Session 1 Exercises

1. **Using FASTX to get statistical report of Illumina sequencing data**
2. **Using Roche gsMapper to align 454 reads to *E.coli* genome.**

   Manual for FASTX can be found at http://hannonlab.cshl.edu/fastx_toolkit/

   Manual for the Roche software is located in the session's data file directory (See step 4).

Step 1. Log into the CAC server.

Windows user

For step-by-step instructions, go to
http://www.cac.cornell.edu/wiki/index.php?title=V4_Linux_Cluster#From_Windows

- launch the software Putty. (You can get Putty from http://putty.cs.utah.edu/)
- Under Host Name, Type: linuxlogin.cac.cornell.edu. Make sure you check the "Enable X11 forwarding" checkbox under SSH->X11. You can adjust the color and font under Window-> Colors and Window->Appearance. For example, I like my default Foreground and Bold Foreground set to 0,0,0, and Background and Bold Background set to 255,255,255, so that I could get white background and black fonts.
- You might want to save the session, so that you do not have to type the host name and change settings next time you log in. In order to do that, after you adjust the setting and check the X11 box, click "Session" in the menu, type any name under "Saved Sessions", then click "Save". Next time you only need to double click the session name to login. If this is your personal computer, you can also save the login name under Connection->Data->Auto login user name.
- Enter User Name and password when prompted. The first time you connect to a machine, there will be a warning message to ask you whether you can trust the server, just click "Yes".

Mac user:

For step-by-step instructions, go to
http://www.cac.cornell.edu/wiki/index.php?title=V4_Linux_Cluster#From_a_Mac

- open the terminal window
- type: ssh -X YourUserName@linuxlogin.cac.cornell.edu (replace YourUserName with you CAC account user name, normally your netID)
- Enter password when prompted

Step 2. Create a session1 directory under your home directory on the CAC server. Copy all data files of the project to your session1 directory.

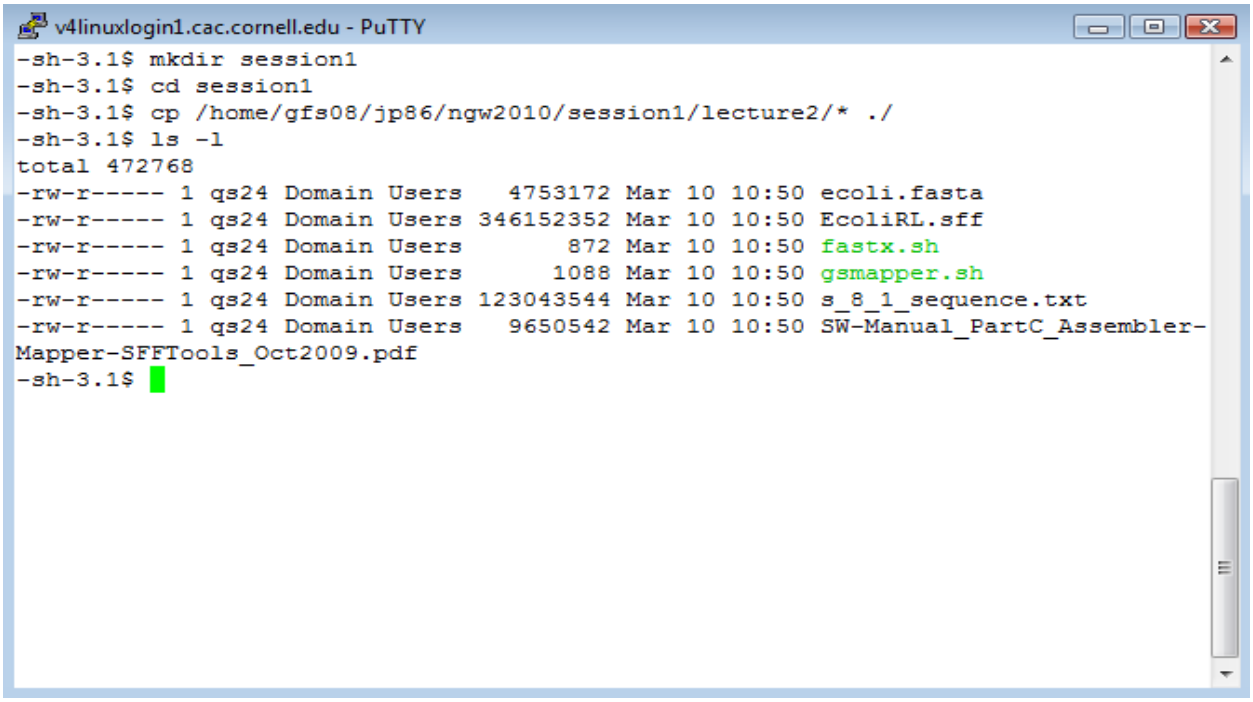
From the terminal window, type the following commands. Each line should be followed by "Enter" key.

