Biol 47800/59500 Homework 7

1. Using the Lander-Waterman equation, calculate the following
   a) How many reads do you need to sequence to get 16X coverage of the E. coli Genome (4.6 Mbase) assuming 800 base reads (typical of first-gen sequencing)?
   b) How many reads do you need to sequence to get 16X coverage of the E. coli Genome (4.6 Mbase) assuming 100 base reads (typical of next gen sequencing)? If next gen sequencing costs 10% as much per base, which is cheaper first-gen or next-gen?
   c) Same as a) for the human genome (3.2Gbase)?
   d) Same as b) for the human genome (3.2Gbase)? If next gen sequencing costs 10% as much per base, which is cheaper?
   e) Is next-gen sequencing an suitable method for determining the whole genome sequence of a novel (never been sequenced before) genome of the size of *Homo sapiens*? Why or why not?

2. You have determined a new eukaryotic genome and run a standard sequence-based gene prediction program (such as genemark). Some of the gene predictions may be excellent, and some may be poor. List 5 methods that you can use to evaluate the gene predictions and decide if they are correct. For each method describe the differences you will observe for real genes (correctly predicted and expressed) and false genes (sequence does not actually encode a gene). For the purpose of this exercise, you do not have to consider detecting partially correct gene models; assume they are either completely true or completely false. You may assume that any other useful experimental information that you need is available for this genome.