"IBS 574 - Computational Biology & Bioinformatics"
Spring 2018, Tuesday (01/30), 2.00-4.00PM

Linux shell & shell scripting - I

Ashok R. Dinasarapu Ph.D
Scientist-Bioinformatics, Dept. of Human Genetics
Emory University, Atlanta

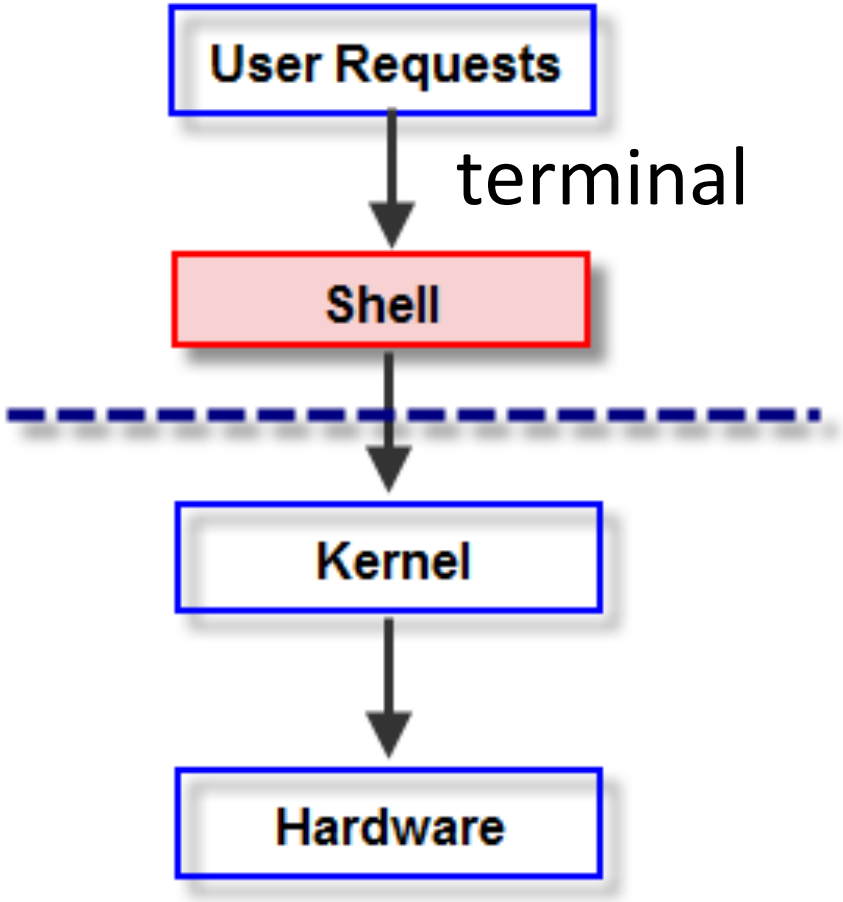
Kernel, Shell & Terminal

frequently used terms in UNIX-like operating systems (OS)

- **Kernel** is the core that provides basic services for all other parts of the OS.
- **Shell** is the outermost part of an OS that interacts with user commands
  - command line interpreter
  - sh: bash (bourne again shell), csh, ksh, dash
- **Terminal emulator** gives us access to the Shell
  - xterm, rxvt, genome-terminal
  - console = physical terminal

