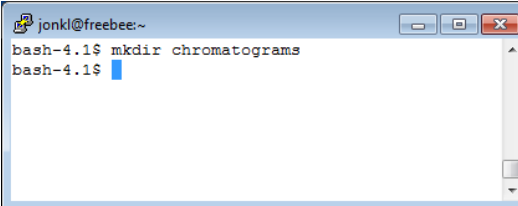# Practical Unix exercise – MBV-INFX410

We will in this exercise work with a practical task that, it turns out, can easily be solved by using basic Unix.
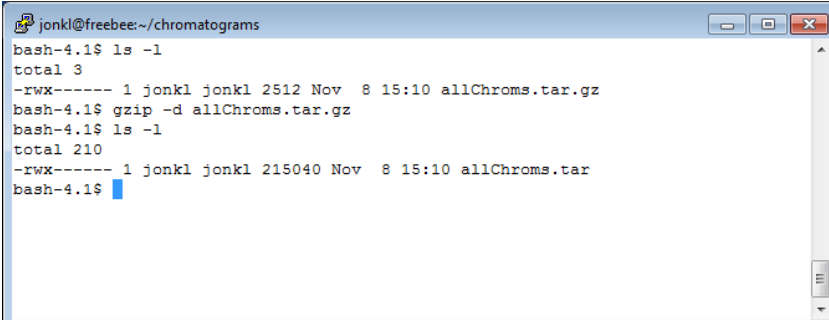
Let us pretend that an engineer in your group has spent several weeks growing and Sanger sequencing several hundred clones. All the successful sequencing data has been "packed" in one single file called allChroms.tar.gz.

1. Log onto freebee.abel.uio.no and create a new directory in your home area. Let us call it "chromatograms".

2. Download the file allChroms.tar.gz from the wiki page and put it in the new directory. How you do this will depend on your laptop. When you have done this, make sure you understand what you did. We will do similar operations many more times during the course (and for the exam…).

3. The file has a double ending, ".tar.gz". This indicates that this is a compressed file that has been compressed, or packed to save space, by the gzip software application (hence the ".gz"). It is also a "tar file", also known as a "tarball", which usually means that many files have been packed into a single file. This is often done to make file transfer and/or file storage easier. Use *ls -l allChroms.tar.gz* to see the size of the compressed file.

4. Uncompress the file by running the command *gzip -d allChroms.tar.gz* (*gunzip allChroms.tar.gz* will do exactly the same and is possibly easier to remember). Do *ls -l* to see what you have now. Notice that the uncompressed file is much bigger than the "gzipped" version. *gzip* and other compression applications are very useful to save disk space and speed up file transfer.

5. Now pack out all the files in the tarball archive file by running *tar -xvf allChroms.tar*. Here "-x" tells *tar* to "extract" all files in the archive, "-f" tells *tar* to extract them from the file allChroms.tar (and not, for example, from a tape station), and "-v" tells *tar* to be "verbose" and print to the terminal what it is doing. Of course, you can read more about *gzip* and *tar* by using the *man* command.

6. Now do *ls -l* to find out what you have in your current directory. You will find that you still have the allChroms.tar archive file, but in addition two directories dirFwdPrimer and dirRevPrimer. We want to keep the tarball file but also save disk space. Let us compress the file again with *gzip allChroms.tar*. Did this shrink the file size?

7. Go into the directories dirFwdPrimer and dirRevPrimer and find out what you have there. Count the number of files in the directories by using some Unix commands. Did you manage?

