|ref|NC_001653.2| Hepatitis delta virus, complete genome |1682
gi|1328406490|ref|NC_036615.1| Influenza D virus (D/swine/Oklahoma/1334/2011)
segment 2 polymerase PB1 (PB1) gene, complete cds |2330
```

To get only the genbank id,

Type:
cut -d "|" -f2 human_viruses.txt

```
116326654
116326656
116536742
1189489108
1189489111
1189489154
1215835591
1215835602
13277517
1328406490
……
```

(b) sort
The sort command is used to sort the input content.

Few options:
-t: field separator
-n: numeric sort
-k: sort with a key (field)
-r: reverse sort
-u: print unique entries

Type:
sort -t "|" -nrk6 human_viruses.txt

```
gi|658609068|ref|NC_024443.1| Roundleaf bat hepatitis B virus isolate RBHBV/GB09-
256/Hip_rub/GAB/2009, complete genome |3368
gi|401829632|ref|NC_018481.1| CAS virus segment S, complete genome |3368
gi|172088171|ref|NC_010562.1| Chapare virus segment S, complete sequence |3357
gi|594139507|ref|NC_023761.1| Boa arenavirus NL isolate 3 segment S, complete
sequence |3351
gi|1464307246|ref|NC_038364.1| Loie River virus isolate R5074 glycoprotein
precursor (GPC) and nucleocapsid protein (NP) genes, complete cds |3344
gi|81097459|ref|NC_007547.1| Rotavirus C segment 1, complete sequence |3309
gi|295441891|ref|NC_014093.1| Torque teno midi virus 2, complete genome |3253
……
```

(c) grep
grep searches the input for a given pattern.

Few options:
-A: after context
-B: before context
-C: before and after context
-c: count
-l: file with match
-i: ignore case
-o: only match
-v: invert match
-w: word match

To get the list of all Hepatitis viruses from human_viruses.txt

Type:
grep "Hepatitis" human_viruses.txt

```
gi|13277517|ref|NC_001653.2| Hepatitis delta virus, complete genome |1682
gi|941241313|ref|NC_003977.2| Hepatitis B virus (strain ayw) genome |3182
```

(d) wc
The command "wc" can be used in 2 ways, which counts lines, words or characters.

Type:
wc -l outbreak.csv

36 outbreak.csv

(or)

cat outbreak.csv | wc -l

36

(e) uniq
The uniq command extracts unique lines from the input. It is usually used in combination with sort to count unique values in the input.

To get the list of countries that has had an outbreak in 2022:

Type:
cut -d, -f3 outbreak.csv | sort | uniq

```
Afghanistan
Australia
China
Democratic Republic of the Congo
Germany
Guinea
Iraq
Israel
```

```
Kenya
Malawi
Mozambique
Multi-Country
Multi-country
Nigeria
Oman
Qatar
Sao Tome and Principe
Saudi Arabia
Somalia
Timor-Leste
Togo
Uganda
United Kingdom of Great Britain and Northern Ireland
United States of America
```

Other text processing commands worth looking at are: tr, rev, sed and paste.

7 I/O control in Linux

When you run a command, the output is usually sent to standard output (stdout) ie. the terminal. However, we can redirect the standard output to a file using ">".

Type:

```
ls > list
cat list
```

```
Exercises
human_viruses.txt
list
outbreak.csv
readME.md
SARS-2-variants.tsv
SARS-CoV-2.fa
SARS-CoV-2.gb
viruses.csv
```

The first command creates a new file called list with all the file names in the directory. If there exists a file already named "list", it is overwritten with the output of the command. Instead, we can append to a file using ">>" redirection.

Another kind of output that is generated by programs is standard error. We must use "2>" to redirect it.

```
ls /foo 2> error
```

To redirect stdout and stderr to a file use "&>"

7.1 Pipes

Piping in Linux is a very powerful and efficient way to combine commands. Pipes (|) in Linux act as connecting links between commands. Pipe redirects output of the first command as an input to the next command. We can nest as many commands as we want using pipes. They ensure smooth running of the command flow and reduces the execution time.

To print 10 smallest viruses

Type:
sort -t "|" -nk6 human_viruses.txt | head -10

