Solutions to Exam 1 from a Past Semester

1. Provide a brief answer to each of the following questions.

a) What do *perfect match* and *mismatch* mean in the context of Affymetrix GeneChip technology? Be as specific as possible in your answer.

A perfect match is an oligonucleotide, 25 bases in length, that is intended to be perfectly complementary to a target sequence of interest. A mismatch is exactly like a perfect match except that the middle (13^{th}) base is changed to the complement. (3 points)

b) True or False: Robots are used to print microarray slides so that gene locations can be easily randomized from slide to slide.

False. It is not easy to randomize gene locations from slide to slide. Usually the genes are in the same relative positions on each slide. (3 points)

c) What is a probe set? Please be as specific as you can.

A probe set is a collection of perfect match and corresponding mismatch oligonucleotide sequences (probe pairs) that are designed to probe the expression of a given gene or target sequence of interest. Millions of each perfect match and each mismatch are contained probe cells that together make up a probe set. (3 points)

d) With a color depth of 16 bits/pixel, how many different signal values are possible?

2^{16} (integers 0 through 2^{16}-1) (3 points)

e) Explain the meaning of *intensity-dependent dye bias*.

Dye bias refers to the idea that for some genes one dye might yield higher fluorescence intensities than the other even when equivalent numbers of transcripts are intended to be associated with each dye. Intensity-dependent dye bias exists if the extent of the dye bias varies with the intensity of the fluorescence level. For example if high-signal genes have a large red-over-green bias while low-signal genes have no bias or perhaps even a green-over-red bias, then intensity-dependent dye bias would be present. (3 points)

2. An experiment was conducted to study the effects of soil temperature on gene expression in developing soybean plants. A total of 18 soybean plants were randomly assigned to 6 containers so that each container held 3 plants. Each container had a separate control that could be used to adjust the soil temperature to any desired level. Three of the 6 containers were randomly selected to be set at a common cold soil temperature. The other 3 containers were set at a normal soil temperature. At the conclusion of the experiment, Affymetrix GeneChips were used to measure RNA levels with one GeneChip per plant.

a) Name the experimental units in this experiment.
containers (3 points)

b) Name the observational units in this experiment.

plants (3 points)

c) Was blocking used in this experiment? If so, describe the blocks.

There is no blocking in this experiment. (3 points)

d) Name the treatment factor(s) in this experiment and list the levels of each treatment factor.

Soil temperature is the treatment factor. (3 points)

e) Write down a model for the data from a single gene. You may use the abbreviated notation described in class and in our course notes. Circle any terms in the model that you would treat as random.

Y = temperature container

I underlined, rather than circled, the random term container. Container should be random because its levels are experimental units which provide multiple observations (three total -- one for each plant). Plant is not needed because there is only one observation per plant so that plant is confounded with the residual. (5 points)

f) Suppose normalized natural-log-scale data for a single gene is as follows:

		Plant			
Container	Soil Temp.	1	2	3	*mean*
1	normal	7	3	8	*6*
2	normal	6	3	3	*4*
3	normal	7	9	8	*8*
4	cold	1	4	1	*2*
5	cold	3	2	1	*2*
6	cold	5	7	3	*5*

(Note this "data" has been set at integer values to make computations easier.)

Provide an estimate of fold change for this gene that describes the effect of soil temperature on this gene's expression level.

average for treatment 1 = (6+4+8)/3=6

average for treatment 2 = (2+2+5)/3=3

fold change estimate is exp(6-3)=exp(3)=20.09

(5 points)

g) Provide a 95% confidence interval for the fold change estimated in part (f). Assume that the *t*-distribution quantile required for the computation of the confidence interval is 2.776.

sample variance for treatment 1 = 4

sample variance for treatment 2 = 3

*pooled estimate of the variance = { (3-1)*4+(3-1)*3) } / { (3-1) + (3-1) } = 3.5*

SE = sqrt{ 3.5 (1/3+1/3) } = sqrt(7 / 3)

confidence interval is

*exp{ 3 – 2.776*sqrt(7/3)} to exp{ 3 + 2.776*sqrt(7/3)}*

0.2892656 to 1394.666

(6 points)

h) Based on your 95% confidence interval, do you think the expression level of this gene was affected by soil temperature? Explain.

It is possible that the expression level was affected by the soil temperature, but it is also quite possible that the soil temperature had no effect on the expression level. The confidence interval for the fold change includes 1, which indicates that no change in expression level in response to treatment is plausible. (4 points)

3. An experiment was conducted to study the effects of soil temperature and chemical treatment on gene expression in developing soybean plants. A total of 18 soybean plants were randomly assigned to 6 containers so that each container held 3 plants. Each container had a separate control that could be used to adjust the soil temperature to any desired level. Three of the 6 containers were randomly selected to be set at a cold soil temperature. The other 3 containers were set at a normal soil temperature. Three chemical treatments (A, B, and C) were randomly assigned to the plants in each container such that one of the three plants was selected for treatment with chemical A, another for treatment with chemical B, and the third for treatment with chemical C. At the conclusion of the experiment, Affymetrix GeneChips were used to measure RNA levels with one GeneChip per plant.

a) Name the experimental units in this experiment.

