Software Tools for Scientific Research

Nathan Dunn
ndunn@cas.uoregon.edu
Craig Rasmussen
rasmus@cas.uoregon.edu
Kristy Potter
kpotter@uoregon.edu
Overview

• Introductions
• Todo
• Syllabus
• A working environment
• ACISS
• Unix Commands
Introductions

• About you
  • Who are you?
  • What are you looking to do?
Requirements

- A laptop with a working connection to the UO network
- No prior computer science coursework or programming knowledge
Todo

- Signup form going around
- Get a laptop with UO network access
- Working Linux / Unix Environment
- [https://aciss.uoregon.edu/newuser](https://aciss.uoregon.edu/newuser)
- Add note / email if you want Cloud: Robert Yelle ryelle@uoregon.edu
Schedule

• We will meet weekly for 90 minutes, but will be available by appointment.

• We will reserve time after each class for Q & A.
Syllabus

- Work environment
- Common Unix Commands: Compressing Files, Searching, Pipes, IO, Filtering, Moving Files Remotely, Advanced Search
- Practical bash scripting
- Source control
- Python / Matlab programming
- ACISS: Cluster and Cloud
Syllabus

- Scientific visualization
- Visualization tools: matplotlib, google charts, 3d
- Parallel Programming: task and data parallelism
- Parallel iPython and ACISS, Matlab
- Student Project Presentations
A Working Environment

- Shell / Terminal
- Linux / Mac installed
- Windows a few options
**Mac Terminal / X11**

---

The queues are subject to change depending on needs. Let us know what we can do to better meet your needs.

**NEWS:**
- 1/31/2012: The OS has been upgraded on all ACISS nodes.
- 3/27/2012: Latest OS and driver updates applied across all nodes.
- 3/28/2012: Cloud images rolled out across generic nodes.
- Torque upgraded to 2.5.11.
- 4/23/2012: 1 TB drives added to some fatnodes, mounted as /scratch. Use `-q fatnodes -l nodes=1:scratch` to select fatnodes with extra space on scratch.

***Students and first time users, please read the following file for a brief introduction to using ACISS: /INFO/first-time-users.txt***

***Do not run jobs on the head node, please use the queuing system!***

If you have any questions or need any assistance, please send email to systems@aciss.uoregon.edu

[nadunn@hn1 ~]$
Windows Terminal Options

- VirtualBox
- Putty -> ssh client to unix server
- Cygwin -> emulates Linux natively on windows
Windows - Terminal
VirtualBox

- VirtualBox
- free virtualization server
- install application -> install instance
Windows Terminal - Putty

- Putty
- ssh into remote unix server like ACISS
Unix Basics

• File structure
• Navigation
• File manipulation
• Editing
File Structure

```
/  
  root

Users
  keith

Applications
  nigel

Volumes
  USB
  Mac

Unix and Perl course

Applications
  Code
  Data
  Docs
```
Week 1 File Structure
File Structure

- `ls`: list files / directories
- `pwd`: present working directory

```
NathanDunn:folder1% ls
  dirt  rocks
NathanDunn:folder1% pwd
/Users/NathanDunn/hg/casspr/presentations/unix-geophysicists-summer-2012/week1_files/start/folder1
NathanDunn:folder1%
```
Navigation

- cd: change directory
- .. = up
- . = here
Options

• `ls -ltrc` : list in reverse chronological order

• Help:
  • `man <command>`
  • `<command> -help -h --h --help`
  • `google command` (several variants)
Create Directory

- `mkdir <directory>`
Remove File / Directory

- `rm <file / directory>`
- `rm -rf <file / directory>`
  - `-r = recursive , -f = force`

```
NathanDunn:rocks% mkdir kid
NathanDunn:rocks% ls
chemical.txt  igneous  metamorphic  sedimentary
clastic.txt  kid  organic.txt
NathanDunn:rocks% rm -rf kid
NathanDunn:rocks% ls
chemical.txt  clastic.txt  igneous  metamorphic  organic.txt  sedimentary
NathanDunn:rocks%
```
Dump File

