Introduction to Unix

BaRC Hot Topics
Bioinformatics and Research Computing
Whitehead Institute
October 19th 2017
http://barc.wi.mit.edu/hot_topics/
Outline

• Unix overview
  – Logging in to tak4
  – Directory structure
  – Accessing Whitehead drives/finding your lab share

• Basic commands

• Exercises

• BaRC and Whitehead resources

• LSF
What is unix? and Why unix?

• Unix is a family of operating systems that are related to the original UNIX operating system
• It is powerful, so large datasets can be analyzed
• Many repetitive analyses or tasks can be easily automated
• Some computer programs only run on the unix operation system
• TAK4 (our unix server): lots of software and databases already installed or downloaded
Where can unix be used?

• Mac computers
  Come with unix

• Windows computers need Cygwin or PuTTY:
  \grindhouse\Software\Cygwin\Wibr-Cygwin-Installer-1010.exe
  https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html

• Dedicated unix server
  “tak4”, the Whitehead Scientific Linux
Logging in to tak4

• Requesting a tak4 account
  http://iona.wi.mit.edu/bio/software/unix/bioinfoaccount.php

• Windows
  ➢ PuTTY or Cygwin
  ➢ Xming: setup X-windows for graphical display

• Macs
  ➢ Access through Terminal
Connecting to tak for Windows

host: tak4.wi.mit.edu

1) Open putty

2) Write the address of the Host
   2.1) Write the address of the Host
   2.2) Click on SSH->X11-> Enable X11 forwarding

3) Click “Open”

When you write the password you won’t see any characters being typed.

Command
Prompt
user@tak4 ~$
Log in to tak4 for Mac

ssh -Y username@tak4.wi.mit.edu
Unix Directory Structure


nfs and lab are directories specific to Whitehead
Accessing Shared Resources at Whitehead

- Unix
  - /nfs/BaRC_Public
  - /nfs/BaRC_training
  - /lab/solexa_public

- Windows (access using Start Menu → Search)
  - \wi-files1\BaRC_Public
  - \wi-files1\BaRC_training
  - \wi-htdata\solexa_public

- Macs (access using Go → Connect to Server...)
  - cifs://wi-files1/BaRC_Public
  - cifs://wi-files1/BaRC_training
  - cifs://wi-htdata/solexa_public

Where’s my lab’s share?
- [http://it.wi.mit.edu/systems/file-storage/lab-share-paths](http://it.wi.mit.edu/systems/file-storage/lab-share-paths)
Get ready for the exercises

• Use this link to copy paste the commands for the exercises
  http://barc.wi.mit.edu/education/hot_topics/hot_topics/unix_es
tentials_2017/UnixEssentials_HandsOn.txt

• These folders contain today’s slides and the material for the
  exercises
  \\wi-files1\BaRC_Public\Hot_Topics\unixEssentials_2017
cifs://wi-files1/BaRC_Public/Hot_Topics/unixEssentials_2017
Unix Tips

• Use ↑↓ to reuse previous commands

• Ctrl-c: stop a process that is running

• Tab-completion:
  – Complete commands/file names

• Unix is case-sensitive
Basic commands
listing and organizing files/folders

- list the contents of a directory:
  - `ls` [only show names]
  - `ls -l` [long listing: show other information too]
  - `ls -h` [human readable]
- make a directory: `mkdir dirname`
- change directory: `cd dirname`
  - The directories . and ..
  - The current directory (.) `cd .`
  - The parent directory (..)
    - `cd ..` (go to the parent directory) `ls ..` (list to the parent directory)
- print working directory: `pwd`
- Go to your home directory: `cd` or `cd ~`
- Go to the previous directory: `cd -`
Paths

• “Absolute path” example
  
  /nfs/BaRC_training/userName/myWork
  /nfs/BaRC_training/userName/myFig

  cd /nfs/BaRC_training/userName/myWork
  cd /nfs/BaRC_training/userName/myFig

• “Relative path” example
  
  ../dirName
  if I am in “/nfs/BaRC_training/userName/myWork” I can do:

  cd ../userName/myFig
Unix Directory Structure

```
cd /nfs/BaRC_training/userName/myFig

cd ../userName/myFig
```
Permissions

- Use **chmod** to change permissions
  - `user(u), group(g), others(o), all(a)`
  - `chmod u+x foo.pl` (user can execute)
  - `chmod g-w foo.pl` (group can’t write)

```
thiruvil@tak /nfs/BaRC_Public$ ls -l myFile.txt
-rw-r--r-- 1 thiruvil barc 0 2012-10-10 13:32 myFile.txt
thiruvil@tak /nfs/BaRC_Public$ chmod g+w myFile.txt
thiruvil@tak /nfs/BaRC_Public$ ll myFile.txt
-rw-rw-r-- 1 thiruvil barc 0 2012-10-10 13:32 myFile.txt
```
Basic commands

copying, moving files, getting help

• copy: `cp`
• move: `mv`
• remove: `rm`
• remove directory: `rmdir`
• get help on a command: `man {command}`

