IOWA STATE UNIVERSITY

Genome Informatics Facility

ISUGIF

Genome Informatics at Iowa State University

Transforming raw data into informative data for researchers

Andrew Severin
Genome Informatics Facility
Iowa State University

206 Science I
Iowa State University

What do I do?

- Enable researchers to interpret high-throughput data
- Encourage/Contribute to bioinformatics friendly infrastructure
- Explore the latest open source software
- Develop pipelines for efficient analysis
- Contribute to papers and grants (LOS)
- Train and Teach Bioinformatics
- Write Grants

Meet the needs and communicate well with a diverse faculty to help facilitate NGS research on campus.

GIF Team

Arun Seetharam

Usha Muppirala

Margie Carter

Iowa State University

Genome Informatics Facility

ISUGIF

Workshop: Basic UNIX for Biologists

Have you ever….

- tried to open a very large file (like FASTQ) in PC/Mac?
- searching for a specific piece of inform tion from large number of files?
- wanted to renam
- combine lar ?
- got frus favorite gene?
- wanted to run a ogram that isn’t available on your operating system?
- bored of doing same things over and over?

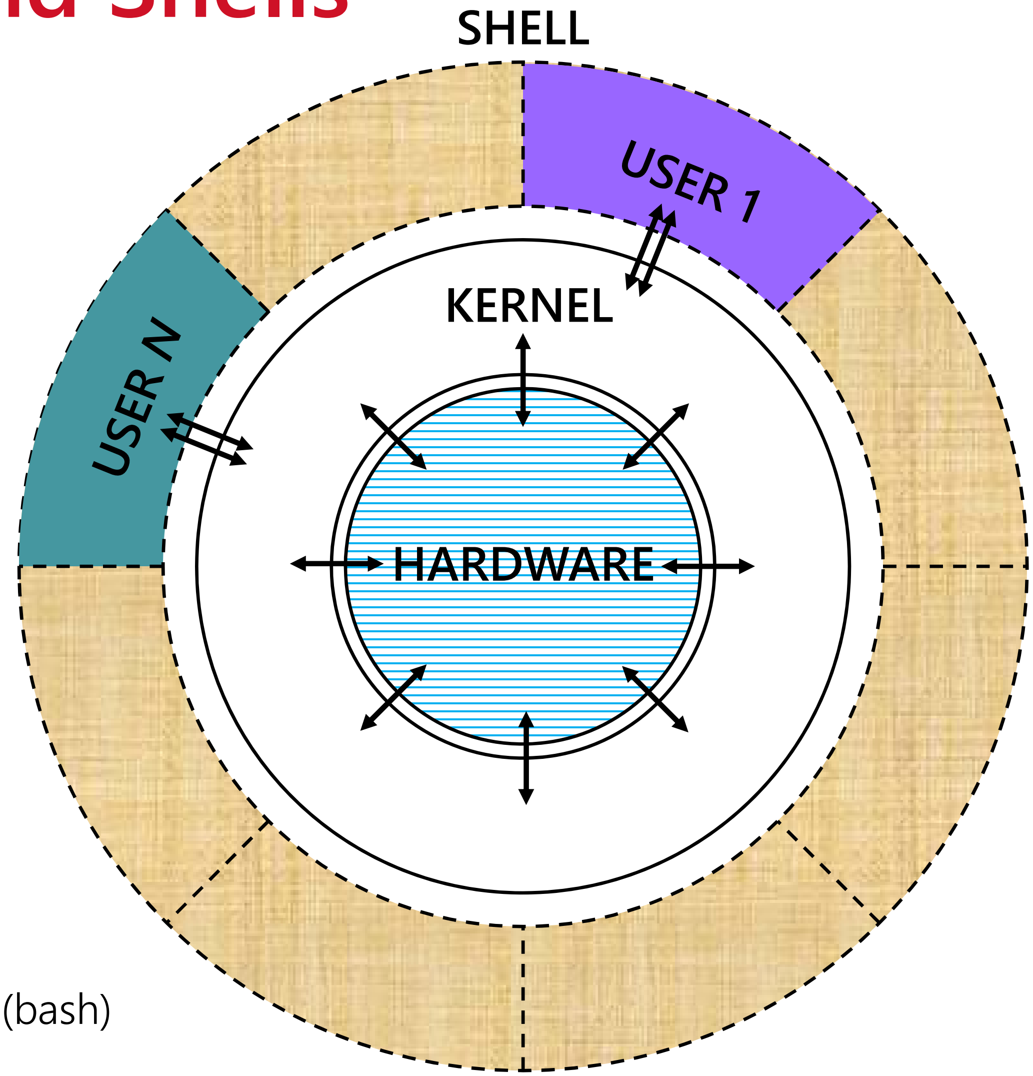
What is UNIX?

- Widely used multiuser operating system
- Linux: free version of UNIX-like operating system
  - Red Hat Enterprise Linux, Ubuntu, and CentOS
- Used on high-end workstations, database servers, web servers and managing shared resources
- Standard features include:
  - Security, reliability, scalability
  - supports multi-user (as in 100s)

Can I learn UNIX?

- Yes! Absolutely. Anyone can if they want.
- No more difficult than learning Word, Excel or Powerpoint
- Biggest difference
  - In Unix: You type the command to execute
  - In Word: You use your mouse to execute a command
  - Remember. In the Terminal, “don’t touch the mouse”

Kernel and Shells



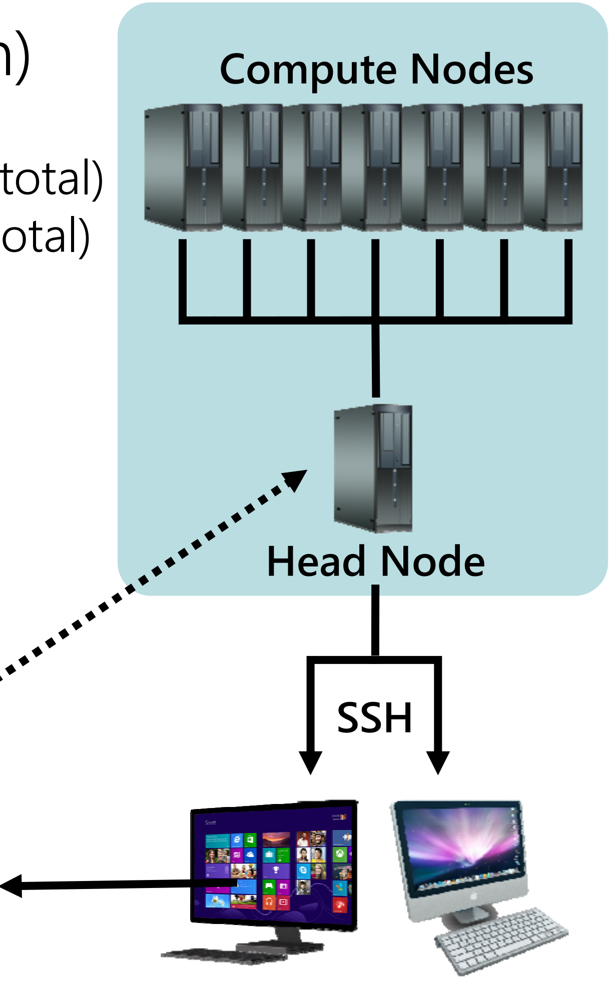
SHELL types:

- Bourne Shell (sh)
- C shell (csh)
- TC shell (tcsh)
- Korn shell (ksh)
- Bourne Again Shell (bash)
- Z shell (zsh)

High Performance Computing (HPC)

HPC-class (education)

- 48 compute nodes
- 16 processors/node (768 total)
- 64Gb RAM/node (3.0Tb total)



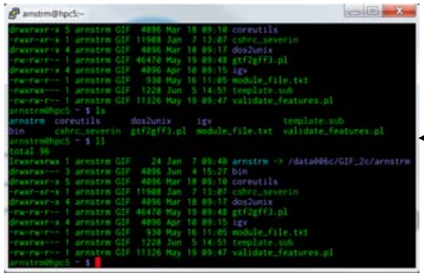
Lightning3 (research)

- 17 compute nodes
- 32 processors/node (384 total)
- 256Gb RAM/node (3.2Tb total)

Condo (up coming)

- 316 compute nodes
- 8 processors/node (2528 total)
- 8Gb RAM/node (3.5Tb total)

