Unix Tutorial for Biologists

Common Commands
Nathan Dunn
CASSPR (ndunn@cas.uoregon.edu)
Overview

- Command-Line
- Compressing
- Searching
- IO
- Filtering
- Moving
- Advanced Search
Command-Line

- `echo $SHELL`
  - should say `/bin/bash`
- `chsh`
- `cntrl-R` is a recursive command search
- `up-arrow`
Compressing: formats

• tar + gzip or bzip2
• zip
Compressing: tar

- tar = <filename>.tar
  - create: tar cvf <filename>.tar <files>
  - extract: tar xvf <filename>.tar
  - view: tar tvf <filename>.tar
Compressing: tar+gzip

- tar+gzip = <filename>.tgz, <filename>.tar.gz
  - tar cvfz <filename>.tgz <files>
  - tar xvfz <filename>.tgz
  - tar tvfz <filename>.tgz
- gzip = <filename>.gz
  - gzip <filename>.gz <files> #create
  - gunzip <filename>.gz #extract
Compressing: tar+bzip2

- tar+bzip2 = <filename>.tar.bz2
  - tar cvfj <filename>.tar.bz2 <files>
  - tar xvfj <filename>.tar.bz2
  - tar tvfj <filename>.tar.bz2
- bunzip2 = <filename>.bz2
  - bzip2 <filename>.bz2 <files>
  - bunzip2 <filename>.bz2
Compressing: ??

- **ls -lh**

  ```
  NathanDunn:unix-biologists% ls -lh
  total 157640
  -rw-r--r-- 1 NathanDunn staff 29M May 14 13:14 small_sample_1.fq
  -rw-r--r-- 1 NathanDunn staff 8.5M May 14 13:15 small_sample_1.tgz
  -rw-r--r-- 1 NathanDunn staff 29M May 14 13:14 small_sample_2.fq
  -rw-r--r-- 1 NathanDunn staff 8.4M May 14 13:15 small_sample_2.tgz
  ```

- **du -sh**

  ```
  NathanDunn:unix-biologists% du -sh small_sample_*
  29M  small_sample_1.fq
  8.5M  small_sample_1.tgz
  29M  small_sample_2.fq
  8.4M  small_sample_2.tgz
  ```

- **df -h**

  ```
  NathanDunn:unix-biologists% df -h
  Filesystem   Size  Used  Avail Capacity Mounted on
  /dev/disk0s2  931G  227G  703G  25%   /
  devfs        194K   194K    0B   100%   /dev
  ```

- **wc -l <files>**

  ```
  NathanDunn:unix-biologists% wc -l small_sample_*.fq
  1000000 small_sample_1.fq
  1000000 small_sample_2.fq
  2000000 total
  ```
Searching: find

- find <directory> -name <pattern>
- find . -name uogen_present.key
- find . -name \*.key  # look for suffix

```
NathanDunn@casspr% find . -name \*.key
/presentations/unix-biologists/week1-presentation.key
/presentations/unix-biologists/week2-presentation.key
/presentations/unix-biologists/week3-presentation.key
/presentations/uogenesis/uogen_present.key
NathanDunn@casspr% 
```

- find / -name \*.key  # look everywhere!
Searching: grep

• Search file for pattern
  
grep <pattern> <where>

• grep -l <pattern> <where>
  • just list the name

• grep -r <pattern> <where>
  • search recursively

```
NathanDunn presentations% grep -rl "@IRIS:7:1:18:1260#0/1" *
unix-biologists/small_sample_1.fq
```
Searching: grep

• `grep -irl <pattern> <where>`
  • case insensitive, recursive, list only name

• `grep -v <pattern> <where>`
  • exclude matching pattern!!
Searching: head

- head -<N> <file>

- show me the first N lines of <file>

```
NathanDunn:unix-biologists% head -3 small_sample_1.fq
@IRIS:7:1:17:394#0/1
+TCAGGACAAGAAAGACAANTCCATATTATG
@IRIS:7:1:17:394#0/1
```
Searching: tail

- tail -<N> <file>
  - show me the last N lines of <file>

NathanDunn:unix-biologists% tail -3 small_sample_1.fq
GAATGATAGTNCTTGGTGGTTATGCAAAGAAAAAT
+IRIS:7:9:317:1519#0/1
aaaab_baaa^[D[a``Ya/X^]\_a_]^a`a_a

- tail -f <file>
  - show me output of <file> as it changes
Searching: less

- less filename
- ‘/’ look for expression
  - ‘n’ = next, ‘shift-n’ = previous
Searching: less

- page
  - down: ‘space’, ‘page-down’, ‘cntrl-d’
  - up: ‘page-up’, ‘cntrl-u’
- all the way down: ‘shift-g’
- all the way up: ‘g’
- ‘left/right-arrow’ for long text
Searching: gzipped