(Note: You can copy-paste selected text instead of typing. Putty users can do copy-paste by right-click mouse in the Putty window. First you copy a command from this instruction file by highlight then press Ctrl-C. Then paste in the Putty window by right click the mouse. Mac users can do copy-paste the regular way using Apple-C or Apple-V.)

mkdir session1

cd session1

cp /home/gfs08/jp86/ngw2010/session1/lecture2/* ./



After you finish these steps, make sure you see the following 6 files by typing "ls -l" followed by Enter key. "ls -l" command would give you the size of the file (in bytes), and last time it was modified.

ecoli.fasta (fasta file of the E.coli genome sequence)

EcoliRL.sff (454 sequencing data file of the E.coli K12)

s_8_1_sequence.txt (Illumina sequencing data of a bac clone PhiX)

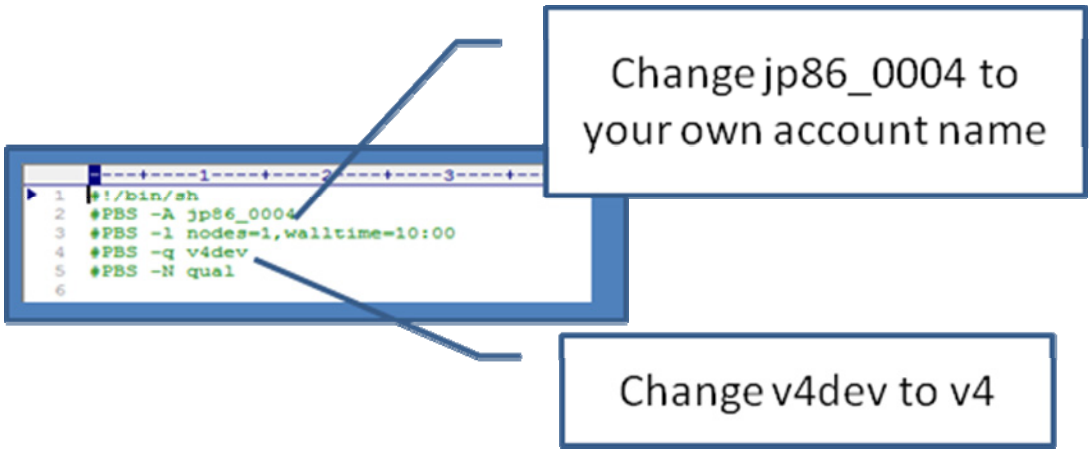
fastx.sh (a script for submitting a job to run FASTX)

gsmapper.sh (a script for submitting a job to run command line version of gsMapper)

SW-Manual_PartC_Assembler-Mapper-SFFTools_Oct2009.pdf (manual of 454 software)

Step 3. Submit a job to get the quality report of a Illumina data file.

- Inspect the fastq file. Type the command: "more s_8_1_sequence.txt" and press Enter. Press "space bar" to move to the next page; Press "q" to exit the "more" command".
- Modify the first five lines in the fastx.sh file. If you do not know how to do it, follow the instructions for Windows or Mac users below.



Modify the line "#PBS -A jp86_0004". Change "jp86_0004" to your own project name. The project name is different from your login NetID. When you open the CAC project manager, you would see your project name. If you do not know what it is, please contact CAC (help@cac.cornell.edu). Change the cluster name from v4dev to v4. You can also modify the walltime (#PBS -l nodes=1,walltime=10:00) or job name (#PBS -N qual), but it is not necessary for this project.

(Default walltime is 10:00, which means maximum allowed running time 10 minutes. If you want to change to 1 hour, change it to "walltime=1:00:00. Default name is "qual". You can change it to any name you want).