8. One possibility is to do like this:

```
jonkl@freebee:~/chromatograms
bash-4.1$ ls -l
total 211
-rwx------ 1 jonkl jonkl 215040 Nov  8 15:10 allChroms.tar
drwx------ 2 jonkl jonkl    100 Nov  8 14:45 dirFwdPrimer
drwx------ 2 jonkl jonkl    100 Nov  8 14:45 dirRevPrimer
bash-4.1$ gzip allChroms.tar
bash-4.1$ ls -l
total 4
-rwx------ 1 jonkl jonkl 2512 Nov  8 15:10 allChroms.tar.gz
drwx------ 2 jonkl jonkl  100 Nov  8 14:45 dirFwdPrimer
drwx------ 2 jonkl jonkl  100 Nov  8 14:45 dirRevPrimer
bash-4.1$ cd dirFwdPrimer
bash-4.1$ ls
chromExp45-100  chromExp45-117  chromExp45-134  chromExp45-151  chromExp45-168  chromExp45-185
chromExp45-101  chromExp45-118  chromExp45-135  chromExp45-152  chromExp45-169  chromExp45-186
chromExp45-102  chromExp45-119  chromExp45-136  chromExp45-153  chromExp45-170  chromExp45-187
chromExp45-103  chromExp45-120  chromExp45-137  chromExp45-154  chromExp45-171  chromExp45-188
chromExp45-104  chromExp45-121  chromExp45-138  chromExp45-155  chromExp45-172  chromExp45-189
chromExp45-105  chromExp45-122  chromExp45-139  chromExp45-156  chromExp45-173  chromExp45-190
chromExp45-106  chromExp45-123  chromExp45-140  chromExp45-157  chromExp45-174  chromExp45-191
chromExp45-107  chromExp45-124  chromExp45-141  chromExp45-158  chromExp45-175  chromExp45-192
chromExp45-108  chromExp45-125  chromExp45-142  chromExp45-159  chromExp45-176  chromExp45-193
chromExp45-109  chromExp45-126  chromExp45-143  chromExp45-160  chromExp45-177  chromExp45-194
chromExp45-110  chromExp45-127  chromExp45-144  chromExp45-161  chromExp45-178  chromExp45-195
chromExp45-111  chromExp45-128  chromExp45-145  chromExp45-162  chromExp45-179  chromExp45-196
chromExp45-112  chromExp45-129  chromExp45-146  chromExp45-163  chromExp45-180  chromExp45-197
chromExp45-113  chromExp45-130  chromExp45-147  chromExp45-164  chromExp45-181  chromExp45-198
chromExp45-114  chromExp45-131  chromExp45-148  chromExp45-165  chromExp45-182  chromExp45-199
chromExp45-115  chromExp45-132  chromExp45-149  chromExp45-166  chromExp45-183
chromExp45-116  chromExp45-133  chromExp45-150  chromExp45-167  chromExp45-184
bash-4.1$ ls | wc
    100     100    1500
bash-4.1$
```

ls lists all the files and *wc* will count the number of words. There are 100 files in each of the directories dirFwdPrimer and dirRevPrimer. Thus, 200 files in two directories were packed in the tarball. These are not real Sanger sequencing chromatogram files (that would have taken a lot of space), but let us pretend they are.

9. The group engineer has named the clones 100, 101, 102, etc. She has put all the files generated by using forward primers in the directory dirFwdPrimer, and files from reverse primers in dirRevPrimer. Consequently, the files dirFwdPrimer/chromExp45-123 and dirRevPrimer/chromExp45-123 correspond to the same clone (number 123)
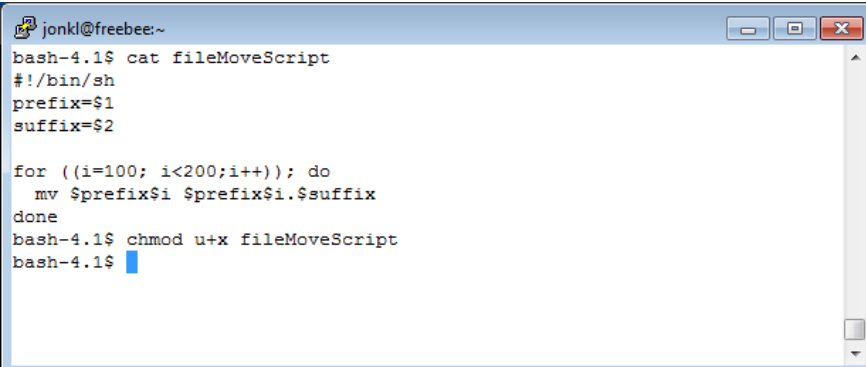
but from forward and reverse primers, respectively. The engineer has now gone to Bali for a 3 week holiday and left the analysis job to you…