```
gi|73921565|ref|NC_007380.1| Influenza A virus (A/Korea/426/1968(H2N2)) segment
8, complete sequence |838
gi|752901105|ref|NC_026428.1| Influenza A virus (A/Shanghai/02/2013(H7N9))
segment 8 nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes,
complete cds |841
gi|758899352|ref|NC_026432.1| Influenza A virus (A/California/07/2009(H1N1))
segment 8 nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes,
complete cds |863
gi|1328406502|ref|NC_036621.1| Influenza D virus (D/swine/Oklahoma/1334/2011)
segment 7 nonstructural protein 2 (NS2) and nonstructural protein 1 (NS1) genes,
complete cds |868
gi|32140160|ref|NC_004906.1| Influenza A virus (A/Hong Kong/1073/99(H9N2))
segment 8, complete sequence |890
gi|1189489111|ref|NC_034481.1| Nyando virus strain MP401 segment S nucleoprotein
and nonstructural protein NSs genes, complete cds |902
gi|751868323|ref|NC_026282.1| Maprik virus isolate MK7532 segment S, complete
sequence |911
gi|1189489154|ref|NC_034501.1| Kaeng Khoi virus strain PSC-19 segment S
nucleoprotein and nonstructural protein NSs genes, complete cds |975
gi|758899349|ref|NC_026431.1| Influenza A virus (A/California/07/2009(H1N1))
segment 7 matrix protein 2 (M2) and matrix protein 1 (M1) genes, complete cds |982
gi|22256027|ref|NC_004110.1| La Crosse virus segment S, complete genome |984
```

We will be working on other examples during the course, where we use pipes to combine more than two commands.

8 Process control

Some commands take time to finish the assigned job. For example, if you would like to compress a huge file with gzip command that takes a few minutes to finish running, you can run it in the background by appending the command with "&" (Another way is to suspend a command by pressing Ctrl+Z and typing "bg"). The completion of the task is indicated by "Done".

Type:
gzip list &

We can get list of currently running jobs in the terminal by "jobs" command. This will give you all the background jobs running in the current terminal. If you want to see all the running processes in the system, use "top". You can get user specific details in top using "-u" option.

Type:
top

Few of the important columns in top output:

PID: Process Id, this is a unique number used to identify the process.
COMMAND: Command Name
S: Process Status: The status of the task which can be one of:
– D = uninterruptible sleep
– R = running
– S = sleeping

– T = traced or stopped
– Z = zombie

If you want to stop a running background job use "kill" command followed by the process id.

kill 1234

This command kills the job with the process id 1234. As a user you can kill only your jobs. You do not have permission to run this command on the process ids of other users.

Command line shortcuts

- Up/Down arrows: Previous commands
- !!: Reruns previous command
- Tab: Auto complete
- Tab+Tab: All available options
- Ctrl+a: Move cursor to start of line
- Ctrl+e: Move cursor to end of line
- Alt+: Alternates between terminals
- Ctrl+l: Clear screen (or Command+k on Mac)
- Ctrl+c: Terminates the running program
- Ctrl+z: Suspends the running program
- Ctrl+w: Removes a previous word
- Ctrl+d: Logout
- Ctrl+d(in a command): Removes a character
- Ctrl+u: Removes till the beginning

Exercises:

1. Open a new terminal and navigate into Exercises directory (cd Exercises/).
2. Extract first 15 lines from the file "HM067743.1_cds_ADQ37313.1_1.fa" and save the output into "output.fa"
3. How many fasta files are there in the directory?
4. Extract all header lines from the file all.fa
5. How many sequences are there in the file all.fa?
6. Get the list of countries (excluding multi country outbreaks) that had an outbreak in 2022 (Input: outbreak.csv)
7. Find the number of outbreaks (exclude multi country outbreaks using invert match grep (-v)) in each month of 2022.

Quiz:

1. How will you find the 99th line of a file using only tail and head command?

2. Given a file how will you find the count of lines containing word "ABC".

3. How do you stop a process with pid 5678? (Hint: Type "man kill" to find out)

4. Which command can you use to re-execute a previous command?

5. .....is the command used to create new directory.

6. Command used to create an empty file.

7. Which is the command used to remove or delete file without confirmation prompt?

8. "cat" is the command used to .....

9. ..... command is used to count the total number of lines, words and character in a file

10. Which command would you use to know the location of your current working directory?

11. Which command would you use to extract 2nd, 5th, 7th column of a text file?

12. Extract first 10 lines of outbreak.csv, sort them and save as outbreak_1.csv. Extract first 20 lines of outbreak.csv, sort them and save as outbreak_2.csv

13. Extract common lines between the files outbreak_1.csv and outbreak_2.csv (use "comm" command, type man comm to get information)

14. Long list all the contents of the folder "Exercises" and save the output as a file named "list".

15. How many lines are there in the file "SARS-2-variants.tsv" excluding the header?

16. What would the following command do?
cp ../file .

17. Which command would you use to find the word "pattern" from the file, "filename.txt"? Using that command, extract the "BioProject" information from the file SARS-CoV-2.gb.

18. Which option with the command "rm" is required to remove a directory?

19. The command used to display the manual pages for any command is

20. Which option of "ls" will show the hidden files?