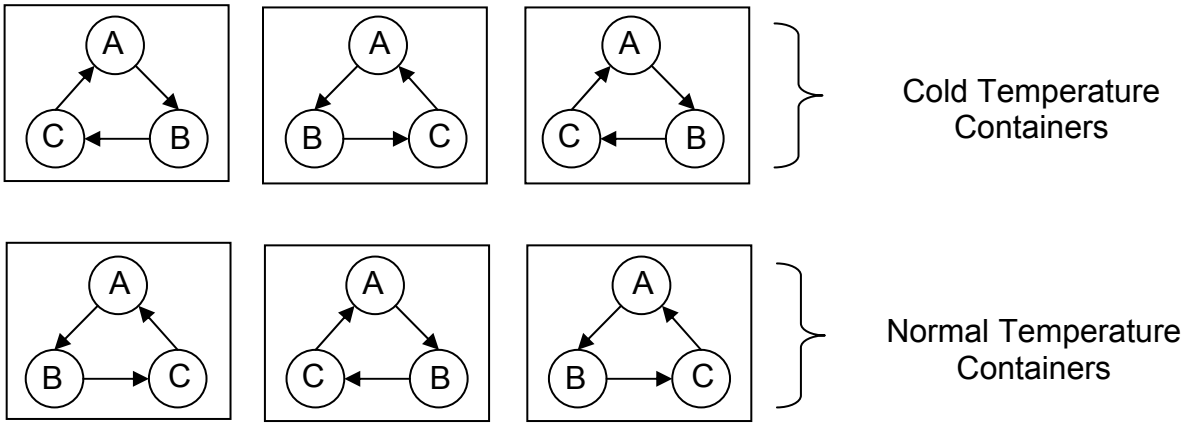
This is a split-plot experiment. With respect to the treatment factor "soil temperature," the experimental units are containers. With respect to the treatment factor "chemical," the experimental units are individual plants. (4 points)

b) Write down a model for the data from a single gene. You may use the abbreviated notation described in class and in our course notes. Circle any terms in the model that you would treat as random.

Y=temp chem temp:chem container

I underlined, rather than circled, the random term container. Container should be random because its levels are whole-plot experimental units. Plant is not needed because there is only one observation per plant so that plant is accounted for by the residual. (6 points)

c) Suppose that 18 two-color microarray slides will be used to measure expression instead of 18 GeneChips. Sketch a plot using our microarray circle and arrow notation to indicate how you would use two-color microarrays to measure RNA levels in the plants if the researchers are primarily interested in detecting differences among chemical treatments within each temperature.



(8 points)

d) Write down a model for the two-color array data from a single gene based on the design you have specified above. You may use the abbreviated notation described in class and in our course notes. Circle any terms in the model that you would treat as random.

Y=temp chem temp:chem dye container plant slide

We need to add dye as a fixed factor and plant and slide as random factors. The factor plant is now necessary because there are two observations for each plant (and plants are the split-plot experimental units). (6 points)

e) In the context of this example, explain the meaning of temperature-by-chemical interaction.

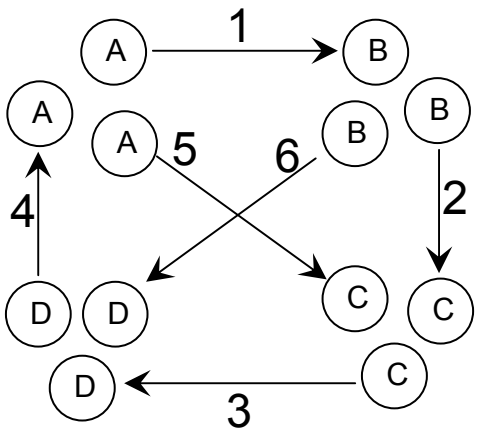
A temperature-by-chemical interaction would mean that differences among the levels of expression associated with each chemical under cold temperature would be different from the differences among the levels of expression associated with each chemical under normal temperature. (5 points)

4. Suppose a two-color microarray experiment is to be conducted to compare the effect of four treatments (A, B, C, and D) on gene expression in maize. Suppose that treatment D is a control and that researchers are primarily interested in understanding which of the treatments A, B, and C differ from the control treatment D in terms of mean expression for each gene. The

researchers have 12 experimental units and 6 slides available for the experiment. For any given gene, denote the mean expression of an observation as a function of dye and treatment according to the following table:

Treatment	Dye	Mean Expression
A	Cy3	$u+d_3+a$
B	Cy3	$u+d_3+b$
C	Cy3	$u+d_3+c$
D	Cy3	$u+d_3$
A	Cy5	$u+d_5+a$
B	Cy5	$u+d_5+b$
C	Cy5	$u+d_5+c$
D	Cy5	$u+d_5$

a) Consider the balanced incomplete block design (depicted below) that compares each treatment to each other treatment on exactly one slide. Provide the appropriate X matrix for this design. Assume that we will use the Cy5-Cy3 difference on each slide as our response variable and that our parameter vector for this analysis is $[d_5-d_3,\ a,\ b,\ c]'$. (Note that the numbers in the diagram below correspond to slide numbers. Please enter the rows of X accordingly.)



