LSL Bioinformatics Series

Basic bioinformatics for beginners

Introduction to Linux command-line interface

Tsai-Ming Lu

Assistant Research Scientist

ICOB Bioinformatics Core

tmlu@gate.sinica.edu.tw

2022-12-07

Linux

Linux is a family of **open-source** Unix-like operating systems based on the Linux kernel released in 1991, by Linus Torvalds.

There are various Linux distributions,

e.g., RHEL (Red Hat), CentOS, Fedora, Ubuntu, Debian.

Why Linux?

Many open-source bioinformatics tools are command-line and are only available in Linux.

Free; easy to create analysis pipeline by integrating multiple tools

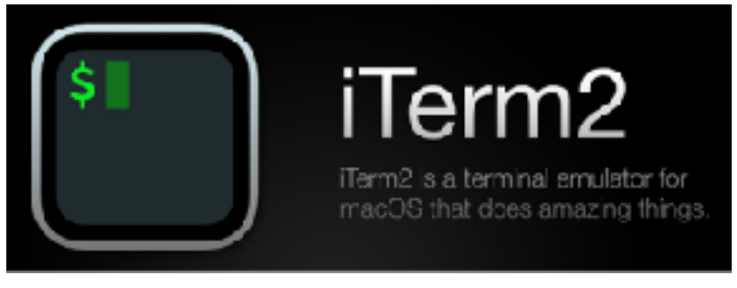
Shell, the command-line interface

What is the shell?

In Linux, the textual interface to communicate with the kernel via commands, e.g., bash, zsh, etc.

To open a shell ***prompt*** (where you can type commands),
you first need a *terminal*.

in the /Applications/Utilities folder



PATH; /directory/file location

```
tsai@TMLiMac ~ %
```

pwd | print working directory

tsai@TMLiMac ~% pwd

```
# Print the current directory:
/Users/tsai
```

Absolute path

Relative path

./ Current directory

../ Upper directory

~/ Home directory

tsai@TMLiMac ~% pwd

```
# Print the current directory:
/Users/tsai
```

cd | change directory

tsai@TMLiMac ~% cd example_folder ./example_folder

tsai@TMLiMac ~% pwd

```
/Users/tsai/example_folder
```

list directory contents

ls | list

tsai@TMLiMac ~% ls

```
fastaLengths.pl  gff3-py-1.0.1  parse_metrics.sh  row2column.awk
```

List all files, including hidden files:
ls -a
Long format list (permissions, ownership, size, and modification date):
ls -l
Long format list with size displayed using human-readable units (KiB, MiB, GiB):
ls -lh
Long format list sorted by size (descending):
ls -lS
Long format list of all files, sorted by modification date (oldest first):
ls -ltr

Command usage

man | manual

tsai@TMLiMac ~% man ls

```
LS(1)                          General Commands Manual                          LS(1)

NAME
     ls – list directory contents

SYNOPSIS
     ls [-@ABCFGHILOPRSTUWabcdefghiklmnopqrstuvwxy1%,] [--color=when] [-D format] [file ...]

DESCRIPTION
     For each operand that names a file of a type other than directory, ls displays its name as well
     as any requested, associated information.  For each operand that names a file of type
     directory, ls displays the names of files contained within that directory, as well as any
     requested, associated information.

     If no operands are given, the contents of the current directory are displayed.  If more than
     one operand is given, non-directory operands are displayed first; directory and non-directory
     operands are sorted separately and in lexicographical order.

     The following options are available:
```

tldr pages

> Simplified and community-driven man pages

The tldr pages are a community effort to simplify the beloved man pages with practical examples.

```
> tldr tar

  tar
  Archiving utility.
  Optional compression with gzip / bzip.

- Create an archive from files:
  tar cf target.tar file1 file2 file3

- Create a gzipped archive:
  tar czf target.tar.gz file1 file2 file3

- Extract an archive in a target folder:
  tar xf source.tar -C folder

- Extract a gzipped archive in the current directory:
  tar xzf source.tar.gz

- Extract a bzipped archive in the current directory:
  tar xjf source.tar.bz2

- Create a compressed archive, using archive suffix to determine the compression program:
  tar caf target.tar.xz file1 file2 file3

- List the contents of a tar file:
  tar tvf source.tar
```

