

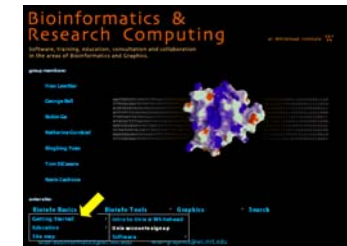
Introduction to Unix: most important/useful commands & examples

Bingbing Yuan
Jan. 19, 2010

1

Where can UNIX be used?

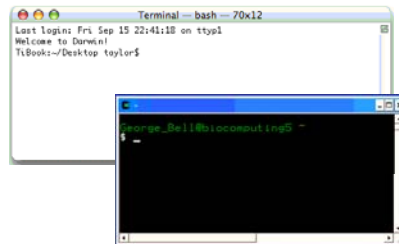
- Real Unix computers
 - “tak”, the Whitehead Scientific Linux server
 - Apply for an account on the BaRC page
- Mac computers
 - Come with Unix
- Windows computers
 - Need Cygwin:
Free from
<http://www.cygwin.com/>



2

Getting to the terminal

- Macs:
 - Go to Applications => Utilities => Terminal or X11
- Windows:
 - Click on Cygwin
- To log in to tak:
 - `ssh -l userName tak.wi.mit.edu`



3

Where are you?

List all files/directories

`ls` [only show names]

`ls -l` [long listing: show other information too]

```
byuan@tak:~$
byuan@tak:~/format.pl$ ls
brub.txt  readme.txt  results/  s_3_sequence.txt.tar.gz  s_4_sequence.txt.tar.gz  s_7_sequence.txt.tar.gz
byuan@tak:~$ ls -l
total 152
-rw-r--r-- 1 byuan user 3611 2009-06-09 15:33 bootia/youngformat.pl*
-rw-r--r-- 1 byuan user 2882 2009-06-09 13:04 bootia/youngformat_v1.pl*
-rw-r--r-- 1 byuan user 9749 2009-06-09 11:59 brub.txt
-rw-r--r-- 1 byuan user 935 2009-06-09 16:36 readme.txt
-rw-r--r-- 1 byuan user 1024 2009-06-11 10:05 results/
lrwxrwxr-x 1 byuan user 91 2010-01-13 15:55 s_3_sequence.txt.tar.gz -> /lab/solexa_public/Young/949527_WICRP-0520X_3039AA00/qualityscore/s_3_sequence.txt.tar.gz
lrwxrwxr-x 1 byuan user 91 2010-01-13 15:55 s_4_sequence.txt.tar.gz -> /lab/solexa_public/Young/949527_WICRP-0520X_3039AA00/qualityscore/s_4_sequence.txt.tar.gz
lrwxrwxr-x 1 byuan user 91 2010-01-13 15:55 s_7_sequence.txt.tar.gz -> /lab/solexa_public/Young/949527_WICRP-0520X_3039AA00/qualityscore/s_7_sequence.txt.tar.gz
```

`ln -s /lab/solexa_public/.../QualityScore/s_7_sequence.txt.tar.gz .`

4

Changing permissions

- Who can **read**, **write**, or **execute** files?
 - **User (u)**, **group (g)**, or **others (o)**?
 - 9 choices (rwx or each type of person; default = 644)


0 = no permission	4 = read only
1 = execute only	5 = r + x
2 = write only	6 = r + w
3 = x + w	7 = r + w + x
- Default: `-rw-r--r--`
- `-rw-rw-r-- chmod 664 myFile (chmod g+w myFile)`
 - `-rw----- chmod 600 myFile (chmod go-r myFile)`
 - `-rwxr-xr-x chmod 755 myProgram (chmod a+
myProgram)`

5

Where do you want to go?