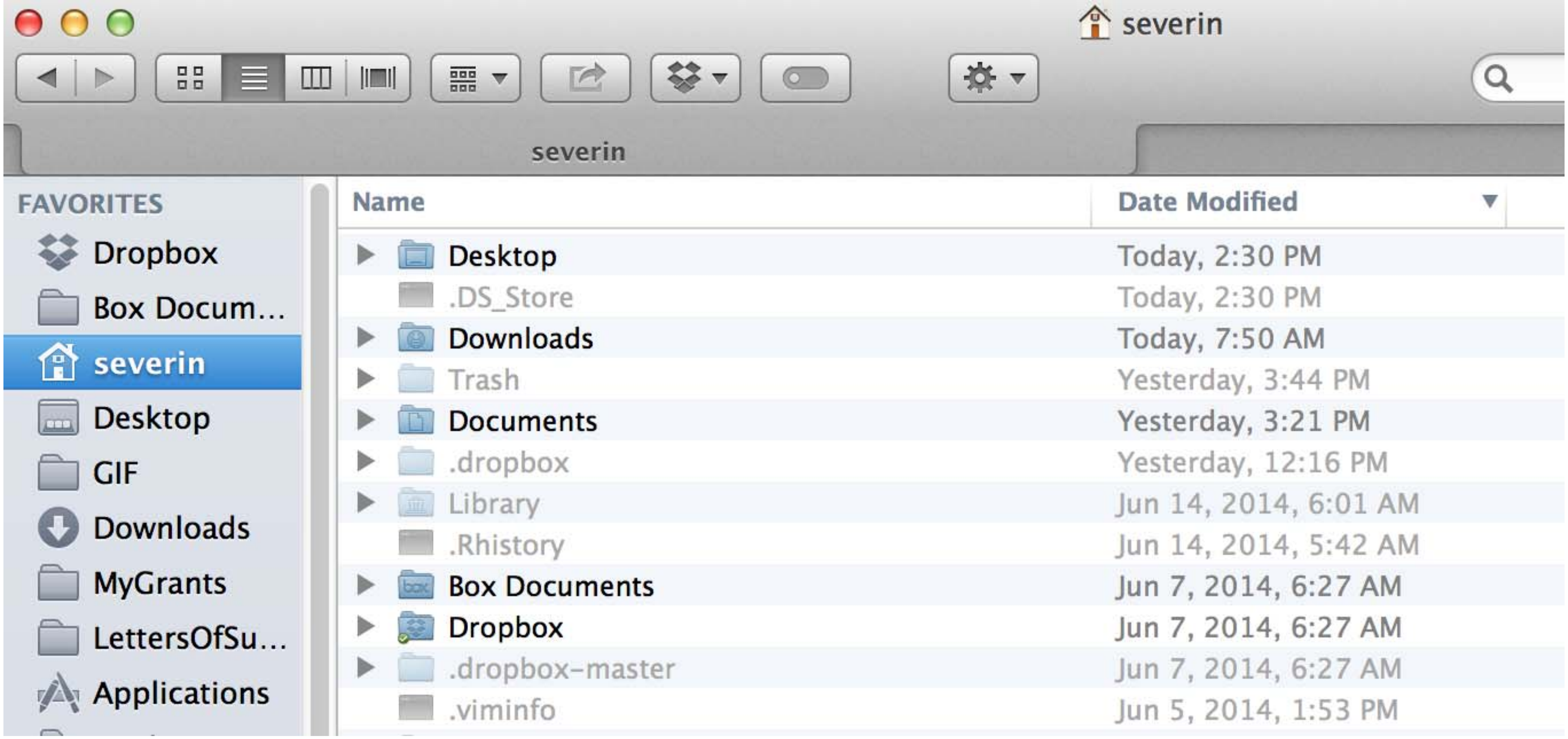
Terminal



```
arnstrm@hpc5:~
drwxrwxr-x 5 arnstrm GIF  4096 Mar 18 09:10 coreutils
-rwxr-xr-x 1 arnstrm GIF 11908 Jan  7 13:07 cshrc_severin
drwxrwxr-x 4 arnstrm GIF  4096 Mar 18 09:17 dos2unix
-rw-rw-r-- 1 arnstrm GIF 46470 May 19 09:48 gtf2gff3.pl
drwxrwxr-x 4 arnstrm GIF  4096 Apr 10 09:15 igv
-rw-rw-r-- 1 arnstrm GIF   930 May 16 11:05 module_file.txt
-rwxrwx--- 1 arnstrm GIF  1228 Jun  5 14:51 template.sub
-rw-rw-r-- 1 arnstrm GIF 11326 May 19 09:47 validate_features.pl
arnstrm@hpc5 ~ $ ls
arnstrm  coreutils      dos2unix     igv              template.sub
bin      cshrc_severin  gtf2gff3.pl  module_file.txt  validate_features.pl
arnstrm@hpc5 ~ $ ll
total 96
lrwxrwxrwx 1 arnstrm GIF    24 Jan  7 09:40 arnstrm -> /data006c/GIF_2c/arnstrm
drwxrwx--- 3 arnstrm GIF  4096 Jun  4 15:27 bin
drwxrwxr-x 5 arnstrm GIF  4096 Mar 18 09:10 coreutils
-rwxr-xr-x 1 arnstrm GIF 11908 Jan  7 13:07 cshrc_severin
drwxrwxr-x 4 arnstrm GIF  4096 Mar 18 09:17 dos2unix
-rw-rw-r-- 1 arnstrm GIF 46470 May 19 09:48 gtf2gff3.pl
drwxrwxr-x 4 arnstrm GIF  4096 Apr 10 09:15 igv
-rw-rw-r-- 1 arnstrm GIF   930 May 16 11:05 module_file.txt
-rwxrwx--- 1 arnstrm GIF  1228 Jun  5 14:51 template.sub
-rw-rw-r-- 1 arnstrm GIF 11326 May 19 09:47 validate_features.pl
arnstrm@hpc5 ~ $
```

Using Linux

- Linux, Mac OS X, Solaris, Cygwin
  - Many variants, same commands
  - I will refer to them all as Linux
- All Macs have Linux under the hood (Darwin)
  - Finder search for X11 or terminal
- Windows requires an emulator (Putty)