X =

(slide 1)

(slide 2)

(slide 3)

(slide 4)

(slide 5)

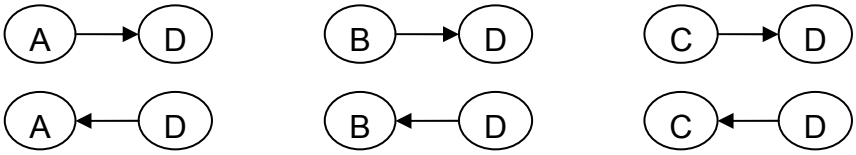
(slide 6)

1 -1 1 0

1 0 -1 1

$X = \begin{matrix} 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{matrix}$ *(6 points)*

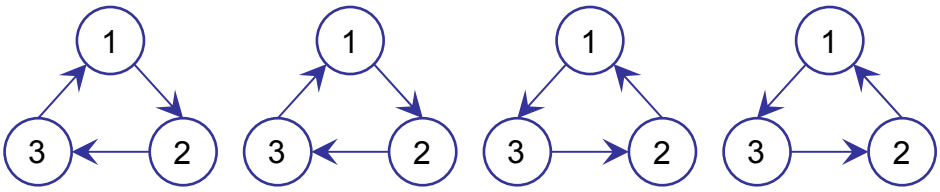
b) There exists at least one 6-slide 12-experimental-unit design that will dominate the design in part (a) for testing for differences between the control treatment (D) and each of the other three treatments (A, B, and C). Draw a diagram for a design that you believe will dominate the design in part (a). You will receive full credit if your chosen design does indeed dominate the design in (a). You do not need to do any calculation to show that your design dominates the design in (a); simply provide a diagram for such a design.



Recall that each experimental unit is represented by a circle and each slide by an arrow. Thus every drawing should have had 12 circles and 6 arrows. There was no restrction on the number of experimental units for each treatment, given that 12 experimental units total were used. Some of you seemed to incorrectly assume that it was necessary to have 3 for each treatment.

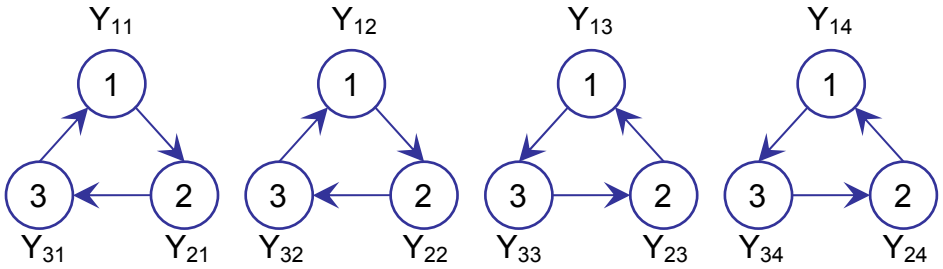
(8 points)

5. Consider a completely randomized experiment with three treatments denoted 1, 2, and 3 and four experimental units per treatment. Suppose mRNA levels are measured with two-color microarrays using the following design.



In class we discussed a mixed linear model for the 24 observations obtained for a single gene. This model included an overall mean μ and fixed factors treatment and dye. Denote the treatment effects by τ_1, τ_2, and τ_3, and denote the dye effects by δ_1 and δ_2. The model we discussed included variance components for the random factors slide, experimental unit, and residual. Denote these variance components by σ_s^2, σ_u^2, and σ^2 respectively. Assume this model is correct throughout this problem.

Suppose that instead of analyzing the 24 observations or instead of taking red-green differences as discussed in class, the two observations for each experimental unit are averaged to obtain a total of 12 averages denoted by Y_{ij} in the figure below.



a) Determine the expected value (mean) of Y_{11} in terms of the model parameters.

$\mu + \tau_1 + (\delta_1 + \delta_2)/2$ *(3 points)*

Please see me if you don't understand how to obtain this answer or any of the other answers below.

b) Determine the expected value (mean) of $Y_{11} - Y_{21}$ in terms of model parameters.

$\tau_1 - \tau_2$ *(3 points)*

c) Determine the variance of Y_{11} - Y_{21} in terms of the model parameters.

$\sigma_s^2/2 + 2\sigma_u^2 + \sigma^2$ *(4 points)*