- Print the **w**orking **d**irectory: `pwd`
- Change **d**irectories to where you want to go: `cd dir`
- Going up the hierarchy: `cd ..`
- Go back home: `cd` or `cd ~`
- Root: first /
- Gobo: /nfs/ or /lab/

```
byuan@tak ~$ pwd
/home/byuan
byuan@tak ~$ cd /nfs/BaRC
byuan@tak /nfs/BaRC$ cd ../
byuan@tak /nfs$ cd ../
byuan@tak /$ cd
byuan@tak ~$ pwd
/home/byuan
```



6

Combining commands

- In a pipeline of commands, the output of one command is used as input for the next
- Link commands with the “pipe” symbol: |

ex1: `ls *.fa | wc -l`

ex2: `grep ">" *.fa | sort`

7

Save files

- Defaults: stdin = keyboard; stdout = screen
- output examples
 - `ls > file_name` (make new file)
 - `ls >> file_name` (append to file)
 - `ls foo >| file_name` (overwrite)

8

Read files

`more file_name`

- **Display first n lines of file: n=50**
`head -50 file_name`
- **Display last 100 lines of file: n=100**
`tail -100 file_name`
- **Display all except header line**
`tail --line=+2 file_name`
- **Display lines between 600 and 1000 lines:**
`head -1000 file_name | tail -400`
`awk 'NR==600, NR==1000' file_name`

9

Print lines matching a pattern

grep

`byuan@tak$ more FILE`

```
U0 chr19.fa 4126539 R
U0 chr6.fa 81889764 R
U0 Chr6.fa 77172493 R
byuan@tak$ grep -v 'chr19' FILE
U0 chr6.fa 81889764 R
U0 Chr6.fa 77172493 R
```

`byuan@tak$ grep 'chr6' FILE`

```
U0 chr6.fa 81889764 R
byuan@tak$ grep -i 'chr6' FILE
U0 chr6.fa 81889764 R
U0 Chr6.fa 77172493 R
byuan@tak$ grep -n -i 'chr6' FILE
2:U0 chr6.fa 81889764 R
3:U0 Chr6.fa 77172493 R
```

<code>-v</code>	select non-matching lines
<code>-i</code>	ignore case
<code>-n</code>	line number

10

Print lines matching a pattern

grep

- `grep ">" seqFile.fa`

```
>AM293347.1 Schmidtea
mediterranea mRNA for msh2
protein
```

- `>` : is required to be at the beginning of the header line in fasta sequence

- `grep -A 3 ">" seqFile.fa`

```
>AM293347.1 Schmidtea mediterranea
mRNA for msh2 protein
ACAATCAATAAAATAAAATCATTGATCTCATA
GCCTCATTGGCTAATTGAATTGACTGCTTGA
AGCCTATCAGAAATTTTACAGCGGAA
```

- `-A NUM`
 - Print NUM of lines After the matching line
- `-B NUM`
 - Print NUM of lines Before the matching line
- `-C NUM`
 - Print NUM of lines Before and After the matching line

11

cut sections from each line of files

cut

- `more FILE`

```
Read2 GAAGTGGATTAGAGTGTGAATTGGCC U0 1 0 0 chrX.fa 78426100 R
Read8 ATACCTGGATCTTCCAGCTGGGGAC U0 1 0 0 chr1.fa 77055965 F
```

- `cut -f1,2,7-9 FILE`

```
Read2 GAAGTGGATTAGAGTGTGAATTGGCC chrX.fa 78426100 R
Read8 ATACCTGGATCTTCCAGCTGGGGAC chr1.fa 77055965 F
```

<code>-f</code>	output only these fields
<code>-d</code>	field delimiter Default: TAB

paste

merge lines of files

`paste file_1 file_2 file_3 >all_files`

12

cut and paste

```
byuan@tak$ head -3 exp_2
```

```
Genbank Acc UniGene ID exp Gene Symbol & Name
BC044791 Mm.208618 109181 Trip11; thyroid hormone receptor interactor 11
AK029748 Mm.183137 16678 Krt2-1; keratin complex 2, basic, gene 1
```

```
byuan@tak$ paste exp_2 exp_3 exp_4 |head -1
```

```
Genbank Acc UniGene ID exp Gene Symbol & Name Genbank Acc
UniGene ID exp Gene Symbol & Name Genbank Acc UniGene ID exp
Gene Symbol & Name
```

```
byuan@tak$ paste exp_2 exp_3 exp_4 |cut -f1,2,3,7,11,12 |head -3
```

```
Genbank Acc UniGene ID exp exp exp Gene Symbol & Name
BC044791 Mm.208618 109181 109184 109187 Trip11; thyroid hormone
receptor interactor 11
AK029748 Mm.183137 16678 16679.2 16680.4 Krt2-1; keratin complex 2,
basic, gene 1
```