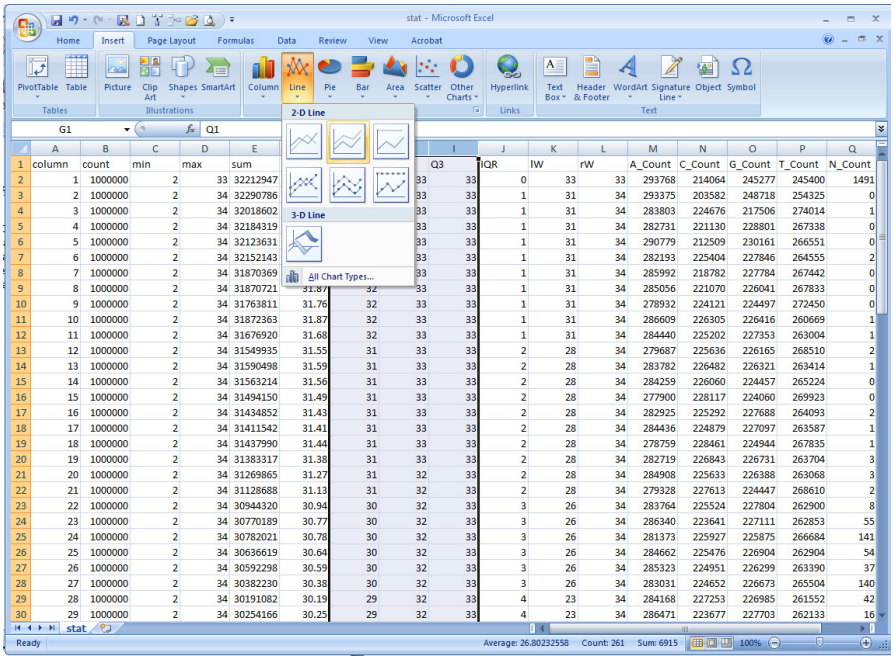
Windows users can open the fastx.sh file on the CAC server by following the instructions at : http://www.cac.cornell.edu/wiki/index.php?title=Using_EditPlus

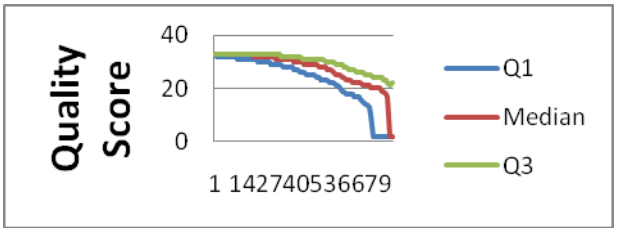
The file you need to modify is "fastx.sh" under "session1" directory. After you finish modification, click to save the file. You will be prompted "Do you want to change backup option", just click "No".

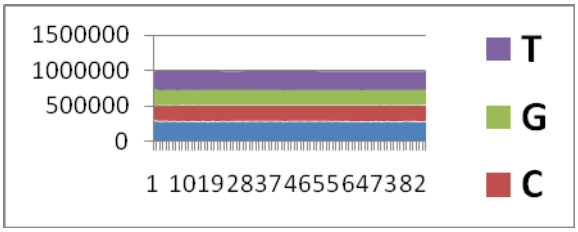
Mac users can open the fastx.sh file using TextWrangler. Instructions of using TextWrangler can be found at http://www.cac.cornell.edu/wiki/index.php?title=TextWrangler_Step_by_Step . After you finish the modification, click "File" -> "Save to FTP/SFTP server", click "Save".

- Submit the job by typing this command line followed by Enter.
  nsub fastx.sh
- The job would take a few minutes to finish. You can monitor the progress by "qstat" command. After the job is finished, you will see two new files created under the session1 directory: stat.xls and stat.png. If you do not see these files, your job has failed. The error message is in the xxxxxxxscheduler.v4linux.ER file. You can use "more xxxxxxscheduler.v4linux.ER" command to see the error messages. Move these two files back to your local computer. Windows users can use WinSCP (http://www.cac.cornell.edu/wiki/index.php?title=WinSCP_Step-by-Step) . Mac users can use FETCH to move files (http://www.cac.cornell.edu/wiki/index.php?title=Fetch_Step-by-Step).
- After you move the files to your own computer, open the stat.xls file in Excel. If Excel complains that this is not the correct, just click "Yes". First plot the Q1, median and Q3 columns using line plot in Excel, then plot the A C G T count using area plot in Excel. The columns are:
  - column = cycle number (1 to 80 for a 80-cycles read solexa file)
  - count = number of bases found in this column.
  - min = Lowest quality score value found in this column.
  - max = Highest quality score value found in this column.
  - sum = Sum of quality score values for this column.
  - mean = Mean quality score value for this column.
  - Q1 = 1st quartile quality score.
  - med = Median quality score.
  - Q3 = 3rd quartile quality score.
  - IQR = Inter-Quartile range (Q3-Q1).
  - lW = 'Left-Whisker' value (for boxplotting).
  - rW = 'Right-Whisker' value (for boxplotting).
  - A_Count = Count of 'A' nucleotides found in this column.
  - C_Count = Count of 'C' nucleotides found in this column.
  - G_Count = Count of 'G' nucleotides found in this column.
  - T_Count = Count of 'T' nucleotides found in this column.
  - N_Count = Count of 'N' nucleotides found in this column.
  - max-count = max. number of bases (in all cycles)

Plotting steps: Open stat.xls in Excel ; Select columns GHI; Click Insert->Line->2-D Line to get the quality score plot. Select columns MNOPQ; Click Insert->Area->2-D Area to get the Nucleotide distribution plot.







Step 4. Submit a job to align 454 reads to the *E.coli* genome.

