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

Section–3: UNIX Commands for Data Manipulation

Content Representation–1

Files, Directories, Directory Structure, Paths

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

Content Representation–2

Regular Expressions(File Naming Patterns)

- An asterisk (*) – matches one or more occurrences of any character, including no character.
- Question mark (?) – represents or matches a single occurrence of any character.
- Bracketed characters ([ ]) – matches any occurrence of character enclosed in the square brackets. It is possible to use different types of characters (alphanumeric characters): numbers, letters, other special characters etc.
- Curly brackets ({}) – Terms are separated by commas and each term must be the name of something or a wildcard. This wildcard will copy anything that matches either wildcard(s), or exact name(s) (an "or" relationship, one or the other).

Content Creation and Removal–1

Files

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

Content Creation and Removal–2

Directories

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

Accessing Content–1

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

Accessing Content–2

- more [filename] → Viewing content
- less [filename] → Scroll through a file using arrow keys or(spacebar = advance page | b = reverse page | q = quit)
- wc → word count

Redirecting Content

Standard Files

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

Querying Content

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

Comparing Content

- diff $\rightarrow$ display difference
- comm $\rightarrow$ display common lines among files

Archiving Content–1

Compress

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

Archiving Content–2

Decompress

- unzip dirname.zip → decompress zipped file
- gunzip dirname.gz → decompress gzipped files
- tar -xzvf dirname.tar.gz → 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