File Format Tutorial

For this tutorial we will work in the folder called **fileformat_tutorial**

```
$ cd
$ mkdir fileformat_tutorial
$ cp ~/Course_Data/{All_HIV_Ref.fas,sample.bam,test.fastq} ~/fileformat_tutorial
$ cd fileformat_tutorial
$ ls -lh
```

Opening and Editing files

Opening a FASTA file

```
$ nano All_HIV_Ref.fas
```

Edit the sequences in any way you want. Ctrl + X will exit nano. Save the file when you exit.

Q: How many sequences are there in the FASTA file?

```
$ grep -c '>' All_HIV_Ref.fas
```

Opening a FASTQ file

```
$ nano test.fastq
```

Or if we just want to look at the first few reads of the file

```
$ head test.fastq
```

Q: How many reads are there in the FASTQ file

We know that the FASTQ header starts with '@'. Can we use this information to find how many reads are there in a file.

```
$ grep -c '@' test.fastq
```

If not what is the other way to count?

```
$ cat test.fastq | echo $((`wc -l`/4))
```

Opening a SAM/BAM file

Viewing header information of a SAM file

```
$ samtools view -H sample.bam
```

Q: What is the reference sequence used for this alignment?

Viewing aligned reads in a SAM file

```
$ samtools view sample.bam | head
```