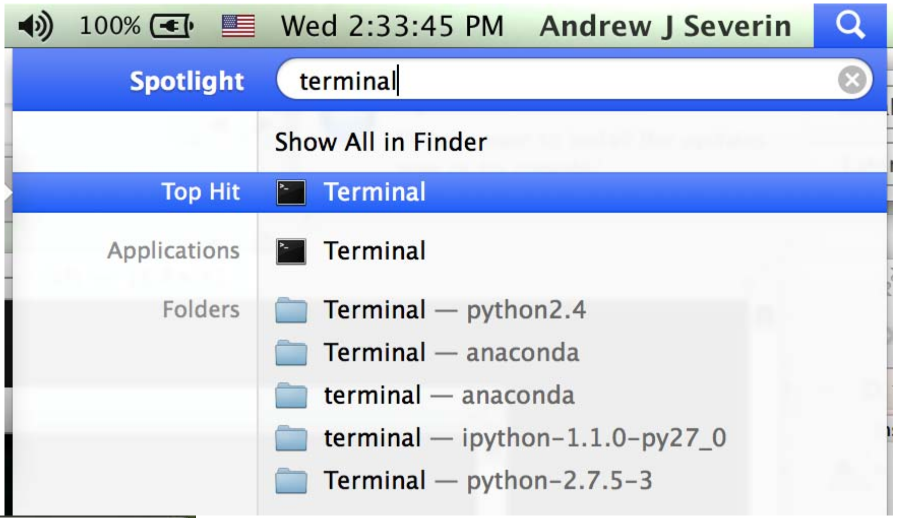


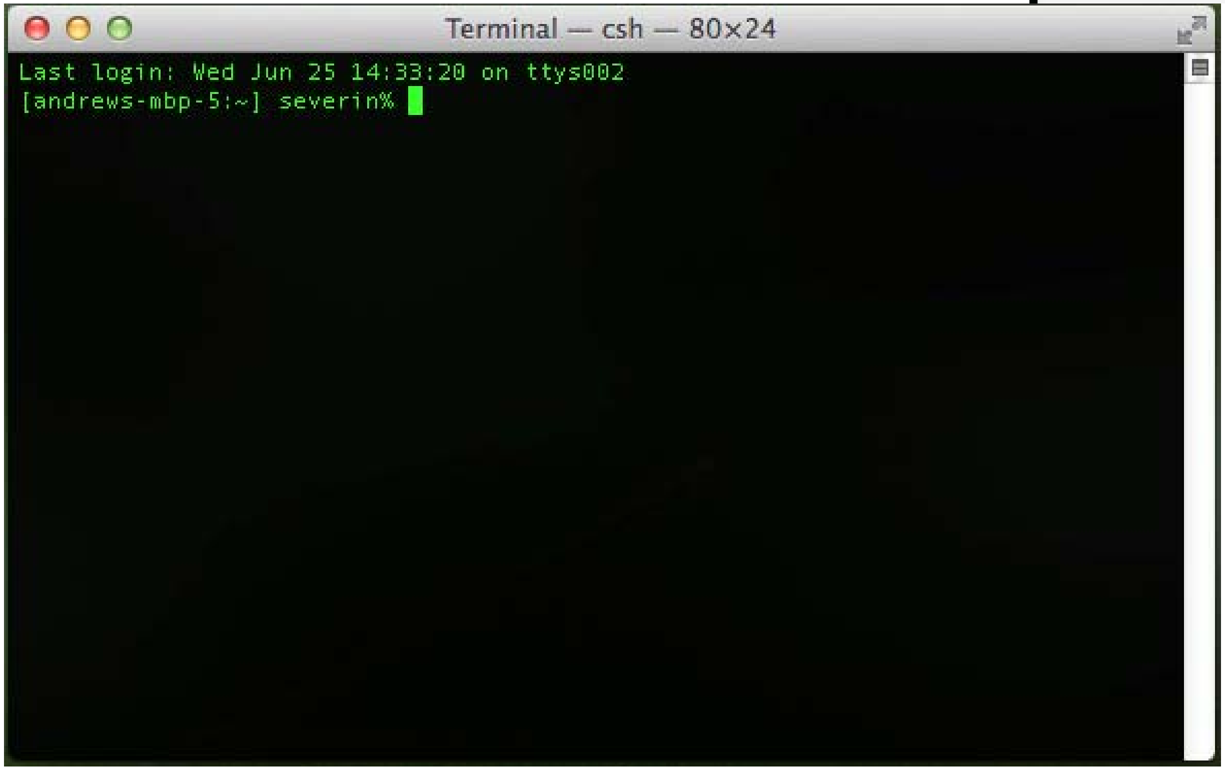
severin

FAVORITES

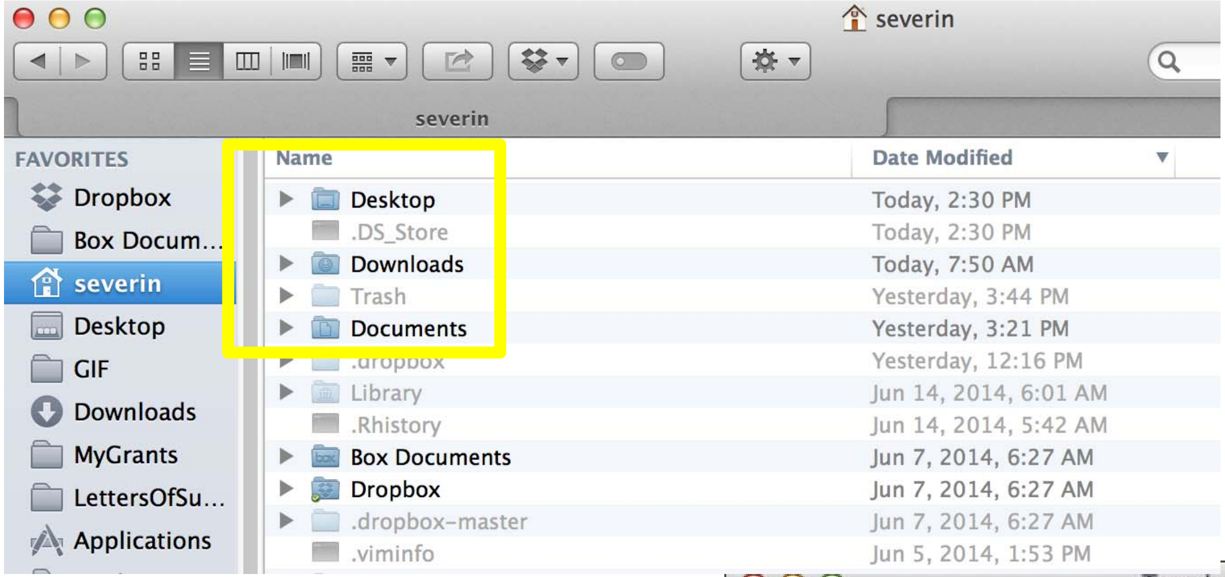
- Dropbox
- Box Docum…
- severin
- Desktop
- GIF
- Downloads
- MyGrants
- LettersOfSu…
- Applications

Name	Date Modified
▶ Desktop	Today, 2:30 PM
.DS_Store	Today, 2:30 PM
▶ Downloads	Today, 7:50 AM
▶ Trash	Yesterday, 3:44 PM
▶ Documents	Yesterday, 3:21 PM
▶ .dropbox	Yesterday, 12:16 PM
▶ Library	Jun 14, 2014, 6:01 AM
.Rhistory	Jun 14, 2014, 5:42 AM
▶ Box Documents	Jun 7, 2014, 6:27 AM
▶ Dropbox	Jun 7, 2014, 6:27 AM
▶ .dropbox-master	Jun 7, 2014, 6:27 AM
.viminfo	Jun 5, 2014, 1:53 PM

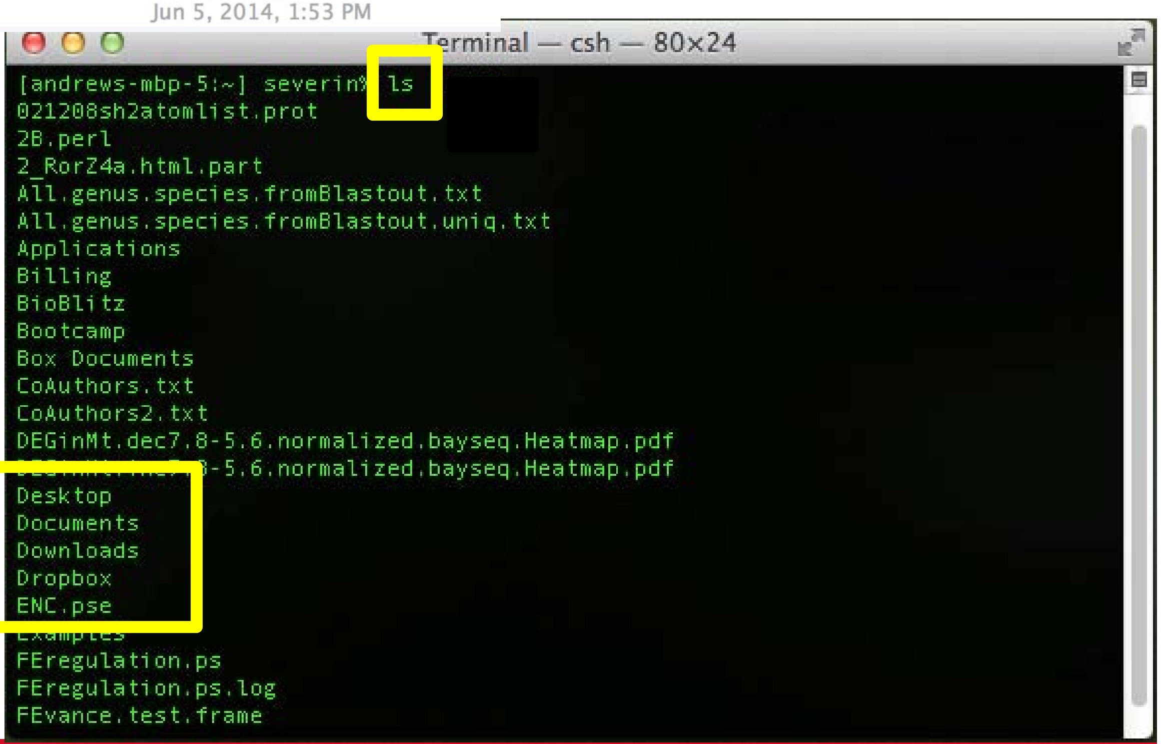




```
Last login: Wed Jun 25 14:33:20 on ttys002
[andrews-mbp-5:~] severin%
```



Name	Date Modified
Desktop	Today, 2:30 PM
.DS_Store	Today, 2:30 PM
Downloads	Today, 7:50 AM
Trash	Yesterday, 3:44 PM
Documents	Yesterday, 3:21 PM
.dropbox	Yesterday, 12:16 PM
Library	Jun 14, 2014, 6:01 AM
.Rhistory	Jun 14, 2014, 5:42 AM
Box Documents	Jun 7, 2014, 6:27 AM
Dropbox	Jun 7, 2014, 6:27 AM
.dropbox-master	Jun 7, 2014, 6:27 AM
.viminfo	Jun 5, 2014, 1:53 PM



```
[andrews-mbp-5:~] severin% ls
021208sh2atomlist.prot
2B.perl
2_RorZ4a.html.part
All.genus.species.fromBlastout.txt
All.genus.species.fromBlastout.uniq.txt
Applications
Billing
BioBlitz
Bootcamp
Box Documents
CoAuthors.txt
CoAuthors2.txt
DEGinMt.dec7.8-5.6.normalized.bayseq.Heatmap.pdf
DEGinMt.[illegible]8-5.6.normalized.bayseq.Heatmap.pdf
Desktop
Documents
Downloads
Dropbox
ENC.pse
Examples
FEregulation.ps
FEregulation.ps.log
FEvance.test.frame
```

Running list of commands

- Pull out a sheet of paper and as we learn new commands right them down as we go to refer back to.

Commands in Part I

Navigation	cd, ls, pwd
File creation	touch,nano,mkdir,cp,mv,rm,rmdir
Reading	more,less,head,tail,cat
Compression	zip,gzip,bzip2,tar,compress
Uncompression	unzip,gunzip,bunzip2,uncompress
Permissions	chmod
Help	man

Terminal Window or Prompt

- The > is where commands can be entered
    - The command line in the prompt.
- Very Basic Linux Commands
  - cd = change directory
  - ls = list
  - pwd = path of working directory

Downloading Class Materials

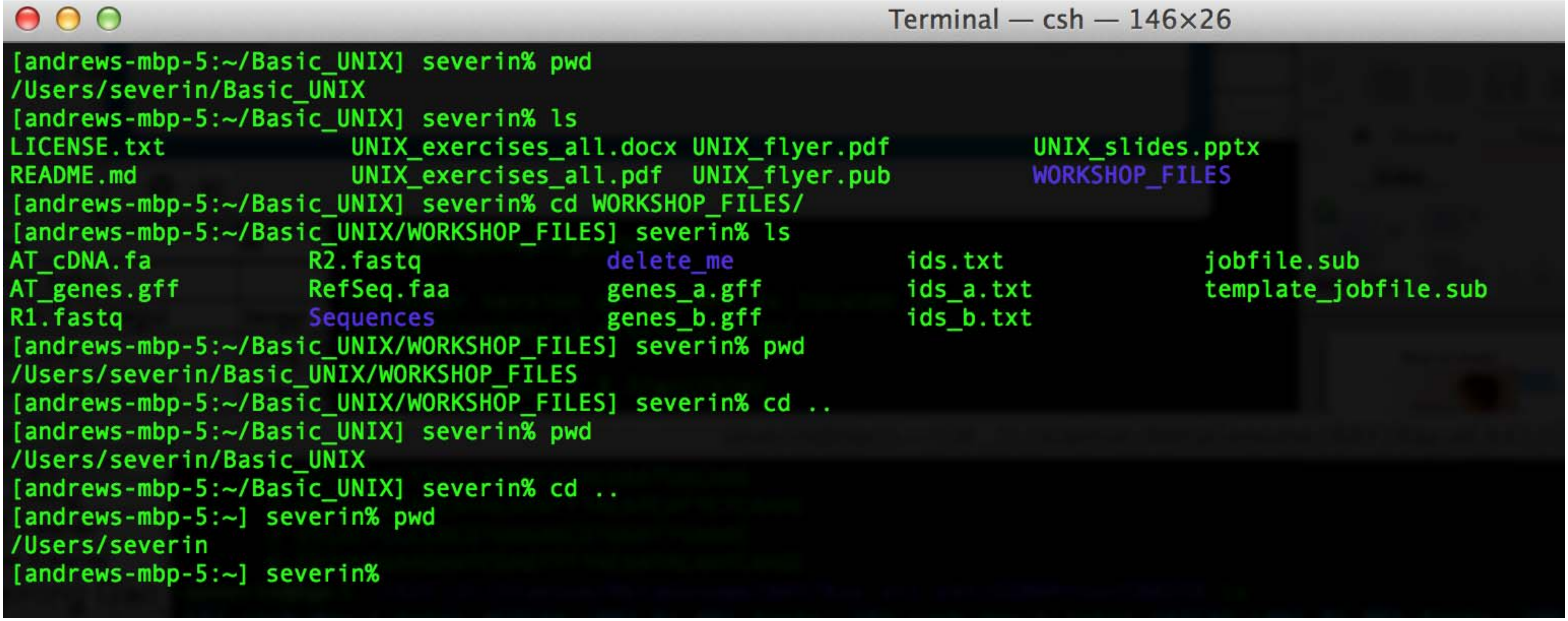
- Hand-outs/files are available for download at https://github.com/ISUgenomics/Basic_UNIX