13

Sort lines of text files: **sort**

```
byuan@tak$ head -1 mapped.txt
```

```
SRR015146.1_WICMT-SOLEXA_8_3_1_908_882_length=26 - chrX 79418719
GGCCAATTCACACTCTAATCCACTTC IDIIIIIIIIIIIIIIIIIIII 0
```

```
byuan@tak$ cut -f2-5 mapped.txt |head -3
```

```
- chrX 79418719 GGCCAATTCACACTCTAATCCACTTC
+ chr1 77169391 ATACCTGGATCTCCAGCTGGGGAC
- chr13 38726605 TGGGGCTCCAACACTAGTCCCAATTC
```

```
byuan@tak$ cut -f2-5 mapped.txt |sort -k 2,2d -k 3,3n |head -3
```

```
+ chr1 3007991 TGATCTAACTTTGGTACCTGGTATCT
+ chr1 3009967 TTTCCATTTCCATTTCTTTGATT
+ chr1 3009967 TTTCCATTTCCATTTCTTTGATT
```

```
byuan@tak$ cut -f2-5 mapped.txt |grep "chr15" |sort -k 2,2d -k 3,3n |head -3
```

```
+ chr15 3003325 GCCCAGAGTCCCACAGCCTGCTGCCT
+ chr15 3005096 GCAGTGGAAATTTCTTTTGTGTTAC
+ chr15 3009156 GAATTGATGCAGGAAATAGATTGTTCC
```

-k Field	-t field-separator. Default: space -t; -t\ -t'	-r reverse
-d dictionary-order	-n numeric sort lines of text	

14

Remove duplicate lines

uniq

- **more FILE**

```
chr6.fa 34314346 F
chr6.fa 52151626 R
chr6.fa 81889764 R
chr6.fa 52151626 R
```

- **uniq FILE**

```
chr6.fa 34314346 F
chr6.fa 52151626 R
chr6.fa 81889764 R
chr6.fa 52151626 R
```

-u	unique
-d	repeated

- **sort FILE**

```
chr6.fa 34314346 F
chr6.fa 52151626 R
chr6.fa 52151626 R
chr6.fa 81889764 R
```

- **sort FILE |uniq**

```
chr6.fa 34314346 F
chr6.fa 52151626 R
chr6.fa 81889764 R
```

- **sort FILE | uniq -d**

```
chr6.fa 52151626 R
```

- **sort FILE |uniq -u**

```
chr6.fa 34314346 F
chr6.fa 81889764 R
```

15

Print number of lines in files: **wc -l**

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |sort -k 2,2d -k 3,3n |head -2
```

```
+ chr15 3003325 GCCCAGAGTCCCACAGCCTGCTGCCT
+ chr15 3005096 GCAGTGGAAATTTCTTTTGTGTTAC
```

```
# seq only
```

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |cut -f4 |head -1
```

```
GTAAAACTTTATCTGCTGGCTGTCC
```

```
# seq count in chr15
```

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |cut -f4 |wc -l
```

```
101529
```

```
# count unique seq
```

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |cut -f4 |sort |uniq -u |wc -l
```

```
89604
```

```
# count duplicated seq
```

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |cut -f4 |sort |uniq -d |wc -l
```

```
4575
```

```
# total seq
```

```
byuan@tak /nfs/BaRC/byuan$ cut -f2-5 mapped.txt |grep "chr15" |cut -f4 |sort |uniq |wc -l
```

```
94179
```

16

awk

Alfred Aho, Peter Weinberger and Brian Kernighan

Awk program has the general form:

```

BEGIN                {<initializations>}
<search pattern 1>  {<program actions>} or
{if <search pattern 1> <program actions>}
END                  {<final actions>} ' file_name

```

Default: field separated by space,
Action: default print line (record)

17

awk

Alfred Aho, Peter Weinberger and Brian Kernighan

Relational Operators

Operator	Meaning
==	Is equal
!=	Is not equal to
>	Is greater than
>=	Is greater than or equal to
<	Is less than
<=	Is less than or equal to

Binary Operators

Operator	Type	Meaning
+	Arithmetic	Addition
-	Arithmetic	Subtraction
*	Arithmetic	Multiplication
/	Arithmetic	Division
%	Arithmetic	Modulo

Regular Expression Operators

Operator	Meaning
~	Matches
!~	Doesn't match

Boolean operators

Operator	Meaning
&&	AND
	OR

18

awk

Alfred Aho, Peter Weinberger and Brian Kernighan

