

Introductory Linux Tutorial for Life Sciences

Session 7: Util commands

Linus Östberg

history

- enables the repeating of commands entered earlier in the session
- the GNU History library keeps track of all lines typed in the terminal

```
$ history
1997 find . -name '*.tex' | sort -n | wc -l
1998 exit
1999 cd myTeaching/linux_introduction/
2000 kile slides/linux_session6.tex&
2001 cd ..
...
2017 man history
2018 history
```

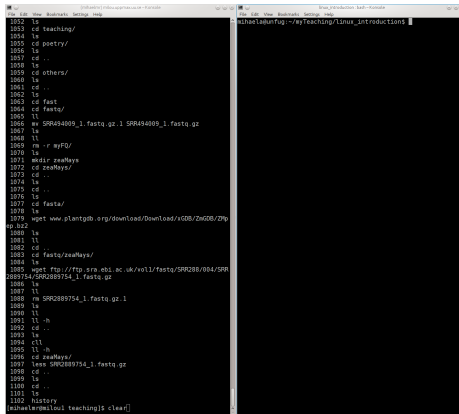
- use  and  to navigate through the history

clear

- clears the screen and shows only the prompt

```
$ clear
```

- is equivalent to  + L



```
1052 ls
1053 cd teachling/
1054 ls
1055 cd poetry/
1056 ls
1057 cd ..
1058 ls
1059 cd others/
1060 ls
1061 cd ..
1062 ls
1063 cd fastq/
1064 cd fastq/
1065 ll
1066 mv SRR494009_1.fastq.gz.1 SRR494009_1.fastq.gz
1067 ls
1068 ll
1069 rm -r myFQ/
1070 ls
1071 mkdir zealays
1072 cd zealays/
1073 cd ..
1074 ls
1075 cd ..
1076 ls
1077 cd fasta/
1078 ls
1079 wget www.plantgdb.org/download/Download/IGDB/ZmGDB/ZmG
1080 ls
1081 ll
1082 cd ..
1083 cd fastq/zealays/
1084 ls
1085 wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR286/004/SRR
1086 ls
1087 ll
1088 mv SRR2869754_1.fastq.gz.1
1089 ls
1090 ll
1091 ll -h
1092 cd ..
1093 ls
1094 cll
1095 ll -h
1096 cd zealays/
1097 less SRR2869754_1.fastq.gz
1098 cd ..
1099 ls
1100 cd ..
1101 ls
1102 history
[michael@miloul teaching]$ clear
```

Checking disk space

- **du** – shows how much disk space is taken by your files

```
$ cd /proj/g2015039/nobackup/nov2015/tuesday  
$ du -hs .  
4,2M .
```

-h — “human readable”

→ displays combined size of all files in the current directory and recursively in all its subdirectories

Checking disk space

- **du** – shows how much disk space is taken by your files

```
$ cd /proj/g2015039/nobackup/nov2015/tuesday
$ du -hs .
4,2M .
```

-h — “human readable”

→ displays combined size of all files in the current directory and recursively in all its subdirectories

→ displays combined size + the size of each subdirectory

```
$ du -h --max-depth=1 .
32K      ./TEST
224K     ./molecules
448K     ./plain_text
3,5M     ./sequences
4,2M     .
```

Checking disk space

- **df** – shows how much disk space is available

```
$ df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/mapper/vg-root 32G   14G   17G  44% /
tmpfs           63G   4,0K   63G   1% /dev/shm
/dev/sda1       2,0G  277M   1,6G  15% /boot
/dev/mapper/vg-scratch 234G   6,0G  216G   3% /scratch
gulo@tcp0:/glob 1,1P  598T  452T  57% /gulo
pica1-v3:/pica/v3 228T  169T   60T  74% /pica/v3
pica1-v2:/pica/v2 228T  165T   64T  73% /pica/v2
```

column

- formats the input into multiple columns that are much easier to read
- usage: `column <options> <file>`

```
$ grep -v "^#" mySample_somaticMutations.vcf | cut -f 1-8 | head
chr1    10043    .        T        G        .        REJECT    .
chr1    10055    .        T        G        .        REJECT    .
chr1    10067    .        T        G        .        REJECT    .
chr1    10079    .        T        G        .        REJECT    .
chr1    10157    .        T        C        .        REJECT    .
chr1    10180    rs201694901    T        C        .        REJECT    DB
chr1    10250    rs199706086    A        C        .        REJECT    DB
chr1    726859    rs139100483    A        G        .        REJECT    DB
chr1    726887    .        A        G        .        REJECT    .
chr1    726895    rs141325488    A        G        .        REJECT    DB
$ grep -v "^#" mySample_somaticMutations.vcf | cut -f 1-8 | column -t | head
chr1    10043    .        T        G        .        REJECT    .
chr1    10055    .        T        G        .        REJECT    .
chr1    10067    .        T        G        .        REJECT    .
chr1    10079    .        T        G        .        REJECT    .
chr1    10157    .        T        C        .        REJECT    .
chr1    10180    rs201694901    T        C        .        REJECT    DB
chr1    10250    rs199706086    A        C        .        REJECT    DB
chr1    726859    rs139100483    A        G        .        REJECT    DB
chr1    726887    .        A        G        .        REJECT    .
chr1    726895    rs141325488    A        G        .        REJECT    DB
```