- cd ↵
- git clone https://github.com/ISUgenomics/Basic_UNIX
- Once your cursor (command prompt) comes back to the original position, type
- ls ↵
- cd Basic_UNIX
- ls ↵

Navigation

- Where am I?
  - pwd
- What is in my folder?
  - ls
- Moving between folders
  - cd WORKSHOP_FILES/
  - ls
  - pwd
- Moving back directories
  - cd ..
  - pwd
  - cd ..
  - pwd

Output of Navigation



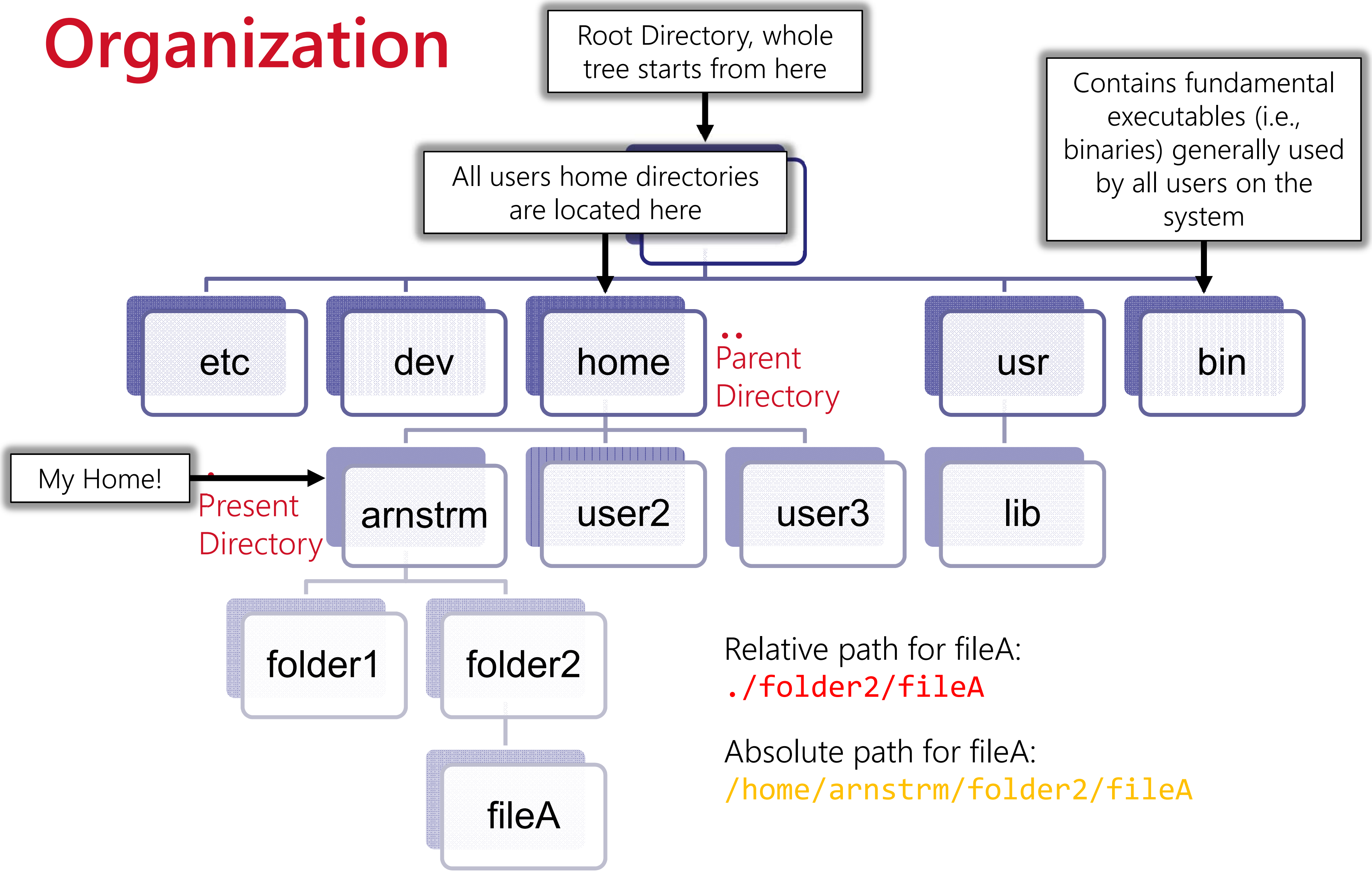
```
[andrews-mbp-5:~/Basic_UNIX] severin% pwd
/Users/severin/Basic_UNIX
[andrews-mbp-5:~/Basic_UNIX] severin% ls
LICENSE.txt             UNIX_exercises_all.docx UNIX_flyer.pdf          UNIX_slides.pptx
README.md               UNIX_exercises_all.pdf  UNIX_flyer.pub          WORKSHOP_FILES
[andrews-mbp-5:~/Basic_UNIX] severin% cd WORKSHOP_FILES/
[andrews-mbp-5:~/Basic_UNIX/WORKSHOP_FILES] severin% ls
AT_cDNA.fa           R2.fastq             delete_me            ids.txt              jobfile.sub
AT_genes.gff         RefSeq.faa           genes_a.gff          ids_a.txt            template_jobfile.sub
R1.fastq             Sequences            genes_b.gff          ids_b.txt
[andrews-mbp-5:~/Basic_UNIX/WORKSHOP_FILES] severin% pwd
/Users/severin/Basic_UNIX/WORKSHOP_FILES
[andrews-mbp-5:~/Basic_UNIX/WORKSHOP_FILES] severin% cd ..
[andrews-mbp-5:~/Basic_UNIX] severin% pwd
/Users/severin/Basic_UNIX
[andrews-mbp-5:~/Basic_UNIX] severin% cd ..
[andrews-mbp-5:~] severin% pwd
/Users/severin
[andrews-mbp-5:~] severin%
```

Go back into the directory

- pwd
- cd Basic_UNIX/
- pwd
- cd WORKSHOP_FILES/
- pwd
- ls

- cd
- pwd
- The last command puts you back to /home/netid

Organization



Root Directory, whole tree starts from here

All users home directories are located here

Contains fundamental executables (i.e., binaries) generally used by all users on the system

My Home!

Relative path for fileA:
`./folder2/fileA`

Absolute path for fileA:
`/home/arnstrm/folder2/fileA`

Repeat

- pwd
- cd Basic_UNIX/
- pwd
- cd WORKSHOP_FILES/
- pwd
- ls

- /Users/netid/Basic_UNIX/WORKSHOP_FILES

Making and Moving

- mkdir playarea
- ls
- cd playarea
- touch myfile
- mkdir mydirectory
- mv myfile mydirectory
- ls
- ls mydirectory

Now, you try!

- touch a b c d e
- touch 1 2 3 4 5

- Your goal is to mkdir two directories: numbers and letters, respectively. Then, mv these new files a-e and 1-5 into them.

Renaming files – The other kind of mv

- touch Zylophone
- ls
- mv Zylophone Xylophone
- ls

Editing files

- Everyone make sure you are in this folder
- /Users/netid/Basic_UNIX/WORKSHOP_FILES/playarea

- touch fav.txt
- nano fav.txt

- Write your 3 favorite things in nano then press control and hit x, hit y, hit enter to exit out of nano.

Reading files

- more fav.txt
- less fav.txt (q to quit)
- cat fav.txt
- head fav.txt
- tail fav.txt

- Let’s find a more interesting example
- Change into this directory
- hint (type pwd and see where you are first)
- /Users/netid/Basic_UNIX/WORKSHOP_FILES

Let’s look at AT_cDNA.fa

• head AT_cDNA.fa		first 10 lines of a file
• less AT_cDNA.fa	(q to quit)	step through/back
• cat AT_cDNA.fa		All lines of the file
• more AT_cDNA.fa		step through
• tail AT_cDNA.fa		last 10 lines of a file

Parameters

- What if you want more than the first 10 lines of a file?
- head AT_cDNA.fa
- First 20 lines?
- head –n 20 AT_cDNA.fa
- Command, parameter, value, file
- Command = head
- Parameter to be change = -n
- Value = 20
- File = AT_cDNA.fa

Command! Parameter! Value! File!

DON'T FORGET THE SPACES!!!!!!!!

Command SPACE parameter SPACE value SPACE file

Manual pages for commands

- man is a command
- man head
- Space bar lets you go farther into the manual
- q quits

- head [-n count | -c bytes] [file ...]

Manual pages for commands

- man ls
- ls
- ls -l
- ls -a
- ls -la
- ls –lha

- Did anyone find a hidden file?

Removing files and directories

Delete the directory named delete_me inside the tutorials directory (to do this you may first want to delete the sample.txt file inside this directory).

- rmdir delete_me
- cd delete_me
- rm sample.txt
- cd ..
- rmdir delete_me

Forcibly removing directories

- rm delete_me_2
- rm –rf delete_me_2

- Use man command to look up what the -r and -f do.

- man rm

Copying files and directories

- ls
- pwd
- cd ..
- pwd /Users/netid/Basic_UNIX

- cp -r WORKSHOP_FILES BACKUP_WORKSHOP
- cp -r WORKSHOP_FILES BACKUP_WORKSHOP2
- cp -r WORKSHOP_FILES BACKUP_WORKSHOP3
- ls

Not enough space? Zip it!

