Answer all five (5) questions.  
Each question is worth 15 marks.  
Full marks can be obtained by correct solutions to all five questions.  
Figures 1, 2 and 3 are attached.  
Calculators are permitted for simple numerical calculations.
1. (a) Figure 1 attached shows an $M$ versus $A$ plot for data from a microarray experiment comparing mRNA from two mutant mouse cell lines. There are 16,128 genes spotted on the microarray and only a relatively small number of genes are expected to be differentially expressed between the two cell lines; the line drawn on the plot is the lowess curve fitted through all the points. Describe any features or patterns you observe in this plot.

(b) Figure 2 shows a spatial plot of the top 500 (positive) and bottom 500 (negative) log ratios for the microarray data depicted in Figure 1. The genes on the slide were spotted using 32 pins (in 8 rows by 4 columns).

(i) Describe any observed spatial effects on the microarray slide.

(ii) State two possible causes of spatial effects in microarray data.

(c) Describe what is meant by the term normalization.

(d) For the microarray data shown in Figures 1 and 2, suggest the most appropriate form of normalization, and justify your choice.

[15 marks]

2. (a) Consider a microarray slide for which there are $i = 1, \ldots, I$ print-tip groups and $j = 1, \ldots, n_i$ genes within each print-tip group. Assume that the log ratios have been location normalized within the $i$th print-tip group, that they are denoted $M_{ij}$, and that they follow a normal distribution with mean zero and variance $a_i^2 \sigma^2$, where $\sigma^2$ is the variance of the log ratios and $a_i^2$ is a scale factor for the $i$th print-tip group.

(i) Write down the likelihood function for the location normalized log ratios $M_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$.

[Note that the normal density function for a random variable $X$ with mean $\mu$ and variance $\sigma^2$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}, \quad -\infty < x < \infty.$$

(ii) Find the maximum likelihood estimating equations, and hence show that

$$n_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{a_i^2 \sigma^2}.$$

(b) (i) Suppose two mRNA cell lines, $A$ and $B$, are hybridized to a slide, and that the experiment is replicated $k = 1, \ldots, n$ times. Assume that the replicates are independent and let $\sigma_j^2$ be the gene-specific variance for the normalized log ratios $M_{jk}$ corresponding to the $j$th gene. Find the variance of the average log ratio, $\bar{M}_j$.

(ii) In a replicated reference design, the mRNA cell line $A$ is hybridized with a pooled reference sample and the hybridization is replicated $n_1 = n$ times; independently,
the mRNA cell line $B$ is hybridized with the same pooled reference sample and this hybridization is repeated $n_2 = n$ times. Again assume that all replicates are independent.

Further assume that $\sigma^2_{A_j} = \sigma^2_{B_j} = \sigma^2_j$, and hence find the variance of $\bar{M}_{A_j} - \bar{M}_{B_j}$.

(iii) How many hybridizations would need to be conducted using the reference design to achieve the same precision as can be achieved with $n$ direct comparisons?

[15 marks]

3. (a) Let $X$ be a continuous random variable with density function $f(x)$ and cumulative distribution function $F(X)$, and assume $F^{-1}$ exists. Find the distribution of the random variable $U = F(X)$.

(b) Figure 3 shows a quantile-quantile plot (i.e., qq-plot) of paired $t$ statistics representing 20,800 genes from an angiogenesis experiment in which mRNA from time zero hours was hybridized with mRNA from time three hours; the experiment was replicated four times.

(i) For each $t$ statistic, state the appropriate null hypothesis.

(ii) What does the straight line through the origin in the plot represent?

(iii) Describe any observed features in this qq-plot.

(iv) Assuming a Type I error of 5% for each test, approximately how many false positives would be expected?

(v) Explain briefly how substantially increasing the level of replication would reduce the number of false negatives.

[15 marks]

4. (a) Define the experiment-wise or family-wise Type I error rate.

(b) Suppose we wish to conduct $j = 1, \ldots, g$ hypothesis tests simultaneously with a desired family-wise error rate of $\alpha$. Assume that the null hypothesis, $H_0$, of no differential expression is true for each test, and let $A_j$ be the event \{Reject test $j$\}.

Use the following inequality to prove that the Bonferroni method can control the family-wise error rate at at most $\alpha$;

$$P(\bigcup_{j=1}^{g} A_j) \leq \sum_{j=1}^{g} P(A_j).$$

(c) If 1,000 hypothesis tests are conducted simultaneously, how could we achieve a family-wise error rate of at most 5%?

(d) Define the adjusted $P$-value, $\hat{p}_j$, for the Bonferroni method.

(e) Explain briefly why the Bonferroni method is not useful for microarray data.

[15 marks]
5. (a) Several important tests in bioinformatics are carried out by examining the observed value of the maximum of several random variables. Let $X_1, \ldots, X_n$ be $n$ independent continuous random variables each having the uniform distribution on $(0, L)$ for some constant $L$.

Find the density function of the maximum, $X_{(n)}$. Then find the mean and variance of this distribution.

(b) The sum of the ranks in group 1 for the Mann-Whitney test can be written as $R = \sum_{j=1}^{n_1+n_2} jI_j$, where $I_j = 1$ if the $j$th ranked observation is in group 1, and $I_j = 0$ otherwise, and $n_1$ and $n_2$ are the numbers of observations in groups 1 and 2 respectively. Under the null hypothesis $H_0$, the two groups of observations come from the same population, and the ranks for group 1 are an i.i.d random sample from $\{1, 2, \ldots, n_1 + n_2\}$.

(i) Find the expected value of $R$ under the null hypothesis, $H_0$.

(ii) Under $H_0$, prove that

$$E(I_jI_k) = \frac{n_1(n_1 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}, \quad j \neq k.$$
Figure 1. $M$ versus $A$ plot for a microarray comparing two mutant mouse cell lines.
Figure 2. Spatial plot of unnormalized log ratios, $M$, for the microarray slide comparing two mutant mouse cell lines. The darker rectangles correspond to positive (i.e., red) log ratios, and the paler rectangles correspond to negative (i.e., green) log ratios.
Figure 3. Quantile-quantile plot of 20,800 single-sample $t$ statistics for the angiogenesis experiment.