job monitor and control

top: similar to windows task manager (space to refresh, q to exit)

w: who is there

ps: all running processes, PID, status, type
ps -ef | grep yyin

bg: move current process to background

fg: move current process to foreground

jobs: list running and suspended processes

kill: kill processes
kill pid (could find out using top or ps)

sort, cut, uniq, join, paste, sed, grep, awk, wc, diff, comm, cat

All types of bioinformatics sequence analyses are essentially text processing.

Unix Shell has the above commands that are very useful for processing texts and also allows **the output from one command to be passed to another command as input using pipe** (“|”).

```
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | cut -f1 | sort -u | wc
```

This makes the processing of files using Shell very convenient and very powerful: you do not need to write output to intermediate files or load all data into the memory.

For example, combining different Unix commands for text processing is like passing an item through a manufacturing pipeline when you only care about the final product

Hands on example 1: cosmic mutation data

- Go to UCSC genome browser website: http://genome.ucsc.edu/
- On the left, find the Downloads link
- Click on Human
- Click on Annotation database
- Ctrl+f and then search “cosmic”
- On “cosmic.txt.gz” right-click -> copy link address
- Go to the terminal and wget the above link (middle click or Shift+Insert to paste what you copied)
- Similarly, download the “cosmicRaw.txt.gz” file

- Under your home, create a folder called class (mkdir)
- Under home/class, create a folder called mar19 (mkdir)
- Move the above downloaded files to home/class/mar19 (mv)
- Change to that directory (cd)

```
zless cosmic.txt.gz (q to exit)
zless cosmicRaw.txt.gz
gzip -d *.gz
less cosmicRaw.txt
```

http://rous.mit.edu/index.php/Unix_commands_applied_to_bioinformatics

```
awk 'condition {action}'
```

```
less cosmicRaw.txt | cut -f2
(ctrl+c to stop, or ctrl+z to suspend and then jobs, then kill -9 %1)

less cosmicRaw.txt | cut -f2 | less
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | less
```

(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cosmicRaw.sql)

```
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | less
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | cut -f1 | sort -u | wc
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | awk '$6=="liver"`
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | cut -f5 | less
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | cut -f5 | sort | uniq -c
less cosmicRaw.txt | cut -f2,3,4,5,8,13 | cut -f5 | sort | uniq -c | sort -k
1,1nr

less cosmicRaw.txt | cut -f2,3,4,5,8,13 | cut -f5 | sort | uniq -c | sort -k
2,2n

less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | cut -f6 | sort | uniq
-c | sort -k 1,1nr

less cosmicRaw.txt | cut -f2,3,4,5,8,13 | awk '$5==22' | cut -f2 | sort | uniq
-c | sort -k 1,1nr | less
```

Hands on example 2: process fasta sequence data

- Download genome data of multiple e.coli k-12 strains

```
wget -q -r ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12* &
```

put the process to background

- List

```
ls -l ftp.ncbi.nih.gov/genomes/Bacteria/
```

- List

```
ls -l
ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/
```

- Find all .faa files

```
find ftp.ncbi.nih.gov/ -name *faa
```

- Count

```
find ftp.ncbi.nih.gov/ -name *faa | wc
```

- Cat all protein sequences into one large file

sed - stream editor

```
find ftp.ncbi.nih.gov/ -name *faa| xargs cat | less
find ftp.ncbi.nih.gov/ -name *faa| xargs cat > ecoli-all.faa
```

- Count how many proteins

```
find ftp.ncbi.nih.gov/ -name *faa| xargs cat | grep '>' | wc -l
```

- View the protein description lines

```
find ftp.ncbi.nih.gov/ -name *faa |xargs cat | grep '>' | less
find ftp.ncbi.nih.gov/ -name *faa |xargs cat | grep '>' | cut -f1 -d ' ' | less

find ftp.ncbi.nih.gov/ -name *faa |xargs cat | grep '>' | cut -f1 -d ' ' | sed
's/>//' | head

find ftp.ncbi.nih.gov/ -name *faa |xargs cat | grep '>' | cut -f1 -d ' ' | sed
's/>//' | cut -f4 -d'|' | head
```

for loop on command line

```
for variable in `command`
do
        command 1
        command 2
done
```

The symbol on the tilde key (~)

`

http://en.wikipedia.org/wiki/Grave_accent

Or backtick

```
for x in `find ftp.ncbi.nih.gov/ -name *faa`
do
        echo $x
        cat $x | grep '>' | wc -l
done
```

```
for x in `find ftp.ncbi.nih.gov/ -name *faa`; do echo $x; cat $x | grep '>' | wc -l; done
```

```
for x in `find ftp.ncbi.nih.gov/ -name *faa`
do
        cat $x >> ecoli-all.faa.2
done
```

```
find ftp.ncbi.nih.gov/ -name *faa| xargs cat > ecoli-all.faa
```