- Modify the first five lines in the gsmapper.sh script as described in step 3. Change the line "#PBS -A jp86_0004" to your own account name.
- Submit the job: nsub gsmapper.sh. It will take a few minutes for the job to finish. A new directory "ecolimapping" will be created. Use the "ls -l" command to check the new directory.
- Inspect the alignment results using GSMapper. Manual of the GSMapper is in the file SW-Manual_PartC_Assembler-Mapper-SFFTools_Oct2009.pdf in your session1 directory, you can move it to your desktop computer and open it. This step is not always easy. Even some experienced Linux users would have problems to set up X-windows. If you cannot get this to work, just skip this step 4. When you need to analyze your own real data, please contact CBSU and we will figure out a solution for you.

  i) Windows Users: make sure you have xming installed on your computer. Launch the Xming by click Start->All Programs->Xming->Xming. If you did not check the "Enable X11 forwarding" box when launching Putty, you need to close putty, and restart Putty with the "Enable X11 forwarding" box checked. Mac users can skip this step.

  ii) Type the command: /opt/nextgen/bin/gsMapper. You should see the gsMapper software open. If not, your X-windows software are not properly set.

iii) Click "Open a Mapper Project", navigate to the "session1" directory, and click open the "ecolimapping" project.

iv) Click the "Alignment results" tab. Inspect the alignment results. Looking for homopolymer regions.

v) Click the "Flowgrams" tab and inspect the flowgram. There will be three panels, the top panel shows the reference, the middle panel shows the read, the bottom one shows the difference.

vi) Click the "Variants" tab and inspect the variants results.

- Close the gsMapper window after you are done.
- Move the 454MappingQC.xls (located in the ecolimapping/mapping directory) back to your local computer. Open it with Excel and inspect the mapping results.

Step 5. Using other functions in FASTX package(optional)

FASTX package has many functions. For examples: fastx_clipper can be used for removing the 3' adaptor sequences, this is useful for small RNA profiling project; fastx_collapser can be used to create a unique set of sequences; fastx_trimmer can be used to trim the sequencing reads shorter; fastx_barcode_splitter.pl is useful for decoding multiplexed samples. A list of available FASTX tools can be found at http://hannonlab.cshl.edu/fastx_toolkit/commandline.html .

In the fastx.sh script, there are two lines specifying what functions to call. You can modify these two lines and call other functions. Then submit the job by "nsub fastx.sh", and you will get the result files.

Change the following two lines:

```
fastx_quality_stats -i s_8_1_sequence.txt -o stat.xls
fastq_quality_boxplot_graph.sh -i stat.xls -t Lane8_Report -o stat.png
```

To one line:

```
fastx_trimmer  -l 70 -i s_8_1_sequence.txt -o trimmed_sequence.txt
```

Then change the following two lines:

```
cp stat.xls $HOME/session1/
cp stat.png $HOME/session1/
```

To one line:

```
cp trimmed_sequence.txt $HOME/session1/
```

Appendix:

A few tips of using Linux system:

1. When you login, you are in your home directory. After a few "cd" commands changing your directory, you should always be aware of what is the directory you are in. If you do not know, use the command "pwd" which would tell you what is your current directory. If you want to process a data file, both the data file and the software need either to be in your current directory, or you need to refer to the software and data file names with full directory path (e.g .

use "/opt/nextgene/bin/gsMapper" instead of "gsMapper"). Use "cd" to change directory. There are a few short cuts for "cd". If you want to go back to your home directory, use "cd" followed directly by "Enter" key. If you want to move to an upper directory, do "cd ..".

2. Copy-paste in Linux terminal is different from Windows or Mac (The exception is when you use Mac built-in terminal to access a Linux server, you can do copy-paste the regular way). In Linux terminal, copy is done by selecting the text followed by right click, then right click again to paste. If you want to copy-paste text between a Windows software (e.g. text from files opened in Microsoft Word) and the Putty terminal window, you should use "Ctrl C" or "Ctrl V" to copy-paste in windows software, "right click" to copy-paste in Linux terminal.
3. If you want to repeat a command that you have executed before, you can keep pressing "upper arrow" to move to that command, then press "Enter".
4. In Linux, "Tab" key can be used to auto-finish a command. For example, you want to see the content of the file "s_1_sequence.txt" by "more" command. Instead of typing "more s_1_sequence.txt", you can type "more s_1" then press "Tab", the terminal would automatically finish the command for you.
5. You can open as many session windows as you want, so that you can do several things simultaneously on the same server.
6. When you use the CAC system, it is VERY IMPORTANT that you do not run any computational intensive software directly from the login node. You should always put the commands that you want to run into a script, and do "nsub yourCommandScript" to run the job. If you are testing a software, you can request a v4dev node interactively and test the software.