Mary S. Silva
Lecture Assignment 1 (Answers are in bold)

1. First 12 lines (resulting in 3 sequence reads from the FASTQ file):
   @HISEQ2500:589:CC04PACXX:7:1101:1235:1968/1
   NAGAGAGTCAGCGAAGGGAGATAGGGGTGTGGCCCTTTTATAGGATTTGGG
   +
   #1BDFFFDFHHHHGIJJJHIGIJHJJJDGGIJIJJJIIJIJHEHCHGHIIJ
   @HISEQ2500:589:CC04PACXX:7:1101:1497:1919/1
   NCTCCAGAACTGTGAGAAGATAAGTGTTTGCCTTTTTTTTTTTTTTTTTTT
   +
   #1=DFFFFHHHHFHIJIJJJIJJJHHHIIJIJJJJIJJJJJJJJHDDDDDDD
   @HISEQ2500:589:CC04PACXX:7:1101:1300:1975/1
   NGGAAAAGGAAATATCTTCACGTAAAAACTAGACAGAAGCACTCTGAGAAA
   +
2. *Run the program script and save the output file.* By setting row.names = FALSE, we don't get the target names.

   a.

transcript	allreads
24.0305779	3.287700894
6.403991858	3.287700894
15.46780449	3.287700894
NA	3.287700894
161.0306849	3.287700894
NA	3.287700894

   b. The "for" loop calculates the coefficient of variation (CV) for total reads every time cycles through. This is redundant. Use an "if" programming routine to have it added to the file only the first time the "for" loop is run.

   Code:

```
for(b in 1:length(tester)){
  restest<-normfunct(countdata,tester[b])
  transcriptcv<-sd(restest$transcript)/mean(restest$transcript)*100
  allcv<-sd(restest$allreads)/mean(restest$allreads)*100
  outfile[b,1]<-transcriptcv;print(transcriptcv)
  if (b==1){
   outfile[b,2]<-allcv;print(allcv)

  }
}
```

   Output:

transcript	allreads
24.03058	3.287701
6.403992	NA
15.4678	NA
NA	NA
161.0307	NA
11.40699	NA
181.0101	NA

   ←all reads calculated for first entry only

3. A.
    a. *Based on the coefficients of variation, which drug target is affected the most?*
    Based on the coefficients of variation, PARP1 shows a CV of 6.4%, which is the lowest extent of variability and can be interpreted as being stably expressed across experiments. CDK7 has the highest extent of variability with a CV of 181% meaning that it is less stably expressed across experiments.

    b. *Is there something about the counts that exaggerates the statistical effect?*
    For CA9 and CDK7, the mean is smaller than the standard deviation. This is the experiments show no expression in most of the experiments/transcripts which exaggerates the statistical effects.

B. *CA9 also has a very high CV in these samples, why would it be clinically instructive to look for this gene in a biopsy sample and what does the large CV tell you about what is happening to the cells in the experiment?*

a. CA9 is overexpressed in many types of cancer including lung cancer where it promotes tumor growth. The high CV tells us that the extent of variability in our expression data is high and less stably expressed across experiments. I think it would be clinically intrusive to look for this gene in a biopsy sample because it would have to be from the tumor cells directly.