Kernel, Shell & Terminal

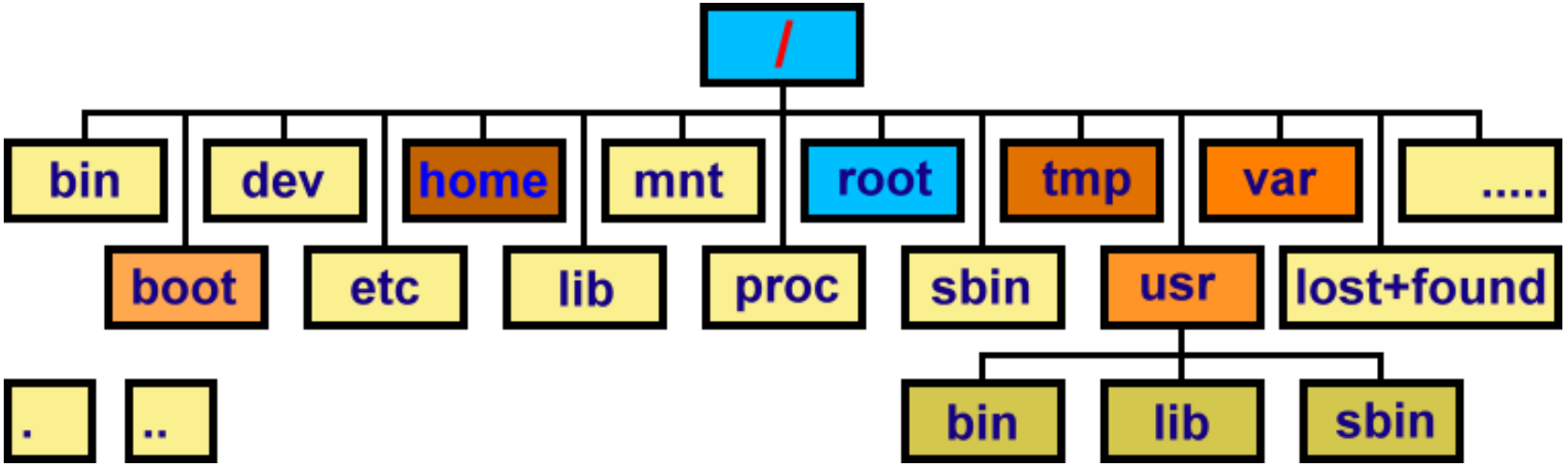
frequently used terms in UNIX-like operating systems (OS)



Console/Terminal

- Shell prompt will usually include

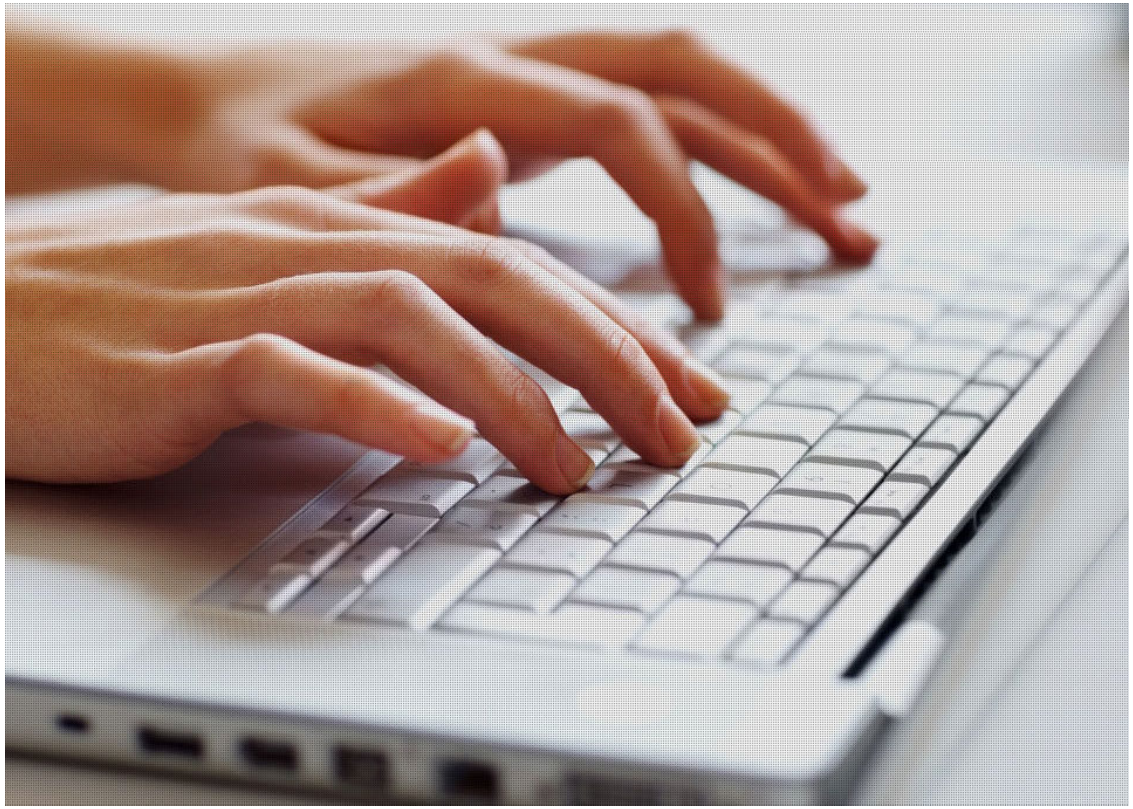
[root@machinename ~]#	/root
[user_name@machinename ~]$	/home/user_name



