Running command-line programs, shell scripting

Exercise:
Use shell commands to calculate the number of times each primer aligns perfectly to chromsome 22.

```
$ awk '{if ($3==100 && $4==23) print}' primerblast.txt | cut -f1 |
sort | uniq -c
```

(awk knows that 100.000 is the same as 100)

Exercise:
Write an awk script to print only the minus strand alignments from the original primerblast.txt file. Pipe this to wc to count the alignments.

```
$ awk '{if ($9 > $10) print}' primerblast.txt | wc -l
     134
```

Exercise
Look through the bowtie2 options to figure out how to align an unpaired fasta file to a reference, and produce a sam file named rawseqs.sam

```
$ bowtie2 -f -x hg38_chr22 -U primers.fa -S rawseqs.sam
15 reads; of these:
  15 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    5 (33.33%) aligned exactly 1 time
    10 (66.67%) aligned >1 times
100.00% overall alignment rate
```

Exercise
Find two ways in UNIX to print the coordinate of each alignment in rawseqs.sam to chromosome 22.

We only aligned to chromosome 22, so we can just print all coordinates (otherwise we'd put a condition on this, or use grep to create a file with only chr22 alignments to begin with):

```
$ cut -f4 rawseqs.sam
$ awk '{print $4}' rawseqs.sam
```

Exercise (python):
Add a conditional statement so that the above for loop works.

```
myfile = file("rawseqs.sam", “r")
for line in myfile:
    fieldlist = line.split(“\t")
    if (len(fieldlist) > 3):
        pos = fieldlist[3]
        print pos
```

Exercise:
Modify your program so that it prints the chromosome, position, and length of each aligned sequence.

```
myfile = file("rawseqs.sam", “r")
for line in myfile:
    fieldlist = line.split(“\t")
    if (len(fieldlist) > 9):
        print fieldlist[2], fieldlist[3], len(fieldlist[9])
```

Exercise:
Modify your program so that it prints the read name and alignment position of every read that aligns between chr22 42141116 and 42150170 (inclusive), but only if the alignment is 51 characters long.

```
myfile = file("rawseqs.sam", “r")
for line in myfile:
    fieldlist = line.split(“\t”)
    start = int(fieldlist[3])
    if (len(fieldlist) > 9 and len(fieldlist[9])==51 and start <=
42150170 and start >= 42141116):
        print fieldlist[2], fieldlist[3], len(fieldlist[9])
```

Exercise:
Do any sequences occur more than once in this file?
The sequence of the alignment is the 10th column in a sam file.

```
myfile = file("rawseqs.sam", "r")
seqs = {}
for line in myfile:
    linelist = line.split("\t")
    if (len(linelist) > 9):
        if linelist[9] in seqs.keys():
            seqs[linelist[9]] += 1
        else:
            seqs[linelist[9]] = 1

seqs.items()
```