• Same operations on gzipped files:
  • zless
  • zcat
  • zgrep
IO: Pipes

- `<command A> | <command B>`
- `# the output of command A is the input to command B`
- `grep IRIS small_sample_1.fq | less`

```
@IRIS:7:1:17:394#0/1
+IRIS:7:1:17:394#0/1
@IRIS:7:1:17:800#0/1
+IRIS:7:1:17:800#0/1
@IRIS:7:1:17:1757#0/1
+IRIS:7:1:17:1757#0/1
```
IO: > >>, <

- ‘>’ write command to file, rewriting
  - `ls -l > file.txt`
- ‘>>>’ write command to file, appending
  - `ls -l >>> file.txt`
- ‘<‘ file becomes input to command
  - `grep IRIS < small_sample_1.fq`
IO: cat

• cat <filename>
  • dumps everything to standard output
• cat small_sample_1.fq small_sample_2.fq | grep +IRIS > iris_ids.txt
Filtering:

- `cut -d<delimiter> -f<column> -s` small_sample_1.fq
- `cut -d<delimiter> -f<column> -s` small_sample_1.fq
- `cut -c<start-stop>`, e.g., `cut -c11-12 small_sample_1.fq`
Filtering: sort

- sort <filename>
- sort small_sample_1.fq | less
- grep +IRIS small_sample_1.fq | sort | less
Filtering: uniq

- `uniq <filename>`
- `dumps out unique output`
- `cut -d: -f4 -s small_sample_1.fq | wc -l 500000`
- `cut -d: -f4 -s small_sample_1.fq | uniq | wc -l 500000`
Filtering: diff

- diff <file1> <file2>
- cut -d: -f4 -s small_sample_1.fq | uniq > out1.txt
- cut -d: -f4 -s small_sample_2.fq | uniq > out2.txt
- diff out1.txt out2.txt
- diff small_sample_1.fq small_sample2.fq
Moving: scp

• scp <filename>  
  <user>@<host>:<directory>

• copy files TO home directory on aciss

```bash
NathanDunn:unix-biologists% scp iris_ids* ndunn@aciss.uoregon.edu:
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
iris_ids_sorted.txt 100% 11MB 10.9MB/s 00:01
```

• copy files TO “/tmp” on aciss

```bash
NathanDunn:unix-biologists% scp iris_ids* ndunn@aciss.uoregon.edu:/tmp
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
iris_ids_sorted.txt 100% 11MB 10.9MB/s 00:01
```
Moving: scp

- scp <user>@<host>::<directory / file> <location / filename>
- copy files FROM home directory on aciss
- copy files FROM “/tmp” on aciss
Moving: sftp

- `sftp <user>@<host>`
- similar but interactive

```
NathanDunn@unix-biologists% sftp ndunn@aciss.uoregon.edu
ndunn@aciss.uoregon.edu's password:
Connected to aciss.uoregon.edu.
sftp> ls
Downloads    blast_test.e100277 blast_test.e100924 blast_test.e100925
blast_test.o100277 blast_test.o100924 blast_test.o100925    hg
iris_ids.txt    iris_ids_sorted.txt    old.local
sftp> put iris_ids*
Uploading iris_ids.txt to /ibrix/home2/ndunn/iris_ids.txt
100% 11MB 10.9MB/s 00:01
Uploading iris_ids_sorted.txt to /ibrix/home2/ndunn/iris_ids_sorted.txt
100% 11MB 10.9MB/s 00:01
```
Moving: wget

- downloads files + web pages
- `wget <url>`

   => `allcontig.agp.gz.1'
Resolving ftp.ncbi.nih.gov... 130.14.250.11
Logging in as anonymous ... Logged in!
==> SYST ... done.   ==> PWD ... done.
==> TYPE I ... done. ==> CWD (1) /genomes/D_rerio ... done.
==> SIZE allcontig.agp.gz ... 1486337
==> PASV ... done.   ==> RETR allcontig.agp.gz ... done.
Length: 1486337 (1.4M) (unauthoritative)

100%[================================================================================================]=> 1,486,337 197K/s in 6.3s

2012-05-14 15:20:47 (231 KB/s) - `allcontig.agp.gz.1' saved [1486337]
Moving: curl

- downloads files
- `curl -o <output> <url>`
Advanced Search: mdfind

- mac only ... but fast ... can search in
- mdfind -name <pattern>
- mdfind <pattern>
- mdfind -name .key | grep key$

NathanDunn@casspr: % mdfind -name ".key" | grep key$
/Users/NathanDunn/Library/Preferences/IntelliJ IDEA 11/idea11.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week1-presentation.key
/Users/NathanDunn/hg/casspr/presentations/uogenetics/uogen_present.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week3-presentation.key
Advanced Search: locate

- Similar to mdfind, but on Linux
- Will need to initialize database if not done
Advanced Search

• sed: stream editor
  • sed 's/\|/,/g' output/phenotypes.txt > output/phenotypes.csv

• awk
Other Resources

• Do it!
• Google it!
• Ask!
• Conery Class: CIS 170
Next Time

• Working examples

• Bash scripting