Linux file system hierarchy

Easiest way to do this is …

Start typing!



Connecting via SSH to your server

ssh *user_name*@blnx1.emory.edu

Alternatively, https://blnx1.emory.edu:22443/

user_name@blnx1:~$

~ means your home dir (/home/*user_name*)

SSH allows you to connect to your server securely and perform Linux command-line operations.

Console/Terminal

- **What is the default Shell for blnx1?**

  Usage: `ls –la /bin/*sh`

  /bin/sh -> dash

- **What is the default terminal emulator?**

  Usage: `ls -la /usr/bin/*terminal*`

  /usr/bin/gnome-terminal

- What is physical terminal?

Console/Terminal

Some simple commands:

- env (current environment)
- date (for date)
- cal (for calendar)
- df or free (disk space or memory)
- exit (close terminal/session)

Usage: `env` (then press enter key)

Navigation of the file system

Commands: **pwd, ls, cd**

- **pwd** (print/current working dir)
- **ls** (listing directories and files)

ls -l (long listing format)
ls -la (with hidden files)

Usage: pwd /home/user_name

Usage: ls -l

Navigation of the file system

Commands: **pwd, ls, cd, mkdir**

- **mkdir** make a sub-directory

Usage: `mkdir document`

mkdir **-p** document/resume
mkdir -p document/{resume,cv,address}

- **cd** change directory

Usage: `cd document/resume` (enter dir)

cd ../.. (return to previous dir)

Navigation of the file system

Commands: **pwd, ls, cd, mkdir**

If a directory name has white space –

```
mkdir ‘My Data’
```

Usage:

```
cd My\ Data
cd "My Data”
cd 'My Data'
```

- A non-quoted backslash, \, is used as an **escape character** in **Bash**.

Navigation of the file system

Commands: **pwd, ls, cd**

cd ~ (change to **home** dir)
cd (change to **home** dir)

cd / (change to **root** dir)

cd document (relative path)
cd ~/document (absolute path)
cd /home/user_name/document

*with $HOME environment variable

Create/Edit text files

Choose a text editor: emacs, Vim

Usage: `vi home.txt`

INSERT mode:
press keys like `i` OR `a` & start typing.

"i" will let you insert text just before the cursor.
"I" inserts text at the beginning of the current line.
"a" will let you insert text just after the cursor, and
"A" will let you type at the end of the current line.

Create/Edit text files

Type the following text:

Street: 201 Dowman Drive
City: Atlanta
State: Georgia
Country: USA
Zip: 30322

Create/Edit text files

SAVE mode:
press `esc` key AND

`:q!` for not to save OR

`:x` to save all typed content.