- zip producedZipFileName WhatYouWant2Zip

- zip BACKUP_WORKSHOP3.zip BACKUP_WORKSHOP3

- mv BACKUP_WORKSHOP tutorials
- ls
- cd tutorials
- ls
- zip AT_genes.gff.zip AT_genes.gff

Zip not enough? Try gzip tar or compress

- **tar** -czvf AT_genes.gff.tar.gz AT_genes.gff
- cp AT_genes.gff AT_genes2.gff
- **compress** AT_genes2.gff
- cp AT_genes.gff AT_genes-gzip.gff
- **gzip** AT_genes-gzip.gff
- cp AT_genes.gff AT_genes-bzip2.gff
- **bzip2** AT_genes-bzip2.gff
- Let’s see how we did
- ls -l AT_genes*

UnZipping

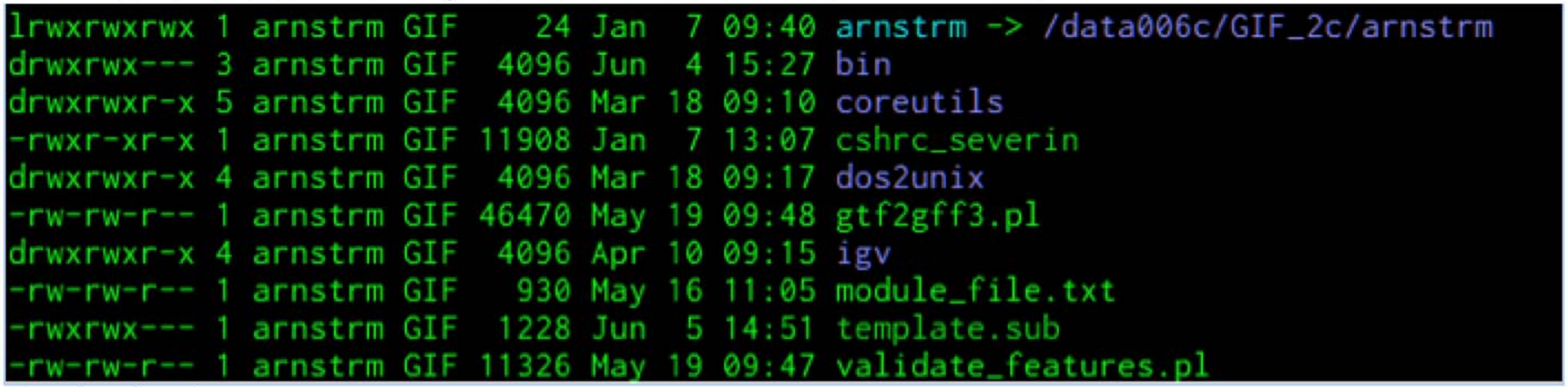
- tar -xvf AT_genes.gff.tar.gz
- unzip AT_genes.gff.tar.gz
- uncompress AT_genes2.gff.Z
- bunzip2 AT_genes-bzip2.gff.bz2
- gunzip AT_genes-gzip.gff.gz

File permissions

PERMISSIONS		RELATIONS	
read	r	owner	u
write	w	group	g
execute	x	others	o
		all users	a

To look at the permissions for any file, you can list the files with l option (`ls -l`).

```
Permissions User   Group Size  Date modified      Name
```



```
lrwxrwxrwx 1 arnstrm GIF    24 Jan  7 09:40 arnstrm -> /data006c/GIF_2c/arnstrm
drwxrwx--- 3 arnstrm GIF  4096 Jun  4 15:27 bin
drwxrwxr-x 5 arnstrm GIF  4096 Mar 18 09:10 coreutils
-rwxr-xr-x 1 arnstrm GIF 11908 Jan  7 13:07 cshrc_severin
drwxrwxr-x 4 arnstrm GIF  4096 Mar 18 09:17 dos2unix
-rw-rw-r-- 1 arnstrm GIF 46470 May 19 09:48 gtf2gff3.pl
drwxrwxr-x 4 arnstrm GIF  4096 Apr 10 09:15 igv
-rw-rw-r-- 1 arnstrm GIF   930 May 16 11:05 module_file.txt
-rwxrwx--- 1 arnstrm GIF  1228 Jun  5 14:51 template.sub
-rw-rw-r-- 1 arnstrm GIF 11326 May 19 09:47 validate_features.pl
```

u g o

(d=directory, l=link, r=read, w=write, x=execute, -=blank, u=user, g=group, o=others)

File permissions example

- chmod 000 YouCannotEnter
- ls -l YouCannotEnter
- cd YouCannotEnter
- Permission denied!!!!!   -- file permission error
- chmod a+rx YouCannotEnter/
- Now you and everyone can enter this directory
- chmod o-rx YouCannotEnter/
- Now you and your group can enter this directory
- chmod g-rx
- Now only you the user can enter this directory

Summary of Part I

Navigation	cd, ls, pwd
File creation	touch,nano,mkdir,cp,mv,rm,rmdir
Reading	more,less,head,tail,cat
Compression	zip,gzip,bzip2,tar,compress
Uncompression	unzip,gunzip,bunzip2,uncompress
Permissions	chmod
Help	man

- Check your sheet, do you have all of these?
- You can now use linux as you do mac or windows OS

Part 2:

Moving data	cat,>,>>,<,|
Regular expressions	/^.*[0-9]+[a-z]*.*$/
Find and replace	grep,sed,tr
Manipulating rows/columns	cut,awk
Comparing files	wc,sort,uniq,diff,comm
Manipulating files	split,join,paste

Moving data

- Everyone should be here
  - /Users/netid/Basic_UNIX/tutorials
  - Use pwd and check
- cat AT_cDNA.fa

Piping | located above enter below delete use shift

- cat AT_cDNA.fa | head

seqlen.awk - Generate sequence ID & sequence length from FASTA

- cat AT_cDNA.fa | head | ./seqlen.awk
  - AT1G51370.2 720

Moving Data

- cat AT_cDNA.fa | head > new.fasta
- more new.fasta
- cat AT_cDNA.fa | tail -n 30 >> new.fasta
- more new.fasta

Create a file named AT_cDNA.len that contains the lengths of each sequence in AT_cDNA.fa

Moving Data

- cat AT_cDNA.fa | head > new.fasta
- more new.fasta
- cat AT_cDNA.fa | tail -n 30 >> new.fasta
- more new.fasta

Create a file named AT_cDNA.len that contains the lengths of each sequence in AT_cDNA.fa

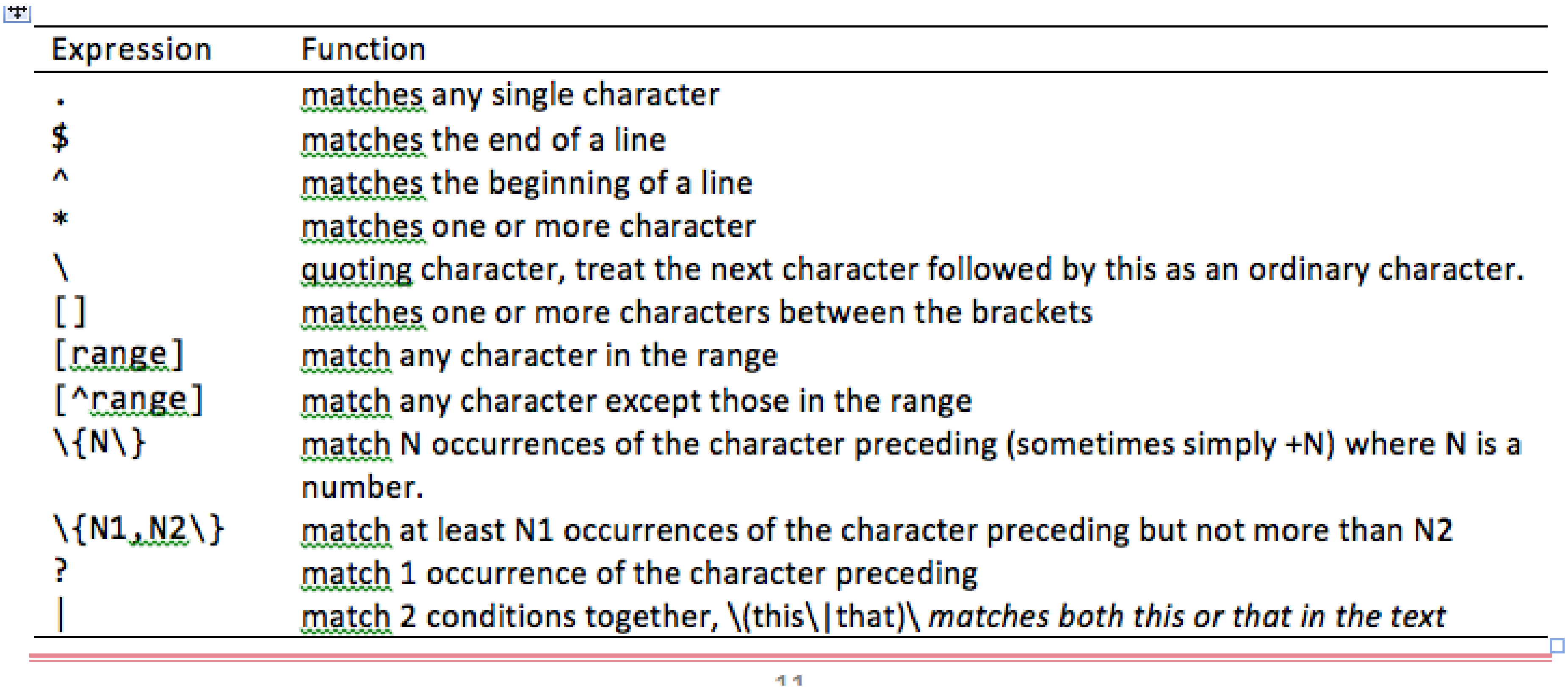
- cat AT_cDNA.fa | ./seqlen.awk > AT_cDNA.len

Find this pattern please!

