Unix Tricks & Text Processing

Please note that MS Word and Google Docs will change the quote chars to ‘curly quotes’ which Unix does not like

Navigation

```
pushd/popd
mkdir -p path/to/newdir
# go to the beginning of your command prompt:
ctrl+a
# go to the end of prompt
ctrl+e
# move within a line
Esc-f    # forward one word
Esc-b     # back one word
Esc-d     # delete one word
```

Commands

!!, !*history#*, !*cmd*, and !$

```
!! # execute last command
history 50 # last 50 commands
!495 # run cmd #495
!srun # last srun cmd
# finding a command executable in the filesystem
which cmd
locate cmd
# look at the details of that file in its location in the filesystem
ls -al `which cmd`
```

Moving/copying files

```
# using rsync, best cmd-line tool for large copy jobs. Pay attention to final slashes
rsync -av sourcedir targetdir     # copy sourcedir and subs into targetdir
rsync -av sourcedir/ targetdir    # copy sourcedir contents into targetdir
rsync -av sourcedir username@host:targetdir
# copying files to regal w/ new datestamps
rsync -a --no-times --size-only --progress sourcedir /n/regal/labname/myfolder/mysubdir

# tar in flight
# very powerful, but be very careful to get all the
tar -C local_source_dir -cpf - . | tar -C /path/to/destination_dir -xvf -
```

Searching in the filesystem: Find

```
find . -name "*Foo*"              # is inherently recursive and case sensitive
find . -maxdepth 2 -name "*.fa"   # limit recursion depth,
find . -ctime +1                  # I created something 2 days ago. where?
```

# Useful for regal users: files **m**odified **time** greater than 90 days (**a**ccess**time** or **c**reation**time**)

```
find /n/regal/mylab -type f -mtime +89
```

# Find all files named foo and remove them

```
find . -name "foo" -type f -print0 | xargs -0 /bin/rm
```

# Find all files named foo and copy them to another location

```
find . -name "foo" -type f -print0 | xargs -0 cp ‘{}’ --target-directory=targetdir \;
```

Counting & Sorting

# how many files in this directory

```
ls -1 | wc -l
```

# sort interact jobs by jobID

```
squeue -p interact | sort -k1 -n        # -k is column; -n is numeric sort
```

# sort usernames

```
squeue -p interact --format=%u | sort   # get squeue to give us custom output
```

# sort username & grab count of unique entries

```
squeue -p interact --format=%u | sort | uniq -c
```

Searching in files: Grep

# getting count of sequences in FASTA file

```
grep -c “^>” sequences.fasta
```

# searching directories recursively for file contents

```
grep -ir "string-to-search" /path/to/search
```

# negating … give me everything *but* the sequence defline

```
grep -v “^>” sequences.fasta
```

# make one large sequence entry for all the sequences in my FASTA file

```
echo “>myLargeSeq” > myLargeSeq.fasta && grep -v “^>” sequences.fasta \
  >> myLargeSeq.fasta # really need the -h so filenames aren’t in grep output
```

Getting data out of files: Awk

# great SwissArmy knife for text processing

# make one large sequence entry for all the sequences in my FASTA file

```
squeue -p interact | sort -k1 -n | awk ‘{print $1”\t”$3}’ > jobIDs_programs.txt
```

# make one large sequence entry for all the sequences in my FASTA file

```
squeue -p interact | sort -k1 -n | awk ‘{print $1”\t”$4}’ > jobIDs_usernames.txt
awk '{s += $3} END { print s}' somefile.txt    #  Adds data from column 3 in file
```

Join

# combine two files

```
join -a1 jobIDs_programs.txt jobID_usernames.txt | head
```

Putting it all together

```
# Delete subdirectories (work on names with spaces?)
ls -l | grep ^d | awk '{print $9}' | xargs  rm -rf
```

```
# In-place file text replacement with sed, using alternative to delim  '/' so paths are easier to manipulate:
sed -i -e 's?/jab/?/akitzmiller/?g' interestingpaths.txt
```

```
# Sum of cpus from sacct using awk
sacct -u akitzmiller --starttime 2015-02-01 --endtime 2015-03-01 \
  --format=JobID,ncpus -n | awk '{sum += $2} END {print sum}'
```

```
# Delete subdirectories
   ls -l | grep ^d | awk '{print $9}' | xargs rm -rf
```

```
# Find files in current directory containing text using find and xargs:
  find . -name "*" -print0 | xargs -0 grep -l 'akitzmiller'
```

Handling large #s of files

```
# renaming
for file in *.fastq; do
  newname=${file//bad/good}
  mv $file $newname
done
```

```
# run cmd on bunches of files, each w/ its own output
for file in *.fastq; do
  output=${file%.fastq}.out
  echo $file
  cmd $file > $output
done
```

```
# submit a bunch of files w/ timestamp log files
# today=$(date +%Y-%m-%d)   # YYYY-mm-dd
# NB! please sleep one second between submissions to make
# Also ensure that each job runs about 5 min or more
now=(date +%Y-%m-%d_%H:%M:%S) for YYYY-mm-dd_hour:min:sec
for file in *.fastq; do
  base=${file%.fastq}
  echo $file
  sbatch -o ${base}_$now.stdout -e ${base}_$now.stderr \
    --wrap=”wc -l $file”
  # or sbatch -o ${base}_$now.stdout -e ${base}_$now.stderr mySLURM.sbatch $file
  sleep 1
done
```

Higher-level Swiss Army knives

# screen

```
screen -S <session_name> # to start a session
                         # Ctrl-a d to disconnect from a session
screen -r <session_name> # to rejoin an existing session.
                         # kill a session you are in, type Ctrl-a D D.
screen -ls # view a list of existing sessions
```

# fasta_tool from MAKER package
# great for manipulating sequences!!

```
source new-modules.sh
module load legacy
module load centos6/maker-2.28
fasta_tool
```

# Scriptome
# Perl one-liners for doing complex work on the command line
# http://archive.sysbio.harvard.edu/csb/resources/computational/scriptome/