Create/Edit text files

Choose a text editor: emacs, Vim

Usage: `vi college.txt`

```
Street: 954 Gatewood Rd
City: Atlanta
State: Georgia
Country: USA
Zip: 30329
```

Listing files, directories & it's contents

- List files recursively

Usage: `ls -R`

document/resume
document/cv
document/address
home.txt
college.txt
My Data

Listing files, directories & it's contents

- List files recursively

Usage: `ls -R`

document/resume
document/cv
document/address
home.txt
college.txt
My Data

Copy
Move
Symlink

“cp” – keeps the original file & makes a duplicate of it

- Copy **a file**

  Usage: cp **home.txt** document/**address**/

- Copy **all files**

  Usage: cp ***.txt** documents/**address**/

CAREFUL! If the destination file already exists, it will be overwritten without a confirmation prompt. Use “cp -i” interactive option for a prompt.

Rename files & directories using “mv” move command

- Move a **file**

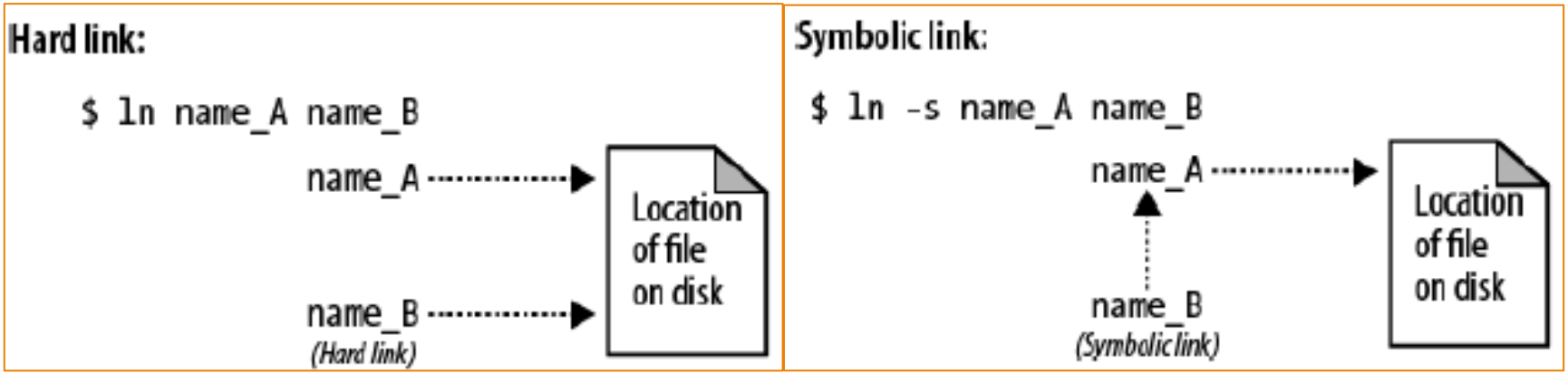
  Usage: mv **home.txt** document/**cv**/

- Move **all files**

  Usage: mv ***.txt** document/**cv**/

CAREFUL! If the destination file already exists, it will be overwritten without a confirmation prompt. Use “mv -n” interactive option for a prompt.

Symbolic link is a special kind of file that points to another file



- **ln** : creates hard link, default

ln ~/document/cv/home.txt **~/document**/resume/

- **ln -s**: symbolic link

ln -s ~/document/cv/college.txt **document**/resume/

The cat command is used for:

- **view/read**
  Usage: `cat home.txt`
- **concatenate & view**
  Usage: `cat home.txt college.txt`
- **concatenate & redirection**
  Usage: `cat home.txt college.txt > add.txt`

WARNING: if add.txt already exists, it will be over-written.
Use >> to append the output to the existing file.

“more or less”: terminal pager commands in Unix-like systems

- less (read with cursor)

  Usage: `less add.txt`
- more (read page by page)

  Usage: `more add.txt`

The program **less** or **more** does not require the whole file to be loaded in memory to view parts of it.

head & tail lets you see the first & last few lines of a text file

- **View** the first 5 lines
  Usage: `head –n5 add.txt`
- **View** the last 4 lines
  Usage: `tail –n4 add.txt`

By **default**, head or tail returns the first or last **10 lines**, respectively of each file that is provided to it.