```

byuan@tak$ head -1 mapped.txt
SRR015146.1_WICMT-SOLEXA_8_3_1_908_882_length=26 - chrX 79418719
GGCCAATTCACACTCTAATCCACTTC IDIIIIIIIIIIIIIIIIII 0
byuan@tak$ awk -F"\t" '{ print $3:"$4 }' mapped.txt|head -2
chrX:79418719
chr1:77169391
# count the occurrence of each position
byuan@tak$ awk -F"\t" '{ print $3:"$4 }' mapped.txt|sort|uniq -c|head -2
1 chr10:100002430
1 chr10:100005747
# max mapped position
byuan@tak$ awk -F"\t" '{ print $3:"$4 }' mapped.txt|sort|uniq -c|sort -k 1,1nr|head -2
1202 chr12:112722237
1202 chr13:112538649

```

19

awk

Alfred Aho, Peter Weinberger and Brian Kernighan

```

byuan@tak$ head -2 myfile
CHROM START STOP STRAND ID1 ID2 DISTANCE REGION
START REGION END PEAK POS PEAK HEIGHT TOTAL
TARGET COUNTS TOTAL BACKGROUND COUNTS
20 604823 590239 -1 NM_03312 BGN 600 589490 589540 589495
11.0 50.0 5.1
# number of genes with peak in chr20
byuan@tak$ awk '{if($1==20) print $6 }' myfile |sort|uniq|wc -l
102
# first gene in chr20 with peak height above 50, show its record and region range
byuan@tak$ tail -line=+2 myfile |awk '{ if($1==20 && $11>50) print $0"\t"$9-$8 }' myfile |head -1
20 48560297 48634493 1 NM_00282 BZD 0 48591510
48592010 48591715 80.0 2295.0 70.0 500

```

20

awk

Alfred Aho, Peter Weinberger and Brian Kernighan

```
byuan@tak$ head -2 data.txt
PROBE Control Exp
1007_s_at 10.14 10.11      Field separated by tab
# exp-control
byuan@tak$ tail --line=+2 data.txt | awk -F"\t" '{ print $0"\t"$3-$2 }' | head -2
1007_s_at 10.14 10.11 -0.03
1053_at 10.35 10.27 -0.08      whole record
# exp > control ?
byuan@tak$ tail --line=+2 data.txt | awk -F"\t" '{ if ($3>$2) print $0"\t"$3-$2 }' | head -2
1316_at 5.35 5.42 0.07
1487_at 8.70 8.77 0.07      number of current record
# which line?
byuan@tak$ tail --line=+2 data.txt | awk -F"\t" '{ if ($3>$2) print NR"\t"$0"\t"$3-$2 }' | head -1
8 1316_at 5.35 5.42 0.07
# max: exp > control
byuan@tak$ tail --line=+2 data.txt | awk -F"\t" '{ if ($3>$2) print NR"\t"$0"\t"$3-$2 }' | sort -k 5,5nr | head -2
44254 235003_at 6.26 9.28 3.02
36121 226864_at 5.36 8.36 3.00
```

21

awk

Alfred Aho, Peter Weinberger, and Brian Kernighan

```
byuan@tak$ awk '{ if($2>10 && $3>10) print $0 }' data.txt | head -3
PROBE Control Exp
1007_s_at 10.14 10.11
1053_at 10.35 10.27
# probe with the highest difference between exp and control and above 10
byuan@tak$ awk '{ if($2>10 && $3>10) print $0"\t"$3-$2 }' data.txt | sort -k 4,4nr | head -1
224691_at 10.10 12.41 2.31
```

```
# sum, average
byuan@tak$ awk '{ sum=sum+$2} END{print sum"\t"sum/NR}' data.txt
345622 6.32127
byuan@tak$ awk '{ conSum=conSum+$2; expSum=expSum+$3} END{print conSum"\t"conSum/NR"\t"expSum"\t"expSum/NR}' data.txt
345622 6.32127 345473 6.31855
```

22

awk

Alfred Aho, Peter Weinberger, and Brian Kernighan