File owner, group, permission

tsai@TMLiMac ~% ls

```
fastaLengths.pl  gff3-py-1.0.1  parse_metrics.sh  row2column.awk
```

tsai@TMLiMac ~% ls -l

```
total 16
-rwxr-xr-x 1 tmlu tmlu  628 Dec  6 17:24 fastaLengths.pl
drwxrwxr-x 6 tmlu tmlu 4096 Dec  6 17:24 gff3-py-1.0.1
-rw-rw-r-- 1 tmlu tmlu  798 Dec  6 17:24 parse_metrics.sh
-rwxrwx--- 1 tmlu tmlu  209 Dec  6 17:25 row2column.awk
```

group

owner

size

file name

r: read
w: write
x: execute

Change permission

tsai@TMLiMac ~% ls -l

```
total 16
-rwxr-xr-x 1 tmlu tmlu  628 Dec  6 17:24 fastaLengths.pl
drwxrwxr-x 6 tmlu tmlu 4096 Dec  6 17:24 gff3-py-1.0.1
-rw-rw-r-- 1 tmlu tmlu  798 Dec  6 17:24 parse_metrics.sh
-rwxrwx--- 1 tmlu tmlu  209 Dec  6 17:25 row2column.awk
```

chmod | change mode

tsai@TMLiMac ~% chmod 700 row2column.awk

tsai@TMLiMac ~% chmod 755 parse_metrics.sh

tsai@TMLiMac ~% ls -l

r: read (4)
w: write (2)
x: execute(1)

```
total 16
-rwxr-xr-x 1 tmlu tmlu  628 Dec  6 17:24 fastaLengths.pl
drwxrwxr-x 6 tmlu tmlu 4096 Dec  6 17:24 gff3-py-1.0.1
-rwxr-xr-x 1 tmlu tmlu  798 Dec  6 17:24 parse_metrics.sh
-rwx------ 1 tmlu tmlu  209 Dec  6 17:25 row2column.awk
```

Create a directory & download

- create a directory “workshop”

mkdir | **make a directory**

~$ mkdir workshop

- download an online file through a link

wget | **download files**

~$ wget https://www.dropbox.com/s/h8tndwpvzyf4xxz/Drerio_GRCz11_partial.gtf.gz

Copy & move a file

- move the .gz file into the "workshop" directory

mv | **move**

[usage] mv /old/PATH/filename /new/PATH/filename

~$ mv Drerio_GRCz11_partial.gtf.gz ./workshop

rm | **remove**

[usage] rm filename ; rm -r directory

- make a backup of gtf.gz

cp | **copy**

[usage] cp /old/PATH/filename /new/PATH/filename

~$ cd ./workshop

~$ cp ./Drerio_GRCz11_partial.gtf.gz ..

(De)compress & read files

- decompress the Drerio_GRCz11_partial.gtf.gz file

gzip | **compress or expand files**

[usage] gzip (-d) filename

~$ gzip -d Drerio_GRCz11_partial.gtf

- read the Drerio_GRCz11_partial.gtf

less | **open a file for interactive reading**

[usage] less filename

~$ less Drerio_GRCz11_partial.gtf

less | reading, scrolling and search

- Page down / up:
  - <Space> (down), b (up)

- Go to end / start of file:
  - G (end), g (start)

- Forward search for a string:
  - /string

Try to look for a gene "ENSDARG00000056498"

- Exit:
  - q

wc | count word, line, character

```
~$ wc Drerio_GRCz11_partial.gtf
```

```
   2500   82676  969320 Drerio_GRCz11_partial.gtf
```

```
# -l  Count lines
```

```
~$ wc -l Drerio_GRCz11_partial.gtf
```

```
   2500
```

head | display the first lines of a file or the standard input

~$ head -n 5 Drerio_GRCz11_partial.gtf

```
chr1  ensembl_havana  gene  27977297  28020042  . + . gene_id "ENSDARG00000100083"; gene_vers
chr1  havana  transcript  27984393  27995611  . + . gene_id "ENSDARG00000100083"; gene_versio
chr1  havana  exon  27984393  27984722  . + . gene_id "ENSDARG00000100083"; gene_version "2";
chr1  havana  exon  27984816  27984885  . + . gene_id "ENSDARG00000100083"; gene_version "2";
chr1  havana  exon  27993185  27993255  . + . gene_id "ENSDARG00000100083"; gene_version "2";
```

tail | display the last lines of a file or the standard input

cut | cut out selected portions of each line of a file

| | pipe the stdout of a program to a new program as stdin

~$ cut -f 1-5 Drerio_GRCz11_partial.gtf | head -n 5

```
chr1  ensembl_havana  gene  27977297  28020042
chr1  havana  transcript  27984393  27995611
chr1  havana  exon  27984393  27984722
chr1  havana  exon  27984816  27984885
chr1  havana  exon  27993185  27993255
```

sort | sort lines of text

uniq | report or filter out repeated lines

- Display each line once:
  sort stdin/file | uniq

- Display only unique lines:
  sort stdin/file | uniq -u

- Display only duplicate lines:
  sort stdin/file | uniq -d

- Display number of occurrences of each line along with that line:
  sort stdin/file | uniq -c