“pipe & redirect”

- sending data from one program to another

  Usage: `cat home.txt | less`

- Redirecting to a file

  Usage: `head home.txt > lines.txt`

  `wc -l < home.txt`

What does "**> /dev/null 2>&1**" mean?

- Arithmetic in POSIX shells is done with $ and double parentheses:

```
echo $((2+2))
```

STDIN	(<0) : keyboard
STDOUT	(>1) : screen
STDERR	(>2) : screen

```
echo $((2+2)) > result.txt
```

(STDOUT redirected into a file, result.txt)

```
echo $((2+2)) > /dev/null 2>&1
```

(To listen only to STDERR (2>&1), STDOUT redirected into /dev/null & STDERR redirected to STDOUT).

What does "> /dev/null 2>&1" mean?

- Arithmetic in POSIX shells is done with $ and double parentheses:

echo $((2+2)) > /dev/tty

STDIN	(<0)
STDOUT	(>1)
STDERR	(>2)

/dev/tty is a special file, representing the terminal for the current process
Use, ps −a (to see processes attached to ttys)
In Linux, the console appears as several terminals (ttys)

echo $((2+2)) > result.txt

(STDOUT redirected into a file, result.txt)

echo $((2+2)) > /dev/null 2>&1

(To listen only to STDERR (2>&1), STDOUT redirected into /dev/null & STDERR redirected to STDOUT).

Create a file (test_error.sh) with the following two lines of code

```
#!/bin/sh
echo $((2+'x'))
```

Run as

```
./test_error.sh
./test_error.sh 2>/dev/tty
./test_error.sh 2>error_file
```

Searching a PATTERN with grep

"global regular expression print"

- **Search a file for keywords**

  Usage: `grep State add.txt`

- **ignore case option**

  Usage:
  ```
  grep state add.txt
  grep -i state add.txt
  ```

Regular expressions are used to search and manipulate the text, based on the **patterns** [Beginning of line (^) ; End of the line ($)]

“join two files”

- Joins the rows of two files which share a common field of data.

Usage: `vi country.txt`

```
1  India
2  USA
3  Ireland
4  UK
5  Canada
```

Usage: `vi city.txt`

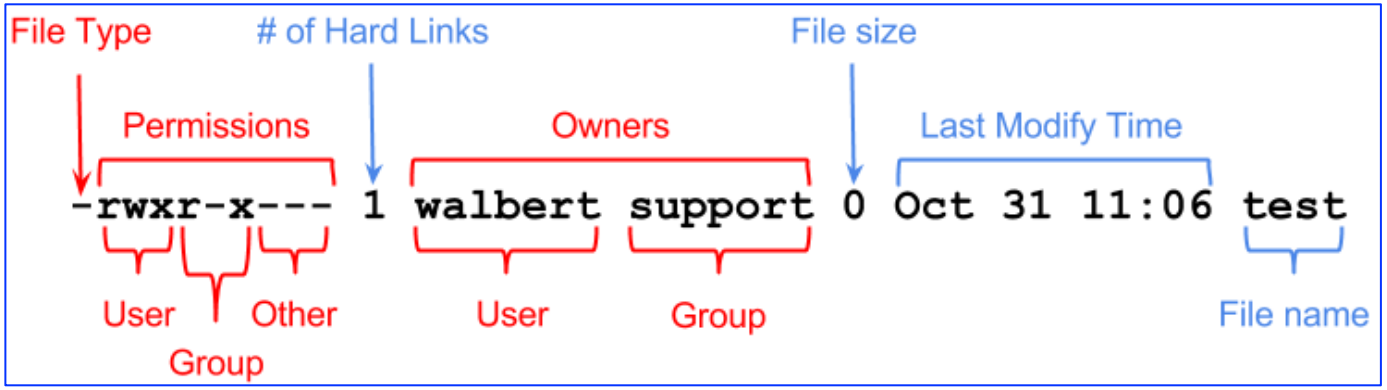
```
1  NewDelhi
2  WashingtonDC
3  Dublin
4  London
5  Toronto
```