Your boss has bought an expensive program that she wants you to use to analyse the data. According to the software manual: "All chromatogram files should be in a single directory. All files should have the format *clonename.suffix*, where *clonename* is a name unique to each clone and *suffix* is "fwd" for forward primers and "rev" for reverse primers."

You clearly have to make a new directory, for example called dirAll. Then you must take the file dirFwdPrimer/chromExp45-100, move it to dirAll, and rename it chromExp45-100.fwd. Next you must take the file dirRevPrimer/chromExp45-100, move it to dirAll, and rename it chromExp45-100.rev. Then do the same for 101, 102, etc…

Needless to say, you are not looking forward to this task. How would you do this if you did not know anything about scripting/programming? Think carefully through this before you proceed.

10. The task at hand is exceptionally boring, because it is the same operations over and over again. If you do this manually it will take a lot of time and most likely you will also do many mistakes. However, this task is perfectly suited for a little computer program or a script. Let's try that! Are you able to write a shell script that can do the job? If not, don't worry. You are not expected to be able to do this, yet. Proceed below to see how it can be done.

11. A little script, here named fileMoveScript, will do the renaming part of the job for you,

```
jonkl@freebee:~
bash-4.1$ cat fileMoveScript
#!/bin/sh
prefix=$1
suffix=$2

for ((i=100; i<200;i++)); do
  mv $prefix$i $prefix$i.$suffix
done
bash-4.1$ chmod u+x fileMoveScript
bash-4.1$
```

Do you understand what the script will do? Make sure you do! Use *nano* and create the script yourself. Type it in and save it to a file called fileMoveScript in the directory ~/chromatograms. Make sure you make the script executable with *chmod* as shown above. Do you understand what *chmod* does? Now use it to rename the forward primer files.

```
jonkl@freebee:~/chromatograms
bash-4.1$ cat fileMoveScript
#!/bin/sh
prefix=$1
suffix=$2

for ((i=100; i<200;i++)); do
  mv $prefix$i $prefix$i.$suffix
done
bash-4.1$ chmod u+x fileMoveScript
bash-4.1$ cd dirFwdPrimer
bash-4.1$ ../fileMoveScript chromExp45- fwd
bash-4.1$ ls
chromExp45-100.fwd  chromExp45-125.fwd  chromExp45-150.fwd  chromExp45-175.fwd
chromExp45-101.fwd  chromExp45-126.fwd  chromExp45-151.fwd  chromExp45-176.fwd
chromExp45-102.fwd  chromExp45-127.fwd  chromExp45-152.fwd  chromExp45-177.fwd
chromExp45-103.fwd  chromExp45-128.fwd  chromExp45-153.fwd  chromExp45-178.fwd
chromExp45-104.fwd  chromExp45-129.fwd  chromExp45-154.fwd  chromExp45-179.fwd
chromExp45-105.fwd  chromExp45-130.fwd  chromExp45-155.fwd  chromExp45-180.fwd
chromExp45-106.fwd  chromExp45-131.fwd  chromExp45-156.fwd  chromExp45-181.fwd
chromExp45-107.fwd  chromExp45-132.fwd  chromExp45-157.fwd  chromExp45-182.fwd
chromExp45-108.fwd  chromExp45-133.fwd  chromExp45-158.fwd  chromExp45-183.fwd
chromExp45-109.fwd  chromExp45-134.fwd  chromExp45-159.fwd  chromExp45-184.fwd
chromExp45-110.fwd  chromExp45-135.fwd  chromExp45-160.fwd  chromExp45-185.fwd
chromExp45-111.fwd  chromExp45-136.fwd  chromExp45-161.fwd  chromExp45-186.fwd
chromExp45-112.fwd  chromExp45-137.fwd  chromExp45-162.fwd  chromExp45-187.fwd
chromExp45-113.fwd  chromExp45-138.fwd  chromExp45-163.fwd  chromExp45-188.fwd
chromExp45-114.fwd  chromExp45-139.fwd  chromExp45-164.fwd  chromExp45-189.fwd
chromExp45-115.fwd  chromExp45-140.fwd  chromExp45-165.fwd  chromExp45-190.fwd
chromExp45-116.fwd  chromExp45-141.fwd  chromExp45-166.fwd  chromExp45-191.fwd
chromExp45-117.fwd  chromExp45-142.fwd  chromExp45-167.fwd  chromExp45-192.fwd
chromExp45-118.fwd  chromExp45-143.fwd  chromExp45-168.fwd  chromExp45-193.fwd
chromExp45-119.fwd  chromExp45-144.fwd  chromExp45-169.fwd  chromExp45-194.fwd
chromExp45-120.fwd  chromExp45-145.fwd  chromExp45-170.fwd  chromExp45-195.fwd
chromExp45-121.fwd  chromExp45-146.fwd  chromExp45-171.fwd  chromExp45-196.fwd
chromExp45-122.fwd  chromExp45-147.fwd  chromExp45-172.fwd  chromExp45-197.fwd
chromExp45-123.fwd  chromExp45-148.fwd  chromExp45-173.fwd  chromExp45-198.fwd
chromExp45-124.fwd  chromExp45-149.fwd  chromExp45-174.fwd  chromExp45-199.fwd
bash-4.1$
```

