Course: COMPSCI 260
Name: Divya Koyyalagunta
NetID: dk160
Problem: 1
Problem Set: 3
Due: Fri 30 Sep 2016, 5pm
Using free extension (yes/no): no

Statement of collaboration and resources used (put None if you worked entirely without collaboration or resources; otherwise cite carefully): None
My solutions and comments for this problem are below.

PROBLEM 1

a) The coverage C can be thought of as the proportion of G that is covered by R reads each of length L. So for example, if we have 5 reads of length 5 for a genome of length 10, each nucleotide is expected to be in a read 2.5 times.

$$C = \frac{R \times L}{G}$$

Intuitively, this makes sense, since:

- as G increases, the coverage decreases
- as R and/or L increase, the coverage increases

b) To find the probability that a specific location will not be found in R reads, we first calculate the probability of *finding* a specific location in a **single** read of length L:

$$p_{find} = \frac{L}{G}$$

Now we can calculate the probability of *not* finding a specific location in a **single** read of length L:

$$p_{nfind\_single} = 1 - \frac{L}{G}$$

From the answer in part a, this probability can be rewritten in terms of only R and C:

$$G = \frac{R \times L}{C}$$

Plugging in G with the value above:

$$p_{nfind\_single} = 1 - \frac{L}{\frac{RL}{C}} = 1 - \frac{LC}{RL} = 1 - \frac{C}{R}$$

Now we must think about what happens as our number of reads increases significantly. Using the following formula:

$$\log_{x\to\infty}\left(1-\frac{a}{x}\right)^{x}=e^{-a}$$

where $x = R$, we can plug in the probability of not finding a given nucleotide in a single read (p_{nfind_single}), and find the real probability as the value of R approaches infinity:

$$p_{not\_found} = log_{x\to\infty}\left(1-\frac{C}{R}\right)^{R} = \boldsymbol{e^{-C}}$$

Intuitively, this makes sense, since as the number of reads increases, the probability of not finding a single nucleotide becomes more and more unlikely.

The number of expected nucleotides that remain unsequenced is then the probability of not finding a nucleotide at a single location, multiplied by the number of locations, or:

$$Ge^{-C}$$

c) We can find the expected number of contigs, by finding the number of reads we expect to be followed by at least one unsequenced nucleotide. This means if we never have an unsequenced nucleotide, we will have one contig. If we have one unsequenced nucleotide, we will get two contigs. Thus the number of expected contigs can be calculated by finding the number of times we expect to find a read followed by an unsequenced nucleotide. Using the probability of a single location in the genome being unsequenced from part b, we get the expected number of contigs:

$$Re^{-C}$$

We can get the expected length of each contig by dividing the number of sequenced nucleotides (length of the genome minus the unsequenced nucleotides) by the expected number of contigs from above:

$$\frac{G-Ge^{-c}}{Re^{-C}}=\frac{G(1-e^{-c})}{Re^{-c}}$$