```
byuan@tak$ awk '{ if($2=="+" && $3=="chr15") print $0 }' mapped.txt | head -1
SRR015146.15_WICMT-SOLEXA_8_3_1_33_728_length=26 + chr15 22686174
GTGGTAAACAATAATCTGCGCATGT I***** 2117
byuan@tak$ awk '{ if($2=="+" && $3=="chr15") print $0 }' mapped.txt | cut -f4 | sort -n | head -3
3000388
3001318
3001504
byuan@tak$ awk '{ if($2=="+" && $3=="chr15") print $0 }' mapped.txt | cut -f4 | sort -n | awk '{ print $1"\t"$1-pre; pre=$1 }' | head -3
3000388 3000388
3001318 930
3001504 186
byuan@tak$ awk '{ if($2=="+" && $3=="chr15") print $0 }' mapped.txt | cut -f4 | sort -n | awk '{ print $1"\t"$1-pre; pre=$1 }' | tail --line=+2 | sort -k 2,2nr | head -3
51360861 61343
67999814 60245
71200190 59915
```

23

split a big file into pieces

split [OPTION] [INPUT [PREFIX]]

- `wc -l FILE`
50000
- `split -l 10000 FILE | wc -l *` (default PREFIX is `x')
50000 FILE
10000 xaa
10000 xab
10000 xac
10000 xad
10000 xae
- `split -l 10000 -d FILE "FILE_" | wc -l FILE*`
50000 FILE
10000 FILE_00
10000 FILE_01
10000 FILE_02
10000 FILE_03
10000 FILE_04

-l	put NUMBER lines per output file
-d	use numeric suffixes instead of alphabetic

24

Concatenate files

cat

- `cat file1 file2 file3 > bigFile`

- more file

```
A   it
B   his
D   her
```

- `cat -A file`

```
A^Iit$
B^Ihis$
D^Iher$
```

-A	show all
^I	TAB (t)
\$	end of line (\$)
^M	carriage return(r)

25

Compress files

- Compress files:
 - `tar -cvf tarfile directory`
 - `gzip file_name`
- Display: `zmore data.txt.gz`
- Compare files: `zdiff data1.gz data2.gz`
- Search expression:
 - `zgrep 'NM_000020' data.gz`
- Decompress files:
 - `gunzip file.gzip`
 - `tar -xvf file.tar`

26

Get organized

- Make a directory
 - `mkdir my_data`
- Remove a directory (after emptying)
 - `rmdir my_data`
- Move (rename) a file or directory
 - `mv oldFile newFile`
- Copy a file
 - `cp oldFile newFileCopy`
- Remove (delete) a file
 - `rm oldFile`

27

Others

- Use up arrow, down arrow to re-use commands
- To get a blank screen: `clear`
- To get help (manual) command: `man`
- Avoid filenames with spaces
 - If necessary to use, refer to with quotes:
 - `"My dissertation version 1 .txt"`

28

commands

<code>ls</code>	<code>pwd</code>	<code>chmod</code>	<code>ln</code>
<code>cp</code>	<code>mv</code>	<code>rm</code>	<code>mkdir</code>
<code>rmdir</code>	<code>more</code>	<code>head</code>	<code>tail</code>
<code>cat</code>	<code>split</code>	<code>cut</code>	<code>paste</code>
<code>sort</code>	<code>uniq</code>	<code>wc</code>	<code>grep</code>
<code>gzip</code>	<code>gunzip</code>	<code>tar</code>	<code>zmore</code>
<code>zdiff</code>	<code>zgrep</code>	<code>man</code>	<code>clear</code>

29

Further Reading

- BaRC: Getting Started with UNIX
 - http://iona.wi.mit.edu/bio/education/unix_intro.html
- BaRC: Connecting to tak and transferring files
 - <http://jura.wi.mit.edu/bio/education/docs/ssh-sftp.html>
- BaRC: Tips and Tricks for bioinformatics
 - <http://iona.wi.mit.edu/bio/bioinfo/scripts/#unix>
- UNIX Tutorial for Beginners
 - <http://www.ee.surrey.ac.uk/Teaching/Unix/>
- Using the UNIX Operation System
 - http://stein.cshl.org/genome_informatics/unix1/index.html
 - http://stein.cshl.org/genome_informatics/unix2/index.html

30