Usage: `join country.txt city.txt > file.txt`

Change the permissions of files

- Three main type of permissions
  - **R**ead (r), **w**rite (w), and e**x**ecute (x)
  - 3 types of users (**u**ser, **g**roup & **o**ther)

Usage: `ls -l`



Change the permissions of files

- Three main type of permissions
  - **R**ead (r), **w**rite (w), and e**x**ecute (x)
  - 3 types of users (**u**ser, **g**roup & **o**ther)

Usage: chmod **u**=rwx,**g**=rx,**o**=r home.txt

Change the permissions of files

- Three main type of permissions
  - **R**ead (r), **w**rite (w), and e**x**ecute (x)
  - 3 types of users (**u**ser, **g**roup & **o**ther)

Usage: chmod **u**=rwx,**g**=rx,**o**=r home.txt

Usage: chmod 754 home.txt

4 stands for "read",
2 stands for "write",
1 stands for "execute", & **0** stands for "no permission"

Delete files & directories

- **rm** removes files blindly, with no concept of 'trash'!!!
- **Remove a file**
  Usage: `rm document/cv/add.txt`
- **Remove directory recursively** along with all of its contents
  Usage: `rm -r document`

Write-protected files prompt the user for a confirmation (with a *y* and an *n*) before removal. Use *-f* (i.e., *force*) to remove all specified files, whether write-protected or not.

How to setup command aliases in Linux

- Open your *.bashrc* vi ~/.bashrc
- **Add the alias**

  alias u="cd ..;ls”
- **Path adjustments**

  export PATH="$PATH:/<here_new_path>”
- Reload the “.bashrc” file

  source ~/.bashrc

Data archiving in Linux

List of archive and/or compression formats

File extension	Official name	Description
.tar	Tape archive	Archiving
.bz2	bzip2	Compression
.gz	gzip	Compression
.tar.gz (.tgz)	tar with gzip	both
.tar.bz2	tar with bzip2	both
.zip	ZIP	both

Data/File compression

- To compress a single file (.gz)

Usage: gzip **home.txt**

- Archive & compress

Usage: tar -**jc**vf home.tar.bz2 **home.txt**

j: use bzip2 compress (z: use gzip compress)
c: compress
v: verbose
f: file

Data/File decompression

Download file using wget

```
wget https://github.com/samtools/samtools/releases/download/1.3.1/samtools-1.3.1.tar.bz2
```