*Do Exercises 1, 2 and 3*
Displaying the contents of a file on the screen

- **cat** filename: Dump a file to the screen
- **more** filename: Progressively dump a file to the screen: ENTER = one line down SPACEBAR = page down q=quit
- **less** filename: like more but with extended capabilities
- **head** filename: Show the first few lines of a file
- **head -n** filename: Show the first n lines of a file
- **tail** filename: Show the last few lines of a file
- **tail -n** filename: Show the last n lines of a file
- **clear**: clear screen
Editing a File

• Command-line editors
  ➢ pico
  ➢ nano
  ➢ emacs (emacs -nw)
  ➢ vi

• Graphical editors (Windows users need an X-windows emulator)
  Note: may not be part of standard installation
  ➢ nedit
  ➢ gedit
  ➢ xemacs

• Put an & at the end of command line to run it in the background when using a graphical editor so that you can continue to use the terminal window

  eg. gedit myFile.txt&
Output Redirection and Piping

• Write output of a command to file
  - Write to output file
    • sort myfile.txt > myfile_sorted.txt
  - Over-write output file (if it exists)
    • sort myfile.txt >| myfile_sorted.txt
  - Append to output file
    • sort myfile.txt >> myfile_sorted.txt

• Piping “|”: use output of one command as input for another command
  - sort myfile.txt | more
Parsing a File: cut
Word count: wc

• Select columns of interest
  `cut -f 9,12-15 myGeneValues.txt > col_9.12to15.txt`
  Options:
  -f output only these fields
  -d field delimiter

• Count number of lines/words/characters in file
  `wc myFile`
  `wc -w (count words only)`
  `wc -l (count lines only)`

Do Exercises 4, 5, 6 and 7
Sorting and removing redundancy
sort and uniq

• Sort on column(s)
  
  \texttt{sort \ -k 3,3 myGeneExpression.txt | more}

  Options:
  - \texttt{-n, --numeric-sort} compare according to string numerical value
  - \texttt{-g, --general-numeric-sort} compare according to general numerical value
  - \texttt{-r reverse}
  - \texttt{-k pos1,pos2} start sorting at pos1, end it at pos2

• Get only unique entries
  
  \texttt{uniq mySortedGenes.txt > myUniqGenes.txt}

  Options:
  - \texttt{-c count entries}
  - \texttt{-d duplicate counts}

  make sure that the file is sorted before running uniq

\textit{Do Exercises 8 and 9}
Searching the contents of a file

- **grep** (global regular expression print)
  Find words, or patterns, occurring in lines of a file

  grep TMEM geneList.txt
  TMEM131
  TMEM9B
  TMEM14C
  TMEM66
  TMEM49

**Options:**
- `-v` select non-matching lines
- `-i` ignore case
- `-n` print line number

Example: get TMEM but exclude TMEM14C

  grep TMEM geneList.txt | grep -v "TMEM14C" | more

*Do Exercise 10*
Getting Files

• Getting files or directories
  ➢ Files
    wget http://data.broadinstitute.org/igv/projects/downloads/2.4/IGVSource_2.4.2.zip

  ➢ Directories from (outside) servers
    scp -r origin destination
    scp -r userName@serverToCopyFrom:/pathToFolderToCopyFrom .
    scp -r jdoe@copper.broadinstitute.org:/broad/lab/works .


(Un)Compressing Files

• .gz file
  Compress: `gzip file.txt` (file.txt.gz will be created)
  Uncompress: `gunzip file.txt.gz` (file.txt will be created)

• .tar.gz file
  Compress: `tar -czvf myFiles.tar.gz myFiles`
  Uncompress: `tar -xzvf myFiles.tar.gz`

  Options
  -c create an archive
  -x extract an archive
  -f FILE name of archive
  -v be verbose, list all files being archived/extracted
  -z create/extract archive with gzip/gunzip

• View compressed files using:
  `zmore`, `zgrep`
BaRC Resources

• barc.wi.mit.edu
BaRC SOPs

http://barcwiki.wi.mit.edu/wiki/SOPs

BaRC Standard operating procedures

These are "how-to's" detailing the methods that BaRC uses and finds to work effectively. Email BaRC if you have any questions about how or why to perform what is described on these pages.

Short read sequencing

- Quality control and preprocessing of short reads (fastq files)
- Mapping short reads
- SAM/BAM summarizing and processing
- SAM/BAM quality control
- Using ChIP-Seq to identify and/or quantify bound regions (peaks)
- Using RNA-Seq to quantify gene levels and assay for differential expression
- Using RNA-Seq to assemble or annotate transcripts
- Integrating expression and Immunoprecipitation experiments
- Creating genome feature heatmaps from sequencing experiments
- Creating an analysis pipeline of compressed files

Variant calling and analysis
Unix Commands and BaRC Scripts
Running Scripts on Unix