- Display number of occurrences of each line, sorted by the most frequent:
  sort stdin/file | uniq -c | sort -nr

How many chromosomes in the Drerio_GRCz11_partial.gtf?

~$ cut -f 1-5 Drerio_GRCz11_partial.gtf | head -n 5

```
chr1  ensembl_havana  gene  27977297  28020042
chr1  havana  transcript  27984393  27995611
chr1  havana  exon  27984393  27984722
chr1  havana  exon  27984816  27984885
chr1  havana  exon  27993185  27993255
```

~$ cut -f1 Drerio_GRCz11_partial.gtf | sort | uniq | wc -l

```
25
```

~$ cut -f1 Drerio_GRCz11_partial.gtf | sort | uniq

-V | Sort version numbers

~$ cut -f1 Drerio_GRCz11_partial.gtf | sort -V| uniq

```
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
chr25
```

grep | Find patterns in files using regular expressions

```
# -w | search “pattern” for as a word
# -i | ignore-case
# -v | invert-match
# -A [num] | Print [num] lines of trailing context after each match.
# -B [num] | Print [num] lines of leading context before each match.
# -C [num] | Print [num] lines of leading and trailing context surrounding each match.
```

~$ grep “gene” Drerio_GRCz11_partial.gtf | head -n 3

```
chr1  ensembl_havana  gene  27977297  28020042  . + . gene_id "ENSDARG00000100083"; gene_versi
chr1  havana  transcript  27984393  27995611  . + . gene_id "ENSDARG00000100083"; gene_versior
chr1  havana  exon  27984393  27984722  . + . gene_id "ENSDARG00000100083"; gene_version "2";
```

~$ grep -w “gene” Drerio_GRCz11_partial.gtf | head -n 3

```
chr1  ensembl_havana  gene  27977297  28020042  . + . gene_id "ENSDARG00000100083"; gene_versi
chr1  ensembl_havana  gene  38398696  38415158  . - . gene_id "ENSDARG00000012016"; gene_versi
chr2  ensembl_havana  gene  6042039 6051836 . - . gene_id "ENSDARG00000063341"; gene_version "
```

~$ grep -w -v “gene” Drerio_GRCz11_partial.gtf | head -n 3

```
chr1  havana  transcript  27984393  27995611  . + . gene_id "ENSDARG00000100083"; gene_versior
chr1  havana  exon  27984393  27984722  . + . gene_id "ENSDARG00000100083"; gene_version "2";
chr1  havana  exon  27984816  27984885  . + . gene_id "ENSDARG00000100083"; gene_version "2";
```

How many genes of each chromosome in the Drerio_GRCz11_partial.gtf?

```
~$ grep -w "gene" Drerio_GRCz11_partial.gtf | cut -f1 | sort -V | uniq -c
```

Plain Text

```
   2 chr1
   4 chr2
   5 chr3
   5 chr4
   4 chr5
   4 chr6
   4 chr7
```

```
~$ grep -w "gene" Drerio_GRCz11_partial.gtf | cut -f1 | sort -V | uniq -c > gene_num.txt
~$ ls -l
```

```
-rw-r--r--  1 Lu  staff  969320 Dec  5 15:00 Drerio_GRCz11_partial.gtf
-rw-r--r--  1 Lu  staff     266 Dec  5 18:00 gene_num.txt
```

awk | a pattern-directed scanning and processing language

Awk's basic syntax:

awk 'optional pattern {some instructions}' filename

-F | specify a field separator

NF | 每一行 ($0) 擁有的欄位總數

NR | 目前 awk 所處理的是『第幾行』資料

FS | 分隔字元，預設是空白鍵

~$ head -n3 Drerio_GRCz11_partial.gtf | awk '{print "line_" NR "\t" NF}'

```
line_1  18
line_2  28
line_3  34
```