- View the contents of a tar.bz2 file

Usage: `tar -j**t**vf samtools-1.3.1.tar.bz2`

- Extract the contents

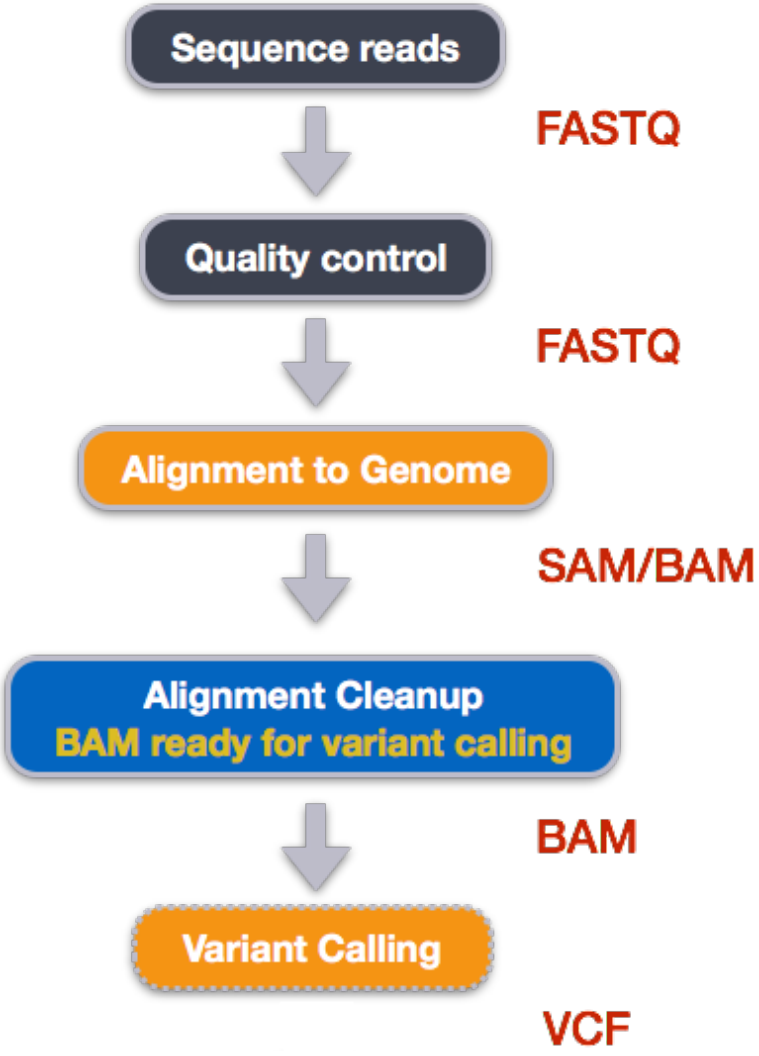
Usage: `tar -j**x**vf samtools-1.3.1.tar.bz2`

t: list the contents; x: extract the contents of an archive

What is SAMTOOLS?

- **SAMTOOLS** is a suite of programs for interacting with high-throughput sequencing data.
- Uses **SAM** (Sequence Alignment/ Map)/**BAM** (Binary Alignment/ Map) files.
- Variant Call **Format** (**VCF**), most likely stored in a compressed manner.

```
samtools mpileup | \
    bcftools call > name.vcf
```



VCF file compression with BGZip & indexing it with Tabix

- Compressing VCF files with **BGZip** and indexing it with **Tabix** is the standard way VCF files are stored.
- **bgzip** – Block compression/decompression utility.
- **tabix** – Generic indexer for TAB-delimited genome position files.
- Download BGZip'ped VCF file from http://vat.gersteinlab.org/datasets.php

VCF file compression with BGZip & indexing it with Tabix

- **View/Read compressed file**

```
gzip -cd CEU.low_coverage.2010_07.indel.genotypes.vcf.gz | head -20
zcat CEU.low_coverage.2010_07.indel.genotypes.vcf.gz | head -20
```

- **Decompress, gzip (compress) & index**

```
gunzip CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
gzip CEU.low_coverage.2010_07.indel.genotypes.vcf
tabix CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
Not a BGZF file: CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
tbx_index_build failed: CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
```

- **Decompress, bgzip (compress) & index**

```
gunzip CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
bgzip CEU.low_coverage.2010_07.indel.genotypes.vcf
tabix CEU.low_coverage.2010_07.indel.genotypes.vcf.gz
CEU.low_coverage.2010_07.indel.genotypes.vcf.gz.tbi
```

Rsync (Remote Sync) command for copying and synchronizing files & directories remotely as well as locally in **Linux**/Unix systems

```
rsync –av ~/script/CorrPlot.R adinasarapu@blnx1.emory.edu:~/
```

adinasarapu@blnx1.emory.edu's password:

building file list ... done
CorrPlot.R

sent 3310 bytes received 42 bytes 103.14 bytes/sec
total size is 3177 speedup is 0.95

Practice Makes Perfect