12. Also rename all the reverse primer files and give them a ".rev" suffix. Create the dirAll directory and move all the renamed files into this directory.

```
jonkl@freebee:~/chromatograms
chromExp45-105.fwd   chromExp45-130.fwd   chromExp45-155.fwd   chromExp45-180.fwd
chromExp45-106.fwd   chromExp45-131.fwd   chromExp45-156.fwd   chromExp45-181.fwd
chromExp45-107.fwd   chromExp45-132.fwd   chromExp45-157.fwd   chromExp45-182.fwd
chromExp45-108.fwd   chromExp45-133.fwd   chromExp45-158.fwd   chromExp45-183.fwd
chromExp45-109.fwd   chromExp45-134.fwd   chromExp45-159.fwd   chromExp45-184.fwd
chromExp45-110.fwd   chromExp45-135.fwd   chromExp45-160.fwd   chromExp45-185.fwd
chromExp45-111.fwd   chromExp45-136.fwd   chromExp45-161.fwd   chromExp45-186.fwd
chromExp45-112.fwd   chromExp45-137.fwd   chromExp45-162.fwd   chromExp45-187.fwd
chromExp45-113.fwd   chromExp45-138.fwd   chromExp45-163.fwd   chromExp45-188.fwd
chromExp45-114.fwd   chromExp45-139.fwd   chromExp45-164.fwd   chromExp45-189.fwd
chromExp45-115.fwd   chromExp45-140.fwd   chromExp45-165.fwd   chromExp45-190.fwd
chromExp45-116.fwd   chromExp45-141.fwd   chromExp45-166.fwd   chromExp45-191.fwd
chromExp45-117.fwd   chromExp45-142.fwd   chromExp45-167.fwd   chromExp45-192.fwd
chromExp45-118.fwd   chromExp45-143.fwd   chromExp45-168.fwd   chromExp45-193.fwd
chromExp45-119.fwd   chromExp45-144.fwd   chromExp45-169.fwd   chromExp45-194.fwd
chromExp45-120.fwd   chromExp45-145.fwd   chromExp45-170.fwd   chromExp45-195.fwd
chromExp45-121.fwd   chromExp45-146.fwd   chromExp45-171.fwd   chromExp45-196.fwd
chromExp45-122.fwd   chromExp45-147.fwd   chromExp45-172.fwd   chromExp45-197.fwd
chromExp45-123.fwd   chromExp45-148.fwd   chromExp45-173.fwd   chromExp45-198.fwd
chromExp45-124.fwd   chromExp45-149.fwd   chromExp45-174.fwd   chromExp45-199.fwd
bash-4.1$ cd ../dirRevPrimer
bash-4.1$ ../fileMoveScript chromExp45- rev
bash-4.1$ cd ..
bash-4.1$ mkdir dirAll
bash-4.1$ cd dirAll
bash-4.1$ mv ../dirFwdPrimer/chromExp45-1* .
bash-4.1$ mv ../dirRevPrimer/chromExp45-1* .
bash-4.1$ ls | head -10
chromExp45-100.fwd
chromExp45-100.rev
chromExp45-101.fwd
chromExp45-101.rev
chromExp45-102.fwd
chromExp45-102.rev
chromExp45-103.fwd
chromExp45-103.rev
chromExp45-104.fwd
chromExp45-104.rev
bash-4.1$ ls | wc
    200    200    3800
bash-4.1$
```