- grep = find this pattern

- Example of a simple search
- ls | grep ids
    - ids.txt
    - ids_a.txt
    - ids_b.txt
- ls | awk '/ids/'
- Grab the first 10 headers in AT_cDNA.fa
- more AT_cDNA.fa | grep ">" | head > AT_cDNA.head.fa

Regular Expressions



Expression	Function
.	matches any single character
$	matches the end of a line
^	matches the beginning of a line
*	matches one or more character
\	quoting character, treat the next character followed by this as an ordinary character.
[]	matches one or more characters between the brackets
[range]	match any character in the range
[^range]	match any character except those in the range
\{N\}	match N occurrences of the character preceding (sometimes simply +N) where N is a number.
\{N1,N2\}	match at least N1 occurrences of the character preceding but not more than N2
?	match 1 occurrence of the character preceding
|	match 2 conditions together, \(this\\|that)\ *matches both this or that in the text*

11

Regular Expressions

you have already seen * = match any character

- `ls AT*`
- `ls genes*`

Let’s grab all the headers that have transposable in their names and start with ATG40 in the gene name.

- `more AT_cDNA.fa | grep ">AT1G" | grep transposable | grep AT1G40`

More succinctly

- `more AT_cDNA.fa | grep "^>AT1G40.*transposable.*"`

Verify they give the same result

Regular Expressions

- Now lets find all fasta ids that have chromosome positions between 15 million and 15 million 300 thousand

-
  ```
  more AT_cDNA.fa | grep "chr1\:15[0-3][0-9]*" | grep transposable
  ```

- There is a lot that can be done with regular expression and I encourage you to learn more on your own via the exercises and other online resources

Find and Replace

Replace Symbol with Andrew in AT_cDNA.head.fa

- `sed 's/Symbols/Andrew/g' AT_cDNA.head.fa`

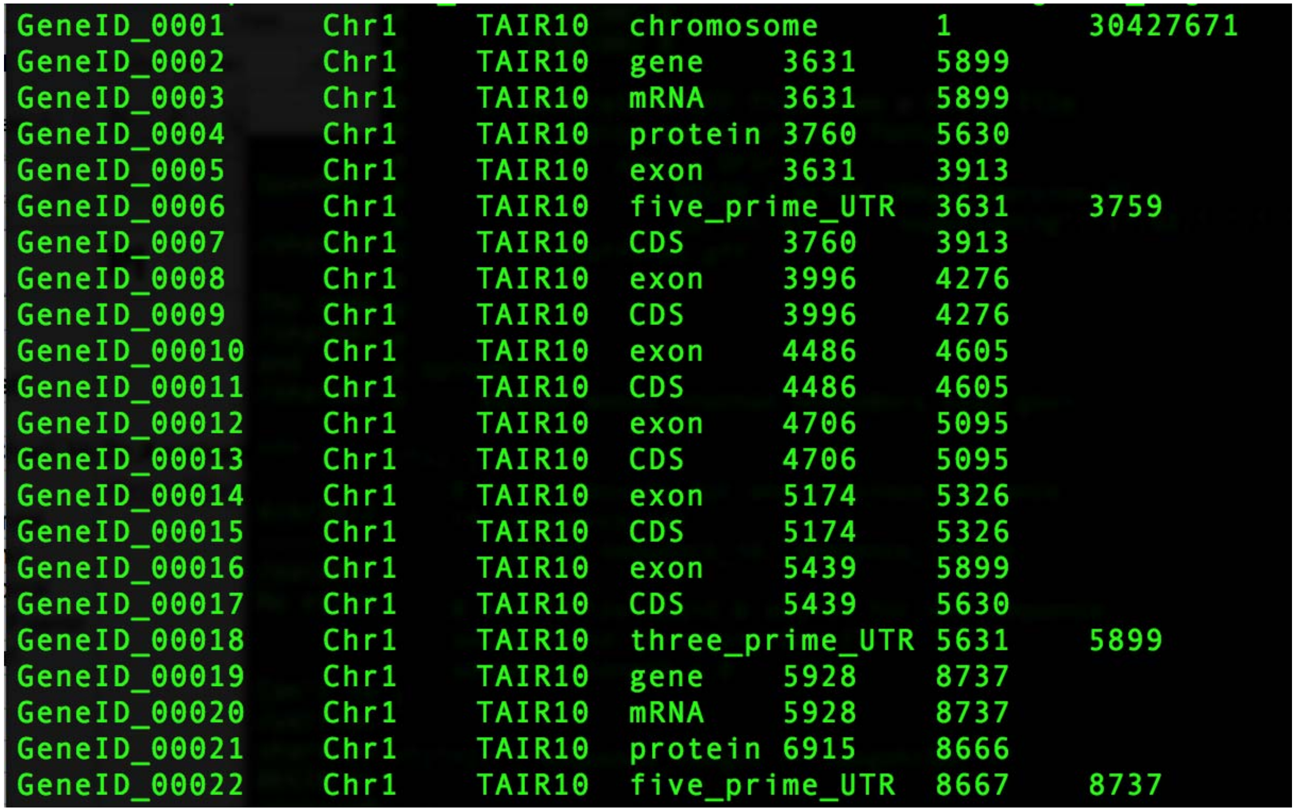
- `perl -pe 's/Symbols/Andrew/g' AT_cDNA.head.fa`

Make all caps

- `tr 'a-z' 'A-Z' < AT_cDNA.head.fa`
- `cat AT_cDNA.head.fa | tr 'a-z' 'A-Z’`

Manipulating rows/columns

- More genes_a.gff

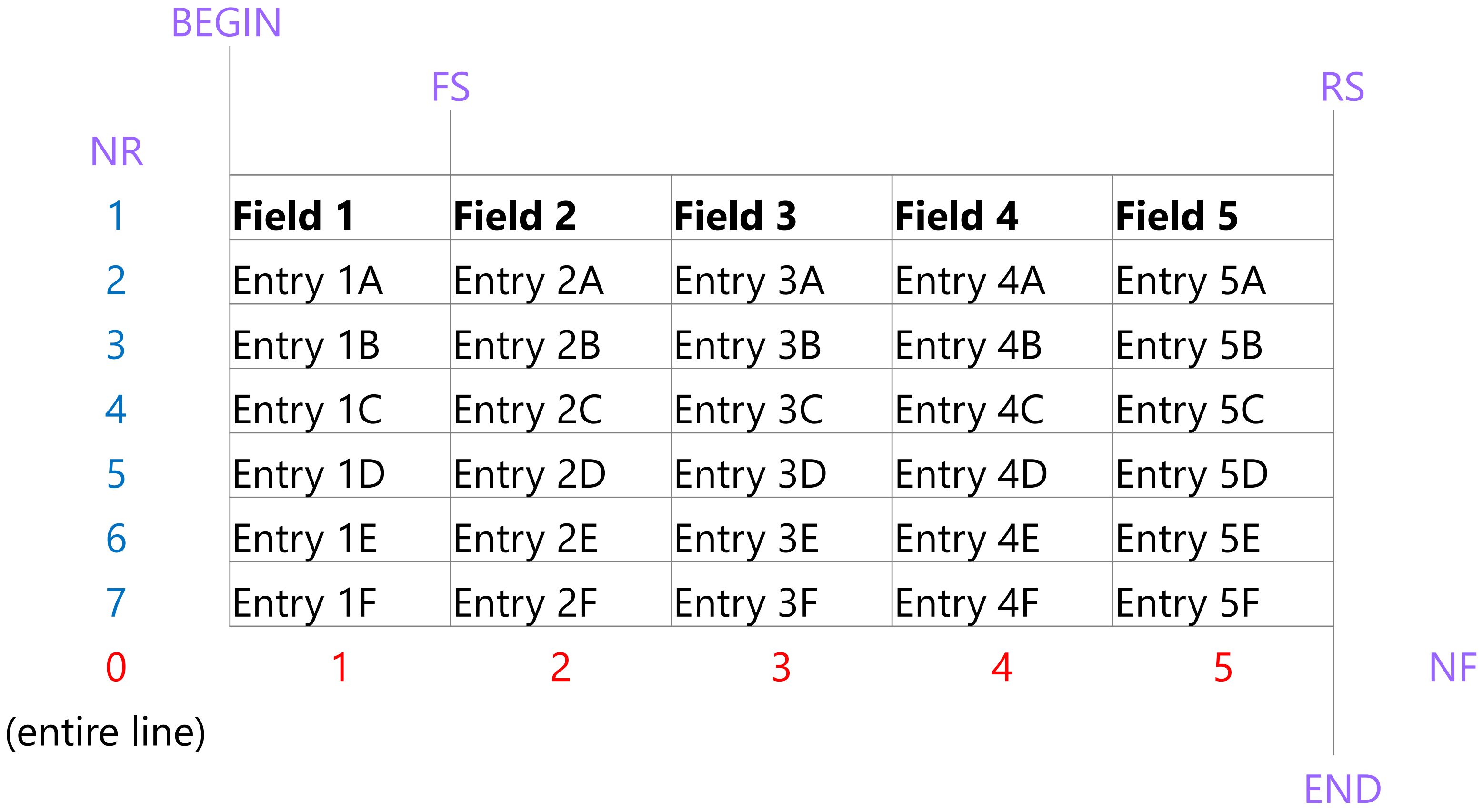


```
GeneID_0001     Chr1    TAIR10  chromosome      1       30427671
GeneID_0002     Chr1    TAIR10  gene    3631    5899
GeneID_0003     Chr1    TAIR10  mRNA    3631    5899
GeneID_0004     Chr1    TAIR10  protein 3760    5630
GeneID_0005     Chr1    TAIR10  exon    3631    3913
GeneID_0006     Chr1    TAIR10  five_prime_UTR  3631    3759
GeneID_0007     Chr1    TAIR10  CDS     3760    3913
GeneID_0008     Chr1    TAIR10  exon    3996    4276
GeneID_0009     Chr1    TAIR10  CDS     3996    4276
GeneID_00010    Chr1    TAIR10  exon    4486    4605
GeneID_00011    Chr1    TAIR10  CDS     4486    4605
GeneID_00012    Chr1    TAIR10  exon    4706    5095
GeneID_00013    Chr1    TAIR10  CDS     4706    5095
GeneID_00014    Chr1    TAIR10  exon    5174    5326
GeneID_00015    Chr1    TAIR10  CDS     5174    5326
GeneID_00016    Chr1    TAIR10  exon    5439    5899
GeneID_00017    Chr1    TAIR10  CDS     5439    5630
GeneID_00018    Chr1    TAIR10  three_prime_UTR 5631    5899
GeneID_00019    Chr1    TAIR10  gene    5928    8737
GeneID_00020    Chr1    TAIR10  mRNA    5928    8737
GeneID_00021    Chr1    TAIR10  protein 6915    8666
GeneID_00022    Chr1    TAIR10  five_prime_UTR  8667    8737
```

