Homework 3

1. **GWAS Study Analysis.** You are given the results of a case /control GWAS study (links to data files are on website). We will now proceed to analyze the results. For this assignment we will use plink (http://pngu.mgh.harvard.edu/~purcell/plink/) and Haploview (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview)

a. Run a chi-square test on the example data. Report the SNPs with the top 10 odds ratio scores

b. The reported p-values have already been Bonferroni corrected, give uncorrected Bonferroni p-values of the top 10 SNPs from A.

c. Using Haploview, generate a Manhattan plot of the data. What are the loci that have very high odds ratio scores? (look by eye, report the chromosome and general location)

d. These regions could indicate SNPs correlated with the disease, but what are some other possible reasons that the SNPs could be correlated? Give justifications and possible real world examples.

2. **DNA sequencing**

FIGURE 1

FIGURE 2

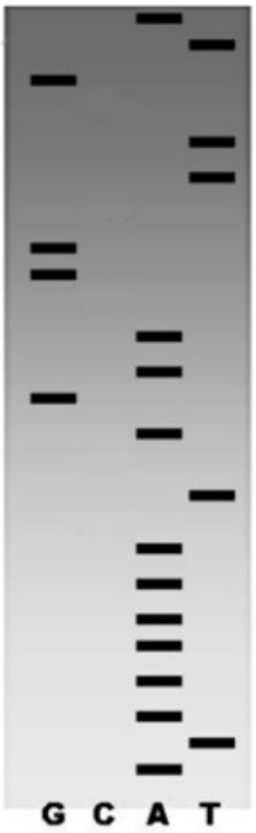
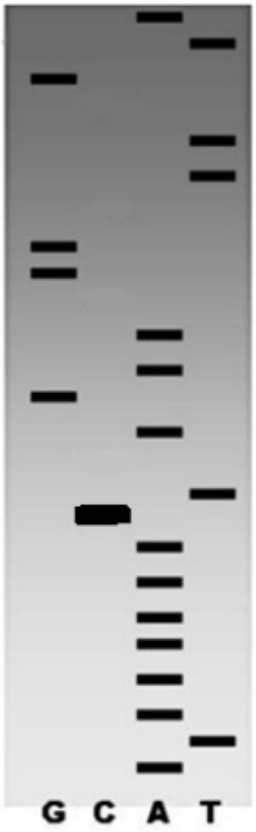


FIGURE 3



a) What are the first three bases of the template strand on the 5' end (Figure 1)? How do you know? *Note, the lane labels correspond to the dideoxy radiolabel added (G = ddGs added).

b) Your advisor tells you to confirm the exact same sequencing experiment with another run, and you get the picture above 2) Provide one possible explanation for this anomaly.

c). Your advisor tells you to get it right this time. You get the picture above (Figure 3). Provide one possible explanation for this anomaly.

Hint: Perhaps this is an overreaction to an earlier mistake.

d) If the same experiment were run in a thicker matrix for the same amount of time, what changes would you expect to see?

e) Derive a general formula for the probability of seeing a fragment of length N where the fraction of ddNTPs in the mix (versus all NTPs) is P? What kind of distribution is this?

3. **Confusion Matrices.** We have developed an algorithm to predict disease based on SNP calls. Each SNP in the genome is either predicted to be assocated with a disease or not. Below is a confusion matrix that has the results of the algorithm. We will analyze the confusion matrix and determine the effectiveness of the algorithm.

	True Disease	True Normal
Predicted Disease	644504	64327
Predicted Normal	522784	2062010

a) Report the Sensitivity, Specificity, Power, and False Discovery Rate (FDR) of the given confusion matrix.

Describe, in English, what each of the values you reported means.

b) What would improve the current algorithm more, adjusting it so it accurately detected more disease cases or detected more normal cases? Provide an argument as to why your answer is correct.

c) Are Sensitivity, Specificity, and FDR independent quantities? If possible, give a confusion matrix that has the same sensitivity and specificity as the one above, but 10 times the FDR.