There you have it! All 200 files renamed according to the correct naming convention and all in the same directory. You can now do the analysis for your boss. She will be very impressed with your work…!

13. Actually, put the directory dirAll with all the files in it in a tarball and gzip it. Call it fixedChroms.tar. Then gzip it and copy it back to a local disk on your laptop. Make sure you are able to do this! You can e-mail it to the engineer in Bali if you feel like it…

```
jonkl@freebee:~/chromatograms
bash-4.1$ ls
allChroms.tar.gz   dirAll   dirFwdPrimer   dirRevPrimer   fileMoveScript
bash-4.1$ tar -cf fixedChroms.tar dirAll
bash-4.1$ gzip fixedChroms.tar
bash-4.1$ ls
allChroms.tar.gz   dirFwdPrimer   fileMoveScript
dirAll             dirRevPrimer   fixedChroms.tar.gz
bash-4.1$
```

Here we did *tar* without the "-v" = "verbose" option and with "-c" that will "create" a new tarball.

Some comments to this exercise:
- This was an example of a task that is very time consuming, exceptionally boring and error prone if you do it "manually", for example by clicking, dragging and typing in a GUI interface like Microsoft Windows. However, the task is very simple, quick, and potentially error free if you know a little bit of programming.
- If you cannot do the programming yourself, you will now at least know that this kind of task *should* be solved programmatically, and you can ask for help from someone that knows how to do this.
- For slightly more complicated tasks, it is most likely better to use a scripting language such as Perl or Python. You will learn a bit of Python tomorrow. If you can use Python, creating scripts in Unix may not be necessary.
- A couple of years ago I got a very similar task to the one presented here from a colleague. Her engineer had used 9(!) different naming conventions for more than 20,000 chromatogram files generated over many months. My colleague was very grateful when I could rename all her files (and make sure I had not done any mistakes) in just over a day.
- After this exercise you *must* be able to:
  - In a Unix shell, create directories, move around between the directories and copy and move files. Use *ls* with options to see which files and directories you have in a directory.
  - Use *cat* and *more* to view the contents of a file, and use *nano* to write text into a file or create a new one.
  - Delete files
  - Transfer files from your laptop to freebee.abel.uio.no and back
  - Download a compressed tarball file from somewhere and store it on your laptop and on freebee.abel.uio.no
  - Uncompress and extract the files from the tarball
  - Make a tarball archive file with tar and compress with *gzip*
  - Make programs executable and files visible or hidden from other users with *chmod*
  - Know how to use *grep*, "pipe" (|), and how to redirect output with ">"