ISCB20.01 – Introduction to LINUX for Biologists

Md. Jubayer Hossain
https://jhossain.me/

November 23, 2020

Lead Organizer, Introduction to Scientific Computing for Biologists
Founder, Health Data Research Organization

HEALTH DATA
RESEARCH ORGANIZATION

Section–3: UNIX Commands for Data Manipulation

Files, Directories, Directory Structure, Paths

- / $\rightarrow$ Root
- cd $\rightarrow$ Change directory
- cd .. $\rightarrow$ Moves one directory up
- pwd $\rightarrow$ Present working directory
- ls $\rightarrow$ List of content of resent working directory
- ls-l $\rightarrow$ Similar as ls, but provides additional info on files and directories
- ls -la $\rightarrow$ Includes hidden files (.name) as well

Regular Expressions(File Naming Patterns)

- * $\rightarrow$ means all
- ? $\rightarrow$ finding pattern
- [] $\rightarrow$ finding pattern
- {} $\rightarrow$ finding pattern

Files

- touch [filename.extention(.txt,.csv, .tsv etc)] $\rightarrow$ create file
- nano [filename.extension] $\rightarrow$ edit file with nano text editor
- cp [path] $\rightarrow$ copy files/directories
- mv [path] $\rightarrow$ move files/directories
- rm [filename.extension] $\rightarrow$ remove file(s)

Directories

- mkdir [dirname] $\rightarrow$ make directory
- cp [path] $\rightarrow$ copy files/directories
- mv [path] $\rightarrow$ move files/directories
- rmdir [dirname] $\rightarrow$ remove empty directory
- rm -r [dirname] $\rightarrow$ remove directory with content
- rm -rf [dirname] $\rightarrow$ remove directory with content

- echo text $\rightarrow$ print text/string
- cat [filename] $\rightarrow$ concatenate file/print content
- head [filename] $\rightarrow$ default displays the first 10 lines
- head -n [filename] $\rightarrow$ displays the firth nth number of lines
- tail [filename] $\rightarrow$ default displays the last 10 lines
- tail -n [filename] $\rightarrow$ displays the last nth number of lines

- more [filename] $\rightarrow$ Viewing content
- less [filename] $\rightarrow$ Scroll through a file using arrow keys or(spacebar = advance page | b = reverse page | q = quit)

Standard Files

- $<$ → standard input(stdin)
- $>$ → standard output(stdout)
- Pipe(|) → pipe is a form of redirection (transfer of standard output to some other destination

- grep "pattern" filename → search a pattern
- sort [file] → sort files(alphabetically)
- uniq [file] → display unique lines
- cut [file] → break files vertically based on fields

- diff → display difference
- comm → display common lines among files

Compress

- zip output.zip inputfile.extension $\rightarrow$ zip files
- zip -r outputdir.zip directory $\rightarrow$ zip directories
- gzip files $\rightarrow$ gzip files
- tar $\rightarrow$ archive and compress files/directories
- tar -czvf output.tar.gz directory $\rightarrow$ compress

Decompress

- unzip dirname.zip $\rightarrow$ decompress zipped file
- gunzip dirname.gz $\rightarrow$ decompress gzipped files
- tar -xzvf dirname.tar.gz $\rightarrow$ extract

A Case Study

- How many chromosomes are there in the genome?
- How many genes and transcript variants?
  - How many genes have a single variant?
  - How many genes have a multiple variant?
- How many genes are there on each of the '+' and '-' strands?
- How many genes(and transcripts) are there on each chromosome?

Thank You

:)