~$ head -n3 Drerio_GRCz11_partial.gtf | awk -F "\t" '{print "line_" NR "\t" NF}'

```
line_1  9
line_2  9
line_3  9
```

~$ awk '{print $9}' Drerio_GRCz11_partial.gtf | head -n3

```
gene_id
gene_id
gene_id
```

~$ awk -F "\t" '{print $9}' Drerio_GRCz11_partial.gtf | head -n3

```
gene_id "ENSDARG00000100083"; gene_version "2"; gene_name "sugt1"; gene_source "ensembl_havana"; gene_bioty
gene_id "ENSDARG00000100083"; gene_version "2"; transcript_id "ENSDART00000171868"; transcript_version "2";
gene_id "ENSDARG00000100083"; gene_version "2"; transcript_id "ENSDART00000171868"; transcript_version "2";
```

~$ awk '$1 >= 5 {print $0}' gene_num.txt

>= 大於或等於

<= 小於或等於

== 等於

!= 不等於

```
5 chr3
5 chr4
5 chr9
7 chr16
7 chr24
```

~$ awk '{print $1, $3, $10}' Drerio_GRCz11_partial.gtf | head -n3

```
chr1 gene "ENSDARG00000100083";
chr1 transcript "ENSDARG00000100083";
chr1 exon "ENSDARG00000100083";
```

~$ awk '{OFS = "\t" ; print $1, $3, $10}' Drerio_GRCz11_partial.gtf | head -n3

```
chr1	gene	"ENSDARG00000100083";
chr1	transcript	"ENSDARG00000100083";
chr1	exon	"ENSDARG00000100083";
```

sed | stream editor

sed 's/ pattern / replacement /g'

~$ awk '{OFS = "\t" ; print $1, $3, $10}' Drerio_GRCz11_partial.gtf | head -n3

```
chr1  gene  "ENSDARG00000100083";
chr1  transcript  "ENSDARG00000100083";
chr1  exon  "ENSDARG00000100083";
```

~$ awk '{OFS="\t";print$1,$3,$10}' Drerio_GRCz11_partial.gtf|head -n3|sed 's/;//'|sed 's/"//'

```
chr1  gene  ENSDARG00000100083"
chr1  transcript  ENSDARG00000100083"
chr1  exon  ENSDARG00000100083"
```

~$ awk '{OFS="\t";print$1,$3,$10}' Drerio_GRCz11_partial.gtf|head -n3|sed 's/;//'|sed 's/"//g'

```
chr1  gene  ENSDARG00000100083
chr1  transcript  ENSDARG00000100083
chr1  exon  ENSDARG00000100083
```

Please make a bed file using Drerio_GRCz11_partial.gtf

~$ head -n5 Drerio_GRCz11_partial.gtf

```
chr1  ensembl_havana  gene  27977297  28020042  . + . gene_id "ENSDARG00000100083"; gene_version "2"; gene_n
chr1  havana  transcript  27984393  27995611  . + . gene_id "ENSDARG00000100083"; gene_version "2"; transcri
chr1  havana  exon  27984393  27984722  . + . gene_id "ENSDARG00000100083"; gene_version "2"; transcript_id
chr1  havana  exon  27984816  27984885  . + . gene_id "ENSDARG00000100083"; gene_version "2"; transcript_id
chr1  havana  exon  27993185  27993255  . + . gene_id "ENSDARG00000100083"; gene_version "2"; transcript_id
```

```
Typical 6-fields bed format
chrom    chromStart    chromEnd    GeneID    score   strand

chr1  27977297  28020042  . ENSDARG00000100083  +
chr1  38398696  38415158  . ENSDARG00000012016  -
chr2  6042039 6051836 . ENSDARG00000063341  -
```

~$ awk '$3=="gene" {OFS="\t"; print $1, $4, $5, ".", $10, $7}' Drerio_GRCz11_partial.gtf | sed s/\"//g | sed s/\;//g | head -n 3

Regular expressions

^	行首
$	行尾
.	任意一個字元
*	重複字元
\w	[a-zA-Z0-9_]
\d	[0-9]
\s	空白
\t	tab
\n	新行

~$ grep "3$" gene_num.txt

```
5 chr3
3 chr13
4 chr23
```

~$ grep "chr2\d" gene_num.txt

```
3 chr20
3 chr21
3 chr22
4 chr23
7 chr24
2 chr25
```