Join

- **join** – used to join different files together by a common column
- works only if both files are sorted by the column to be joined on

```
$ cat example.bed
chr1 26 39
chr1 53 84
chr3 32 99
chr1 9 28
chr2 10 19
$ cat example_length.txt
chr1 58352
chr2 39521
chr3 24859

$ sort -k1,1 example.bed > example_sorted.bed
$ sort -k1,1 example_length.txt > example_length_sorted.txt

$ join -1 1 -2 1 example_sorted.bed example_length_sorted.txt
chr1 26 39 58352
chr1 53 84 58352
chr1 9 28 58352
chr2 10 19 39521
chr3 32 99 24859
```

Comparing files

- `diff` – reports differences between files.
- usage: `diff [OPTIONS] FILE1 FILE2`
- useful options:
 - `-b` – ignore blanks
 - `-w` – ignore white spaces and tabs
 - `-i` – ignore case
 - `-r` – recursively compare all files (when comparing folders)
 - `-y` – side by side comparison of files
 - `-y` – unified diff
- normal output shows only the lines that are different between 2 files: `< FROM-FILE-LINE > TO-FILE-LINE`

diff example

```
$ diff structures.txt structures2.txt
3d2
< 2FZE
5c4
< 1D1T
____
> 1D1S
```

diff example

```
$ diff -y structures.txt structures2.txt
1U3W                                     1U3W
2FZW                                     2FZW
2FZE                                   <
3COS                                     3COS
1D1T                                   | 1D1S

$ diff -u structures.txt structures2.txt
--- structures.txt      2019-11-25 15:55:27.363634645 +0100
+++ structures2.txt     2019-11-25 15:56:37.910306170 +0100
@@ -1,5 +1,4 @@
 1U3W
 2FZW
-2FZE
 3COS
-1D1T
+1D1S
```

Interpreting the output

```
$ diff genList1.txt genList2.txt
gen4c4
< 1AY6:K
_____
> 1AY6:H
7c7
< 1BRR:D
_____
> 1BRR:A
```

- 1st line – **gen4c4**:
 - **c** – changed (**a** – added, **d** – deleted)
 - left number – line numbers of the original file
 - right number – line numbers of the modified file
- 2nd line – **< 1AY6:K** – shows lines from the first file that are different from the second file
- 3rd line – a divider
- 4th line – **> 1AY6:H** – shows lines from the second file that are different from the first one

Downloading data

- **wget** – is a tool for non-interactive download of files from the Web (http, https, ftp)
- **wget 'http://website.url'**

```
$ wget https://www.ebi.ac.uk/pdbe/entry-files/pdb3cos.ent
--2019-11-21 15:19:29-- https://www.ebi.ac.uk/pdbe/entry-files/pdb3cos.ent
Loaded CA certificate '/etc/ssl/certs/ca-certificates.crt'
Resolving www.ebi.ac.uk (www.ebi.ac.uk)... 193.62.193.80
Connecting to www.ebi.ac.uk (www.ebi.ac.uk)|193.62.193.80|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1113102 (1.1M) [text/plain]
Saving to: 'pdb3cos.ent.1'

pdb3cos.ent.1          100%[=====>]  1.06M  1.49MB/s   in 0.7 s

2019-11-21 15:19:30 (1.49 MB/s) - 'pdb3cos.ent.1' saved [1113102/1113102]
```

Example curl

- **curl** – is a tool to transfer data from or to a server (http, https, ftp, imap, pop3)

→ the data is downloaded directly to the screen

curl 'http://website.url'

```
$ curl https://www.ebi.ac.uk/pdbe/entry-files/pdb3cos.ent
% Total    % Received % Xferd  Average Speed   Time    Time     Time    Current
           %         0         0              0             0      0 --:--:-- 0
    HEADER  OXIDOREDUCTASE                29-MAR-08  3COS
TITLE      CRYSTAL STRUCTURE OF HUMAN CLASS II ALCOHOL DEHYDROGENASE (ADH4) IN
TITLE      2 COMPLEX WITH NAD AND ZN
COMPND     MOL_ID: 1;
...
```

```
$ curl "https://www.ebi.ac.uk/pdbe/entry-files/pdb3cos.ent" > 3cos.pdb
$ curl "https://www.ebi.ac.uk/pdbe/entry-files/pdb3cos.ent" -o 3cos.pdb
$ curl "https://www.ebi.ac.uk/pdbe/entry-files/pdb{1ema,1gfl,1g7k,1xmz}.ent" -o
'#1'.pdb
```

Summary

- `history` – returns a history of used command lines
- `clear` – clears the screen
- `join` – join different files together by a common column
- `column` – formats the input into multiple columns
- `df` – shows how much disk space is available
- `du` – shows how much disk space is taken by your files
- `diff` – reports differences between files
- `wget,curl` – download data from an url from the shell