Shotgun Sequencing

For this exercise, you need to have the statistical software R installed on your computer. You can download a version of R for most operating systems from:

http://www.r-project.org/

You also need to download the text file "ShotgunSimulator.R" from the course website.

Once you have R installed, you should put the file "ShotgunSimulator.R" into the working directory of R. To find out the working directory, you can type

getwd()

at the R prompt. To change the directory, you can type

setwd("*directory*")

where *directory* should be replaced with the directory that you want.

If all of this works correctly, you should then be able to execute the R script by typing

source("ShotgunSimulator.R")

This should print a number to your R window and also produce a new window with a histogram.

Initially, the script is set to simulate 1X coverage of a 100-bp genome, where each sequence read is one base long (this is very similar to the "experiment" we did during the lecture). The whole process is repeated 500 times and the mean proportion of unsequenced bases over the 500 replicates is written to the R window. The other window shows a histogram of the proportion of unsequenced bases over the 500 replicates.

To change the coverage, genome length, or the number of replications, use a text editor to change the values of

COVERAGE <- 1
LENGTH <- 100
REPLICATIONS <- 500

in the ShotgunSimulator.R file. Be sure to save the changes, then re-run the script with the same command line given above.

Run the script with coverages of 1X, 2X, 3X, 4X, and 5X.

For each case, what is the mean proportion unsequenced? How does this compare to the expected values (from the Poisson distribution) of 0.37, 0.14, 0.05, 0.02, 0.007, respectively?