- **Perl**
  
  `bed2gff.pl` or `./bed2gff.pl` or `path/bed2gff.pl`

- **R**
  
  `run_rma_customCDF.R`

- **Python**
  
  `myScript.py` or `./myScript.py`

- **Matlab**
  
  `matlab -nodesktop -nosplash myScript.m`

- **Java Archive (JAR)**
  
  `java -Xmx1000m -jar /usr/local/share/IGVTools/igv.jar`
## BaRC code

<table>
<thead>
<tr>
<th>Script type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perl</td>
<td>/nfs/BaRC_Public/BaRC_code/Perl</td>
</tr>
<tr>
<td>R</td>
<td>/nfs/BaRC_Public/BaRC_code/R</td>
</tr>
<tr>
<td>Python</td>
<td>/nfs/BaRC_Public/BaRC_code/Python</td>
</tr>
</tbody>
</table>
Running Programs/Tools on Unix

- **bedtools**
  ```
  bedtools intersect -a myGenes1.bed -b myGenes2.bed
  ```
  Other utilities: http://code.google.com/p/bedtools/wiki/Usage

- **samtools**
  ```
  samtools view myFile.bam
  ```
  Other utilities: http://samtools.sourceforge.net/samtools.shtml

- **Fastx toolkit**
  ```
  fastx_quality_stats -i mySeq.fastq -o fastxStats_mySeq
  ```

- **FastQC**
  ```
  fastqc mySeq.fastq
  ```

- **BLAST**
  ```
  blastp -task blastp -db myProtDB.fa -q myProt.fa -out out.txt
  ```
# Commonly Used Data Locations at Whitehead

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>/nfs/genomes</td>
<td>Genome data: gff, gtf, fasta, bowtie indexed files, blat indexed file, etc. for several organisms</td>
</tr>
<tr>
<td>/nfs/seq/Data</td>
<td>Sequence data, including blast databases, for several organisms</td>
</tr>
<tr>
<td>/nfs/BaRC_datasets</td>
<td>Large (array/NGS) datasets: HBI, HBM 2.0</td>
</tr>
</tbody>
</table>
LSF Commands

• **bsub to submit jobs**

  bsub wc -l reads.fq
  bsub "sort foo.txt > sorted.txt"

  **Options:**
  - `-e error_file`
  - `-o standard_out_file`
  - `-m machine` (send the job to that machine)
  - `-n number` (use that many processors in the cluster)
    - `bsub -n 4 -R "span[hosts=1]"` (use 4 processors all in the same machine in the cluster)

• **bjobs to view your jobs**

  `bjobs`

• **bkill to kill a job**

  `bkill 237878`
LSF cluster jobs

http://tak4.wi.mit.edu/

Welcome to Tak 4.0

Software installation requests or issues:
- Create a new ticket
- Get an Account
- Account Help

Installed software and modules:
- Installed Packages
- Installed Perl Modules
- Installed Python Modules
- Installed R Modules

Available Data Sets:
- Find genomes at /nfs/genomes
- Find biinformatics datasets at /nfs/BaRC_datasets

LSF Cluster Activity

<table>
<thead>
<tr>
<th>User</th>
<th>Running</th>
<th>Pending</th>
</tr>
</thead>
<tbody>
<tr>
<td>bellott</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>pincus</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>corradi</td>
<td>226</td>
<td>0</td>
</tr>
<tr>
<td>thiruv</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>wenguse</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>vcorbin</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>abraham</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Totals:</td>
<td>260</td>
<td>0</td>
</tr>
</tbody>
</table>

Last Updated: Mon, 16 Oct 2017 11:27:15

Host Summary:

<table>
<thead>
<tr>
<th>Host</th>
<th>Max</th>
<th>Running</th>
<th>Load</th>
<th>Threads</th>
<th>CPU</th>
<th>Memory</th>
<th>Uptime</th>
<th>System Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>it-c01b02</td>
<td>23</td>
<td>16</td>
<td>4</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>55 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c01b04</td>
<td>23</td>
<td>8</td>
<td>4</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>35 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c01b11</td>
<td>23</td>
<td>7</td>
<td>3</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>53 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c01b12</td>
<td>18</td>
<td>8</td>
<td>3</td>
<td>24</td>
<td>2.93GHz</td>
<td>160 GB</td>
<td>31 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c01b13</td>
<td>23</td>
<td>18</td>
<td>4</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>61 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c01b14</td>
<td>23</td>
<td>10</td>
<td>5</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>56 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c03b09</td>
<td>23</td>
<td>3</td>
<td>2</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>55 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c03b12</td>
<td>23</td>
<td>9</td>
<td>2</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>49 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c05b06</td>
<td>23</td>
<td>8</td>
<td>3</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>31 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
<tr>
<td>it-c05b10</td>
<td>23</td>
<td>3</td>
<td>3</td>
<td>24</td>
<td>2.93GHz</td>
<td>192 GB</td>
<td>59 days</td>
<td>HP ProLiant BL460c G7</td>
</tr>
</tbody>
</table>
Further Reading

• BaRC: Unix Info
http://iona.wi.mit.edu/bio/education/unix_intro.php

• LSF Cluster (incl. examples)

• Whitehead IT Computing Tutorials
http://it.wi.mit.edu/
Upcoming Hot Topics

• Unix advanced (Oct 25th)
• An Introduction to R - November
• An Introduction to R Graphics - November