- cat <filename>

NathanDunn@rocks% cat sedimentary/organic.txt
Organic: any accumulation of sedimentary debris caused by organic processes. Many animals use calcium for shells, bones, and teeth. These bits of calcium can pile up on the seafloor and accumulate into a thick enough layer to form an "organic" sedimentary roc
Move Files

• `mv <A> <B>` : move files (and directories)

• also renames

```
NathanDunn@rocks% ls
chemical.txt igneous organic.txt
clastic.txt metamorphic sedimentary
NathanDunn@rocks% mv chemical.txt sedimentary
NathanDunn@rocks% ls
clastic.txt metamorphic sedimentary
igneous organic.txt
NathanDunn@rocks% mv *.txt sedimentary
NathanDunn@rocks% ls
igneous metamorphic sedimentary
NathanDunn@rocks% ls sedimentary
chemical.txt clastic.txt organic.txt
```
Copy File / Directories

- `cp <A> <B>`: copy files
- `cp -r <A> <B>`: copy directory

```
NathanDunn:rocks% ls
igneous    metamorphic    sedimentary
NathanDunn:rocks% cp sedimentary/organic.txt igneous/basalt.txt
NathanDunn:rocks% ls igneous
basalt.txt
NathanDunn:rocks% cat igneous/basalt.txt
Organic: any accumulation of sedimentary debris caused by organic processes. Many animals use calcium for shells, bones, and teeth. These bits of calcium can pile up on the seafloor and accumulate into a thick enough layer to form an "organic" sedimentary rock.
```
Create Files

- `touch <filename>`
Browse File: less

- `less <filename>`: view files
- ‘q’ is quit
- ‘space’ to move ahead
- ‘/’ to search ‘n’ for next, ‘shift-n’ for previous
- ‘g’ go to top
- `more <filename>`: same idea
Edit File(s)

- `nano <filename>`: edit files
- instructions listed
- `^` is the control key

```
NathanDunn:rocks% nano igneous/granite.txt
NathanDunn:rocks% 
```

```
GNU nano 2.0.6 File: igneous/granite.txt Modified

Granite is a very hard rock.
```

```
^G Get Help  ^D WriteOut  ^R Read File  ^Y Prev Page  ^K Cut Text  ^C Cur Pos
^X Exit    ^J Justify   ^W Where Is  ^V Next Page  ^U UnCut Text  ^T To Spell
```
Command 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`

- `du -sh`

- `df -h`

- `wc -l <files>`
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-presentations.key
/presentations/unix-biologists/week2-presentation.key
/presentations/unix-biologists/week3-presentation.key
/presentations/uogenesis/uogen_present.key
```

- 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!!

GTCAGGACAAGAAAGACAANTCCACATTATG
+IRIS:7:1:17:394#0/1
aaabaa`]baaaaa_aab]D^^`b`aYDW]abaa`^GGAAACACTACTAGGCTTATAAGATCGGGGTTGCGG
+IRIS:7:1:17:800#0/1..
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
+TCAGGACAAGAAGACAANTCCATTTATG
+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
GAATGTATAGTNCTGGTTGATGCAAGAAAAAT
+IRIS:7:9:317:1519#0/1
aaaaab_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: > >>, <

- ‘>’ 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: 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: cat

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

- `cut -c<start-stop>`
- `cut -c11-12 small_sample_1.fq`
- `cut -d<delimiter> -f<column> -s`
- `cut -d: -f4 -s 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

```
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

```
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

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

- copy files FROM “/tmp” on aciss

```
NathanDunn:unix-biologists% scp ndunn@aciss.uoregon.edu:/tmp/iris_ids.txt .
ndunn@aciss.uoregon.edu's password: 
iris_ids.txt 100% 11MB 10.9MB/s 00:01
```
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.o100277    blast_test.o100924    blast_test.o100925
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/IntelliJIdea11/idea11.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week1-presentations.key
/Users/NathanDunn/hg/casspr/presentations/uogenESIS/uogen_present.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week3-presentation.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week4-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week5-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week6-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week7-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week8-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week9-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week10-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week11-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week12-presentations.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week13-presentations.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`
Summary

• Work on your TODO list!
  • Get a working Linux / UNIX environment
• Review older class files
• Bring in / email us your (technical) problems
• Next time - Bash scripts