Manipulating rows/columns

- awk '{print NF}' genes_a.gff | head -n 1
- awk '{print NR}' genes_a.gff | tail -n 1
- more genes_a.gff
- press up to get the last command and modify with arrows
- more genes_a.gff | **awk '{print $1,$2,$5,$6}'** | more
- more genes_a.gff | awk '{print $1,$2,$5,$6,**$6-$5**}' | more
- more genes_a.gff | awk '{print $1,$2,$5,$6,$6-$5}' | **sort** | head
- more genes_a.gff | awk '{print $1,$2,$5,$6,$6-$5}' | sort **–k 5n** | head
- more genes_a.gff | awk '{print $1,$2,$5,$6,$6-$5}' | sort –k 5**rn** | head
- more genes_a.gff | awk '{print $1,$2,$4,$5,$6,$6-$5}' | awk ’$6>2000'
- more genes_a.gff | awk **’OFS=“\t"** {print $1,$2,$4,$5,$6,$6-$5}' | sort -k 6rn | head

AWK



NR						
1	**Field 1**	**Field 2**	**Field 3**	**Field 4**	**Field 5**	
2	Entry 1A	Entry 2A	Entry 3A	Entry 4A	Entry 5A	
3	Entry 1B	Entry 2B	Entry 3B	Entry 4B	Entry 5B	
4	Entry 1C	Entry 2C	Entry 3C	Entry 4C	Entry 5C	
5	Entry 1D	Entry 2D	Entry 3D	Entry 4D	Entry 5D	
6	Entry 1E	Entry 2E	Entry 3E	Entry 4E	Entry 5E	
7	Entry 1F	Entry 2F	Entry 3F	Entry 4F	Entry 5F	
0 (entire line)	1	2	3	4	5	NF

How many different Items are in column 4?

```
more genes_a.gff
more genes_a.gff | awk '{print $4}'
more genes_a.gff | awk '{print $4}' | sort
more genes_a.gff | awk '{print $4}' | sort | uniq
more genes_a.gff | awk '{print $4}' | sort | uniq –c
more genes_a.gff | awk '{print $4}' | sort | uniq -c | sort –rn
```

You can also use cut to grab a collumn.

```
cut -f 4 genes_a.gff
cut -f 4- genes_a.gff
```

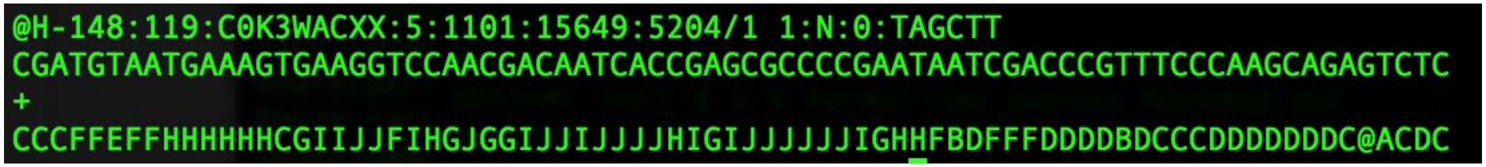
Comparing files

- more AT_cDNA.fa
- more AT_cDNA.fa | grep ">" | more
- more AT_cDNA.fa | grep ">" | awk '{print $1, $NF}' | more
- more AT_cDNA.fa | grep ">" | awk '{print $1,$NF}' | sed 's/LENGTH=//g’
- more AT_cDNA.fa | grep ">" | awk '{print $1,$NF}' | sed 's/LENGTH=//g’
- more AT_cDNA.fa | grep ">" | awk '{print $1,$NF}' | sed 's/LENGTH=//g’ | perl -pe ‘s/>//g’
- more AT_cDNA.fa | grep ">" | awk '{print $1,$NF}' | sed 's/LENGTH=//g' | perl -pe 's/>//g' > AT_cDNA.len2
- cat AT_cDNA.fa | ./seqlen.awk > AT_cDNA.len

Comparing files

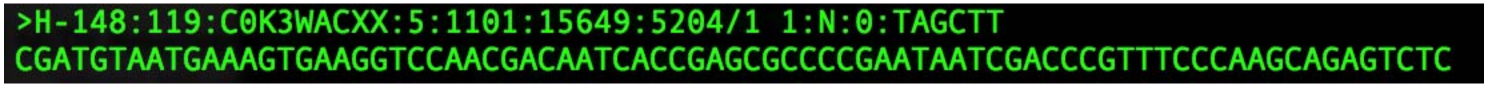
- diff AT_cDNA.len AT_cDNA.len2
- comm AT_cDNA.len AT_cDNA.len2

Manipulating files



```
@H-148:119:C0K3WACXX:5:1101:15649:5204/1 1:N:0:TAGCTT
CGATGTAATGAAAGTGAAGGTCCAACGACAATCACCGAGCGCCCCGAATAATCGACCCGTTTCCCAAGCAGAGTCTC
+
CCCFFEFFHHHHHCGIIJJFIHGJGGIJJIJJJJHIGIJJJJJJIGHHFBDFFFDDDDBDCCCDDDDDDDC@ACDC
```

- more R1.fastq | paste - - - - | more
- more R1.fastq | paste - - - - | awk '{print $1,$2; print $3}’
- more R1.fastq | paste - - - - | awk '{print $1,$2; print $3}’ | sed 's/@/>/g’ > R1.fasta



```
>H-148:119:C0K3WACXX:5:1101:15649:5204/1 1:N:0:TAGCTT
CGATGTAATGAAAGTGAAGGTCCAACGACAATCACCGAGCGCCCCGAATAATCGACCCGTTTCCCAAGCAGAGTCTC
```

Manipulating files

- wc R1.fastq
- split -l 4000 R1.fastq R1_
- ls R1_* | wc

Summary Part 2

Moving data	cat,>,>>,<,|
Regular expressions	/^.*[0-9]+[a-z]*.*$/
Find and replace	grep,sed,tr
Manipulating rows/columns	cut,awk
Comparing files	wc,sort,uniq,diff,comm
Manipulating files	split,join,paste

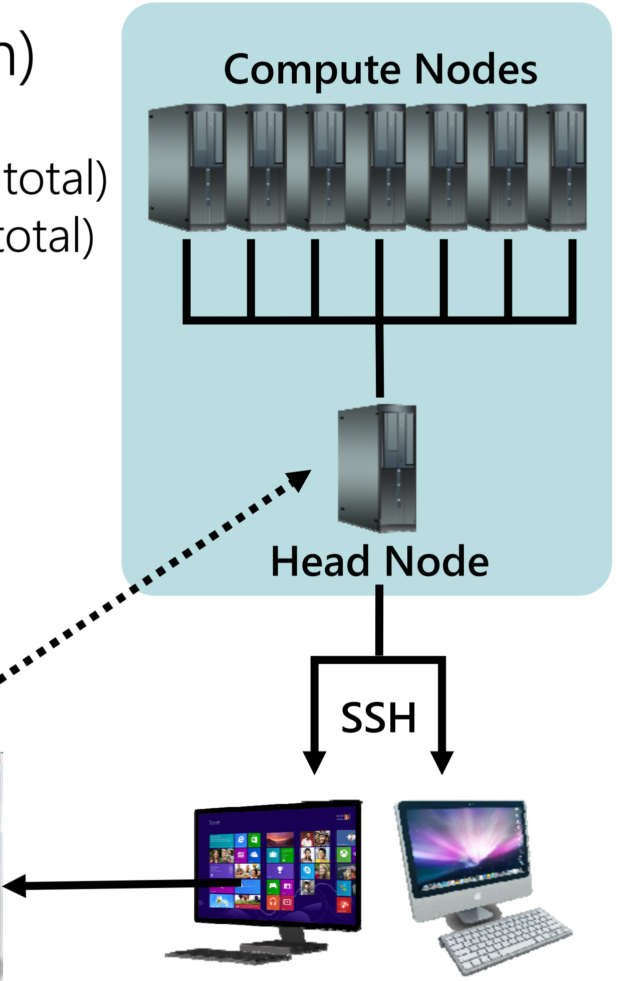
Part 3

- Logging into hpc-class (remote machine)
  - ssh
- Transferring and Downloading data
  - Git,wget,rsync,scp
- Preinstalled software
- Understanding Queues
  - Qstat, qtop
- Submitting Jobs

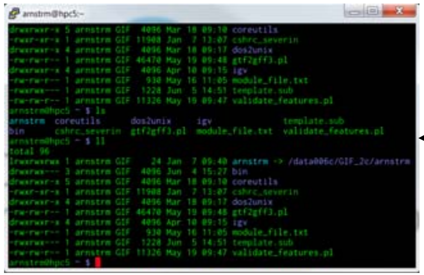
High Performance Computing (HPC)

HPC-class (education)

- 48 compute nodes
- 16 processors/node (768 total)
- 64Gb RAM/node (3.0Tb total)