- Save history of your commands:
history | less
history > hist1

- Send message to other online users
write username (ctrl+c to exit)

- Change your password
passwd

Ctrl+c to tell Shell to stop current process
Ctrl+z to suspend
bg to send to background
Ctrl+d to exit the terminal (logout)

More example:

Go to the webpage of CAZy family GH5: http://www.cazy.org/GH5_all.html

Copy & paste the entire page to an excel spreadsheet (ctrl+a, Ctrl+c, Ctrl+v)

There are four pages; do this for all four pages, append them in one spreadsheet

Copy the col D (GenBank IDs), and paste to vi editor to save the protein ids as a file

Practice sed, sort, uniq etc. to get a unique set of GenBank IDs of GH5

Install external command line softwares

Yanbin Yin

Spring 2013

Programs/tools we often use

- BLAST
- FASTA
- HMMER
- EMBOSS
- lftp
- bioperl
- R
- Galaxy
- Clustalw
- MAFFT
- MUSCLE
- SRA toolkit
- weblogo
- PhyML
- FastTree
- RaxML
- USEARCH
- …

http://gacrc.uga.edu/

Linux-based program types

- Source codes in C, C++, Java, Fortran etc.
  - Need to be compiled before execute the command
- Precompiled executables or binary codes
- Source codes in scripting languages (perl, python, R etc.)
  - Can execute directly

On your own ubuntu machine …

- You are the root and using the sudo command you can install anything you want into the system directory (/usr/bin/, /bin/, /lib/ etc.)
  - apt-get (Advanced Packaging Tool) can do many installations for you from source or binary codes

    http://www.digimantra.com/howto/apple-aptget-command-mac/

- On glu, you are not the root and you can only install things under your home using the "hard" or the "most common" way
  - Download->unpack->install->edit PATH environmental variable
  - Make sure you create folders for each tools, e.g. home/tools/fasta

Install BLAST on your own machine

sudo apt-get install blast2

blastall

which blastall

sudo apt-get remove blast2

blastall

[new version of blast]

sudo apt-get install ncbi-blast+

[Run the following before and after installation]

```
ls /usr/bin/| wc
```

Use apt-get to install

lftp, emboss, hmmer, bioperl, clustaw, muscle, R

fasta

[not available in the package list]

Install FASTA using the common way

[download]
http://fasta.bioch.virginia.edu/

wget -q http://faculty.virginia.edu/wrpearson/fasta/CURRENT/fasta-36.3.5e.tar.gz

mkdir tools
cd tools
mkdir fasta
mv fasta-36.3.5e.tar.gz fasta
cd fasta

[unpack]
tar zxf fasta-36.3.5e.tar.gz

```
[compile/install]
cd fasta-36.3.5e/
ls -l
ls -l ../bin/
less README
cd src
make -f ../make/Makefile.linux_sse2
all
cd ../bin/
ls -l
ssearch
cd
ssearch
```

```
[edit path variable]
vi .bashrc
export PATH="absolute path to fasta
bin folder";
. .bashrc
ssearch
```

Install BLAST using the common way

lftp ftp.ncbi.nih.gov:/blast/executables/LATEST> get ncbi-blast-2.2.27+-ia32-linux.tar.gz

tar -zxf ncbi-blast-2.2.27+-ia32-linux.tar.gz

ls -l

cd ncbi-blast-2.2.27+/bin

ls -l

./blastp -h

Download ncbi-blast-2.2.27+-x64-linux.tar.gz if your machine is 64 bit

uname -a

```
[edit path variable]
vi .bashrc
export PATH="absolute path to blast bin folder";
. .bashrc
blastp
```

Install HMMER

sudo apt-get install hmmer

Hard way:

http://hmmer.janelia.org/software

bioperl

http://www.bioperl.org/wiki/Installing_BioPerl_on_Ubuntu_Server

sudo apt-get install bioperl

The hard way to install bioperl

```
wget -q http://bioperl.org/DIST/current_core_unstable.tar.bz2
tar -xjvf current_core_unstable.tar.bz2
cd bioperl-*
perl Build.PL   # choose the defaults
./Build test
./Build install
```

http://www.bioperl.org/wiki/Installing_BioPerl_on_Ubuntu_Server

Install MAFFT the hard way

wget -q http://mafft.cbrc.jp/alignment/software/mafft-7.029-with-extensions-src.tgz

tar xzf mafft-7.029-with-extensions-src.tgz

cd mafft-7.029-with-extensions/core/

sudo make
sudo make install

http://mafft.cbrc.jp/alignment/software/source.html

Install Galaxy

- http://wiki.galaxyproject.org/Admin/Get%20Galaxy