1. Convince yourself that the Poisson approximation to coverage is accurate by simulating a genome sequencing experiment! Imagine you are sequencing a 1 megabase genome with 10,000 reads that are 100bp long (what’s the coverage?).
2. Import the random module and use random.randint to sample the 5’ end of reads uniformly at random. For each position in the genome, count how many times it is covered by a read.
3. Compute the fraction of sites that are covered 0 times.
4. Repeat this exercise for 50,000 reads and 100,000 reads (you can use the same script, just change the number of reads you simulate)
5. The expected fractions of sites with no coverage under the Poisson approximation are 0.37, 0.0067, and 0.000045 for 10,000 reads, 50,000 reads and 100,000 reads, respectively. How does your simulation compare?
6. Computing the N50 of a draft genome. We will use the genome located at https://www.dropbox.com/s/nv8i0773031rzs6/test.fasta (hint: use wget to download it directly from within bash. If you’re on MacOS, you’ll have to first download homebrew (<http://brew.sh/> ) and then type “brew install wget” into terminal).
7. Write a function that reads a fasta file into a dictionary. The keys should be the contig (or chromosome) names, and the values should be the sequence of that chromosome.
8. Implement a function that computes the N(x) of a genome based on the dictionary you create by reading a fasta file. You should implement the pseudocode I ran through in class.
9. Write the N(x) for x = 5%, 10%, 15%, and so on up to 95% to a file
10. Download the fastq file located at <https://www.dropbox.com/s/lma9iag69mj5fcq/I1303.1240k.fastq>. It consists of ancient DNA from some old human.
11. Ancient DNA tends to be highly degraded. Compute the length of every read.
12. Look up how to plot a histogram of the lengths. Plot a histogram using 100 bins.
13. Compute the base composition per site in reads. i.e. what is the fraction of As, Cs, Gs and Ts in every position of every read.
14. Plot the base composition along reads, with 1 line for each of A, C, G and T and a legend.
15. Compute the average quality score for every position in a read
16. Plot the average quality score along reads
17. Compute the average quality score of ENTIRE reads (so you should have one score for every read) and plot a scatterplot of average quality score against read length.