Terminal



```
drwxrwxr-x 5 arnstrm GIF  4096 Mar 18 09:10 coreutils
-rwxr-xr-x 1 arnstrm GIF 11908 Jan  7 13:07 cshrc_severin
drwxrwxr-x 4 arnstrm GIF  4096 Mar 18 09:17 dos2unix
-rw-rw-r-- 1 arnstrm GIF 46470 May 19 09:48 gtf2gff3.pl
drwxrwxr-x 4 arnstrm GIF  4096 Apr 10 09:15 igv
-rw-rw-r-- 1 arnstrm GIF   930 May 16 11:05 module_file.txt
-rwxrwx--- 1 arnstrm GIF  1228 Jun  5 14:51 template.sub
-rw-rw-r-- 1 arnstrm GIF 11326 May 19 09:47 validate_features.pl
arnstrm@hpc5 ~ $ ls
arnstrm  coreutils      dos2unix     igv              template.sub
bin      cshrc_severin  gtf2gff3.pl  module_file.txt  validate_features.pl
arnstrm@hpc5 ~ $ ll
total 96
lrwxrwxrwx 1 arnstrm GIF    24 Jan  7 09:40 arnstrm -> /data006c/GIF_2c/arnstrm
drwxrwx--- 3 arnstrm GIF  4096 Jun  4 15:27 bin
drwxrwxr-x 5 arnstrm GIF  4096 Mar 18 09:10 coreutils
-rwxr-xr-x 1 arnstrm GIF 11908 Jan  7 13:07 cshrc_severin
drwxrwxr-x 4 arnstrm GIF  4096 Mar 18 09:17 dos2unix
-rw-rw-r-- 1 arnstrm GIF 46470 May 19 09:48 gtf2gff3.pl
drwxrwxr-x 4 arnstrm GIF  4096 Apr 10 09:15 igv
-rw-rw-r-- 1 arnstrm GIF   930 May 16 11:05 module_file.txt
-rwxrwx--- 1 arnstrm GIF  1228 Jun  5 14:51 template.sub
-rw-rw-r-- 1 arnstrm GIF 11326 May 19 09:47 validate_features.pl
arnstrm@hpc5 ~ $
```

Lightning3 (research)

- 17 compute nodes
- 32 processors/node (384 total)
- 256Gb RAM/node (3.2Tb total)

Condo (up coming)

- 316 compute nodes
- 8 processors/node (2528 total)
- 8Gb RAM/node (3.5Tb total)

Logging in

- Microsoft Windows:
- **PuTTY** is an extremely small download of a free, full-featured SSH client.
- **SSH Secure Shell Client**, also a full featured client that is commercial. It is available as part of the Iowa State University site-licensed software.

Logging in

- Macintosh
- ssh -X username@hpc-class.its.iastate.edu

Note: You will not see your password as you type.

Logging in

- ls   nothing in your folder
- ls -la
- pwd   /home/netid
- who   Who else is on this machine
- cp -r /home/severin/Basic_UNIX .
- cd Basic_UNIX/WORKSHOP_FILES

- The remote machine will have a prompt that looks like
- [netid@**hpc-class** WORKSHOP_FILES]$

Transferring and downloading files

- Downloading from a website

- wget http://goo.gl/CDXx15 =soybean annotation
- How do we look at this file?

Transferring and downloading files

- Transferring files from our local machine to the remote machine

- Open up a new terminal

- touch cats.txt
- scp cats.txt netid@hpc-class.its.iastate.edu:/home/netid

- rsync -avz -e ssh cats.txt netid@hpc-class.its.iastate.edu:/home/netid

Pre-installed software

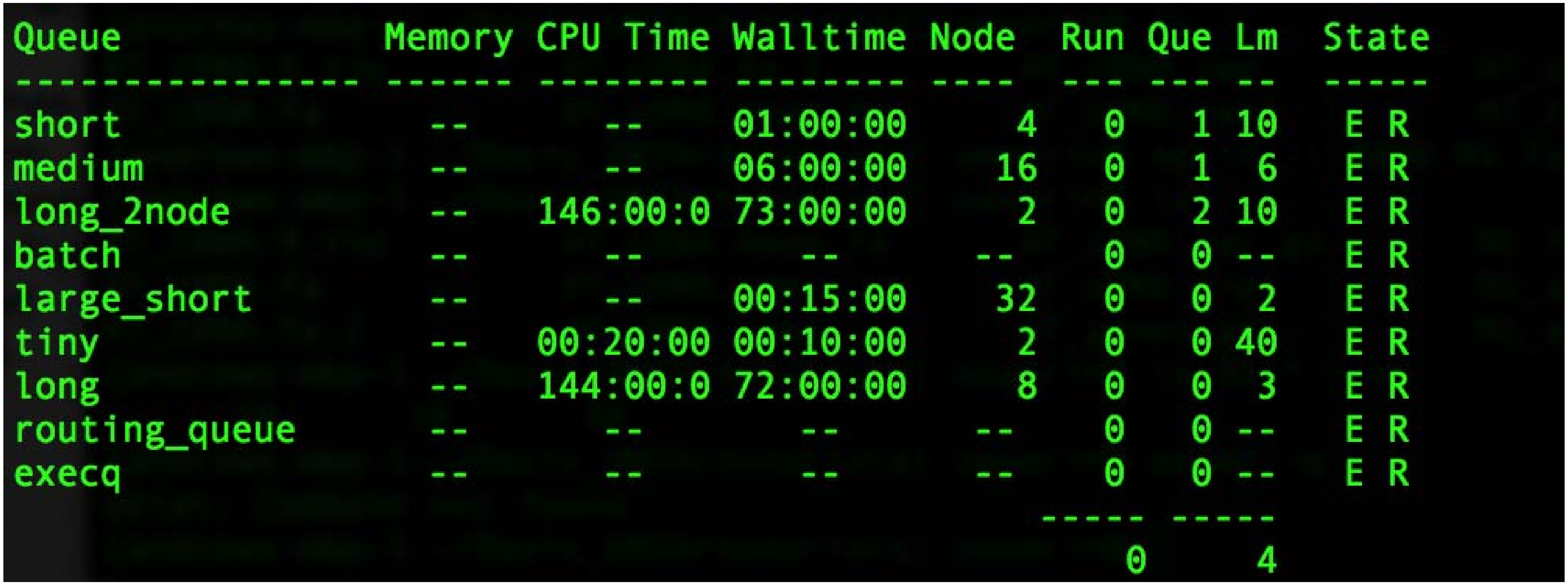
- module use /shared/bioinformatics/modules

- module avail
- module what-is

- fastqc --version
- module load fastqc
- fastqc --version

Understanding Torque and Queueing

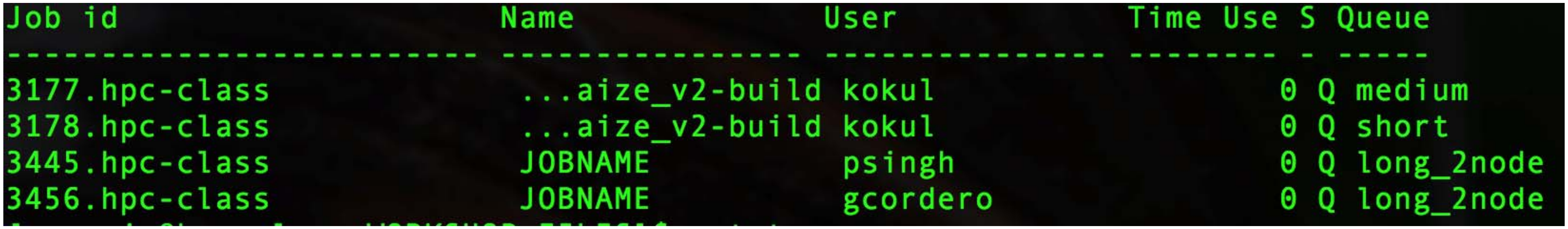
- qstat -q



```
Queue            Memory CPU Time Walltime Node  Run Que Lm  State
---------------- ------ -------- -------- ----  --- --- --  -----
short              --      --    01:00:00     4   0   1 10   E R
medium             --      --    06:00:00    16   0   1  6   E R
long_2node         --   146:00:0 73:00:00     2   0   2 10   E R
batch              --      --       --      --    0   0 --   E R
large_short        --      --    00:15:00    32   0   0  2   E R
tiny               --   00:20:00 00:10:00     2   0   0 40   E R
long               --   144:00:0 72:00:00     8   0   0  3   E R
routing_queue      --      --       --      --    0   0 --   E R
execq              --      --       --      --    0   0 --   E R
                                               ----- -----
                                                   0     4
```

Understanding Torque and Queueing

- qstat -a



```
Job id                    Name             User            Time Use S Queue
------------------------- ---------------- --------------- -------- - -----
3177.hpc-class             ...aize_v2-build kokul                 0 Q medium
3178.hpc-class             ...aize_v2-build kokul                 0 Q short
3445.hpc-class             JOBNAME          psingh                0 Q long_2node
3456.hpc-class             JOBNAME          gcordero              0 Q long_2node
```

Submitting a job

- #!/bin/bash
- #PBS -l vmem=16Gb,pmem=4Gb,mem=16Gb
- #PBS -l nodes=1:ppn=4:compute
- #PBS -l walltime=48:00:00
- #PBS -N *FASTQC* ← You can change this
- #PBS -o ${PBS_JOBNAME}.o${PBS_JOBID} -e ${PBS_JOBNAME}.e${PBS_JOBID}
- cd $PBS_O_WORKDIR

- fastqc R1.fastq ← Your command here

Submitting a job

- /home/netid/Basic_UNIX/WORKSHOP_FILES

- more jobfile.sub
- qsub jobfile.sub
- qstat –a

- R1_fastqc.html
- R2_fastqc.html

- firefox R1_fastqc.html   ← to view the results

Summary: Part 3

- Logging into hpc-class (remote machine)
  - ssh
- Transferring and Downloading data
  - git,wget,rsync,scp
- Understanding Queues
  - Qstat, qtop
- Submitting Jobs

Where to go from here?

- Review the material from the workshop in more detail
  - Exercises (Basic_UNIX/UNIX_exercises_all.docx)
- Unix and Perl primer for Biologists
  - http://korflab.ucdavis.edu/Unix_and_Perl/
- Advanced Unix workshop coming soon!

Post Workshop Survey

- Please tell us what you think about this workshop by completing this short survey (10 questions)
- http://goo.gl/XJq7Bk