Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
1st section synthesis problems
Due Friday, March 14 (π Day!!)

This assignment is designed to help you synthesize what you've learned so far in the class. You may consult any written work (including the computer/web) but no people. Please do not talk to anyone except me about the assignment (including your friends or professors in other departments). I am happy to talk with you about the problems, and I expect you to contact me immediately if any of the questions are unclear in any way. I will be very free with any advice concerning problems done in R. Also, because it makes my life much easier, please type anything that isn't mathematics or equations.

1. Why is complementary base pairing important for micorarrays?

2. What are the main factors influencing the level of fluorescence in a microarray experiment? (This is not a trick question. I want to know the sources of variability that make each of the numbers different on the original spreadsheets you downloaded from SMD. I don't want you to talk about normalization here.)

3. How / why do we use MA plots? In your response, make sure to carefully describe what an MA plot is.

4. An analyst says, "One nice thing about loess regression is that you need not worry about outliers and influential observations." Comment (separately for both the context of outliers in the y direction (big y value) and influential observations in the x direction (big x value)).

5. One "normalization" technique we've talked about is median subtraction. Explain both the technique (how do we do it?) and the reasoning behind the technique (why do we do it?).

6. The following are "M" values for genes # 3201 & # 5989 from Laura Hoopes' data. (If I haven't done so already, remind me to send you these values over email so that you don't have to type them in.)

Gene # 3201			
NA	NA	5.552615469	3.588512653
10.091811010	NA	1.182945377	0.230653842
0.161734418	-0.242071045	0.113605434	-0.154127317
NA	NA	-0.189606855	0.113506120
-0.001467150	NA	NA	NA
NA	NA	NA	NA
NA	NA	2.319764969	NA
7.664714440	NA	NA	

Gene # 5989			
0.54278818	-0.12035038	-0.22198253	-0.01598358
0.35051206	-0.09428862	-0.27269071	0.54936201
0.37331346	0.23772638	1.31325982	-0.05134144
5.35111340	-0.15487961	-0.05484847	-0.04808720
0.24080726	-0.47890400	0.20319238	-0.03722995
-0.14238327	0.19456324	1.11094350	-0.52076122
-0.17906216	-0.15542533	0.33838833	-0.27703942
0.06020039	-0.17617830	-0.42120518	

   (a) Using both Kolmogorov-Smirnov and Looney-Gulledge, test & discuss the normality of these two genes.

   (b) Include a relevant plot for each of the genes.

   (c) Do not be afraid to make decisions about whether or not a particular data value is reliable.

   (d) This question will be graded mostly on the quality of the explanations and demonstrated understanding of the situation (and less on one particular right answer).

7. Graphically represent and comment on the following 2-color microarray designs. Your graphical representation should be well labeled in terms of references, dyes, arrays, randomization, etc. Give at least two advantages and disadvantages of each of the experimental designs suggested.

   (a) An experiment is planned for studying the effect of a plant steroid hormone on gene expression in leaves of maize. Two-week-old maize seedlings will be used to receive the hormone treatment or mock treatment (similar treatment to the hormone treatment except without adding the hormone). The length of treatments will be either 2 or 8 hrs. In total, we have 4 treatment groups (Hormone 2hrs, Hormone 8hrs, mock 2hrs and mock 8hrs). We randomly assign 3 seedlings to each treatment group. After the experiment, mRNA samples are prepared from each seedling, reverse-transcribed to cDNA and labeled with Cy5. The cDNA from another seedling receiving neither hormone nor mock treatment is prepared and labeled with Cy3. Suppose that we have enough of the Cy3-labeled sample and

that is used as a reference here. Each of the Cy5 labeled sample will be applied to a slide together with an equal amount of reference sample for this microarray experiment.

(b) We are interested in profiling the gene expression changes in yeast cells over time after adding inositol (one chemical) to the growth medium. Before the experiment, cells were grown in medium without inositol for at least 12 generations. Then cells were split into 16 dishes and each dish is one experimental unit. Inositol was added to the medium of 12 dishes at a specified concentration at time 0. The 4 dishes without addition of inositol were harvested by filtration at time 0. Dishes with addition of inositol were harvested at 5-, 15-, and 30-min intervals following the addition of inositol (4 dishes at each time point). For each of the 16 samples, mRNA samples were extracted and reverse-transcribed to cDNA. Then half of the samples at each time point (0, 5, 15, 30-min) were labeled with Cy5 and the other half were labeled with Cy3.

Samples that were harvested at adjacent time points were hybridized together on one slide with the Cy3 dye used for the earlier time point and the Cy5 dye for the later time point. In addition, samples from the time-0 cells and the time-30 cells were hybridized together on a slide with the time-0 samples dyed Cy5 and the time-30 samples dyed Cy3. The entire process was repeated one more time. In the repetition, all dye assignments were reversed.

8. Write a question of your own, and answer it. No constraints except for the fact that the question should be thoughtful.