Genome 560
Homework 3
Due April 15, 2008

1. [15 pts] In lecture this past Tuesday we calculated the maximum likelihood estimate of the proportion of recombinant gametes between two heterozygous loci. In class on Thursday we walked through these calculations in R. Now, we are interested in applying this cool statistical method to obtaining the maximum likelihood estimate of the number of spontaneous deleterious mutations in humans. You develop a potential Noble prize worthy method of counting spontaneous deleterious mutations and apply it ten individuals and obtaining the following data:

Individual	**Number deleterious mutations**
1	4
2	3
3	0
4	1
5	9
6	2
7	3
8	8
9	0
10	6

A. Write out the likelihood function for this data. [Hint, we discussed in class the probability distribution that this type of data follows]. Simplify as much as possible.

B. Calculate the maximum likelihood estimate of the average number of spontaneous deleterious mutations per individual from the data above. You can do this in any of the three ways that we discussed in class (graphically, numerically, or calculus).

C. Plot the -log likelihood of the data.

D. Hypothesis testing in the likelihood framework is straightforward and proceeds through a likelihood ratio test (LRT). Specifically, the LRT is defined as:

$$\Lambda = \frac{L(\theta = \hat{\theta}_0)}{L(\theta = \hat{\theta}_A)}$$

where the numerator is the likelihood evaluated at the mle of theta under the null hypothesis and the denominator is the likelihood of the data evaluated at the mle of theta under the alternative hypothesis. Asymptotically, -2 ln Λ follows a chi-square distribution with 1 degree of freedom (at least for this example; technically the degrees of freedom depends on how many parameters you are estimating, which will be one). Perform a LRT that theta is greater than zero (i.e., the

numerator is the likelihood of the data when $\hat{\theta}_0 = 0$ and the denominator is the likelihood of the data for $\hat{\theta}_a$ estimated in part B). What is the test statistic value and p-value?

2. [10 pts] You are interested in the transcriptional changes during early stages of the innate immune response. You obtain lymphoblast cell lines from 10 individuals and for each one measure expression levels at baseline (untreated) and following treatment with the drug immiquimod (which is a TLR8 agonist). The following table shows gene expression levels for a particular transcript.

Individual	**Baseline**	**Stimulated**
1	-0.24	1.74
2	0.25	2.1
3	1.12	1.65
4	-0.06	2.65
5	0.46	3.11
6	0.17	2.31
7	0.02	1.87
8	1.10	3.21
9	0.55	2.19
10	0.98	1.75

A. Perform a one sample t-test to test the hypothesis that baseline expression levels are significantly different than zero. Clearly state the null and alternative hypotheses and submit R code, test statistic value, and p-value.

B. Use a paired t-test to test the hypothesis that gene expression levels are significantly different between baseline and stimulated conditions. Again, clearly state the null and alternative hypotheses and submit R code, test statistic value, and p-value.

C. An alternative way of analyzing the data as opposed to a paired two sample t-test (part B) is to create a new phenotype for each individual defined as the difference between Stimulated and Baseline expression. Formally, let x_i and y_i denote the expression level for the ith individual in baseline and stimulated conditions, respectively. Then define $z_i = y_i - x_i$. Perform a one sample t-test on the vector of z_i values. Clearly state the null and alternative hypotheses and submit R code, test statistic value, and p-value. How does your result compare to that obtained from part B?