**Downloading A FASTQ File**

SRA (<https://www.ncbi.nlm.nih.gov/sra/>) is a great repository of SRA files which could then be converted to FASTQ files. Here is the process to download and extract a FASTQ file on an Ubuntu terminal:

* Download the SRA toolkit
  + Command: sudo apt-get install sra-toolkit
* Download an SRA file
  + Command: prefetch SRR######
  + Replace ###### with the SRR number listed on the SRA website for the file you want to download
* Convert the .sra file to a .fastq file
  + Command: fastq-dump [name of .sra file downloaded in previous step]

**Manipulation of FASTQ files**

Once a fastq file has been downloaded using the above steps, the following commands could be used on the Ubuntu terminal to manipulate it:

* Head [name of .fastq file extracted]
  + Fastq files could be extremely long, therefore this command could be used to view the first 10 lines of the file
* Head -n k [name of .fastq file extracted]
  + Will show the first k number of lines where k is an integer
* Tail [name of .fastq file extracted]
  + Fastq files could be extremely long, therefore this command could be used to view the last 10 lines of the file
* Tail -n k [name of .fastq file extracted]
  + Will show the last k number of lines where k is an integer
* Wc -l [name of .fastq file extracted]
  + Will return the number of lines in the fastq file. Divide this number by 4 to get the number of sequences in the fastq file
* More [name of .fastq file extracted]
  + Will let you scroll through the entire file by line using the enter key or by page using the space key. Not suggested as fastq files could be extremely long and will therefore take a long time to scroll through
* Less [name of .fastq file extracted]
  + Works the same as the more command but is more efficient. Type q to exit out of the file
* Cat [name of .fastq file extracted] | grep [search\_string]
  + Will print all occurrences of the search\_string within the .fastq file. The | symbol means the fastq file is being piped to the grep command for analysis.
  + Options that could be added after grep
    - -v
      * Shows all the lines that do not match the searched string
    - -c
      * Displays only the count of matching lines
    - -n
      * Shows the matching line and its number
    - -i
      * Match both (upper and lower) case
* Sed ‘s/first\_string/second\_string/’ [name of .fastq file extracted]
  + Will replace all occurrences of the first\_string with the second\_string in the fastq file. The file is not updated and the output can only be seen in the terminal
* Sed ‘s/first\_string/second\_string/’ [name of .fastq file extracted] > [name of .fastq file]
  + Will replace all occurrences of the first\_string with the second\_string in the fastq file. The new file is then stored in the new file name after >

**Troubles**

While the SRA toolkit could be used with any operating system, I found that it works best with Linux, however, I did not have a Linux computer. Therefore, I created an Ubuntu Linux virtual machine using VirtualBox to use the SRA toolkit.

**Next Steps**

Next, I need to figure out how to transform the fastq file into a full sequence of the entire sample by putting all the clusters together.