Microarray experimental design: power and sample size considerations

M. C. K. Yang, J. J. Yang, R. A. McIndoe, and J. X. She

1 Department of Statistics, University of Florida, Gainesville, Florida 32611; 2 Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta, Georgia 30912

Submitted 20 March 2003; accepted in final form 30 September 2003

Yang, M. C. K., J. J. Yang, R. A. McIndoe, and J. X. She. Microarray experimental design: power and sample size considerations. Physiol Genomics 16: 24–28, 2003. First published October 7, 2003; 10.1152/physiolgenomics.00037.2003.—Gene expression analysis using high-throughput microarray technology has become a powerful approach to study systems biology. The exponential growth in microarray experiments has spawned a number of investigations into the reliability and reproducibility of this type of data. However, the sample size requirements necessary to obtain statistically significant results has not had as much attention. We report here statistical methods for the determination of the sufficient number of subjects necessary to minimize the false discovery rate while maintaining high power to detect differentially expressed genes. Two experimental designs were considered: 1) a comparison between two groups at a single time point, and 2) a comparison of two experimental groups with sequential time points. Computer programs are available for the methods discussed in this paper and are adaptable to more complicated situations.

gene expression; statistical analysis; functional genomics

HIGH-THROUGHPUT MICROARRAY technologies have created unparalleled opportunities to study the mechanism of disease, monitor disease progression, and evaluate effective therapies. Because tens of thousands of genes are investigated simultaneously with a technology that has not been perfected, there is a large uncertainty in the decision process. Therefore, one must rely on statistical analyses of the data to draw conclusions from microarray experiments. A number of methods have been developed to extract useful and robust information from microarray data sets once the hybridizations are complete. However, not as much attention has been paid to the experimental design before the investigator actually completes the experiment.

One of the principal questions while designing a microarray experiment is: What is the sample size required for a microarray experiment to achieve a user defined statistical threshold? To achieve this goal, one must decide how the data will be analyzed and what conclusions can be made. Since microarrays can be used to answer a variety of questions, different experiments will require different designs and analyses. In this study, we will concentrate on finding expression differences between two treatment groups. The goals of the experiments can be classified into three categories: 1) comparing two groups at a single time point, 2) examining the treatment effect over time when the without treatment effect is known, and 3) comparing two groups at sequential time points. We will investigate the sample size requirements necessary to satisfy a specified false discovery rate (FDR) and power (1 − β) given various biological and experimental variation.

Although there are other reports on how to compute the sample size, none of them is as simple and straightforward as the methods presented here. In our experience, we often observe a large number of genes that show expression differences in a typical microarray analysis. The key is that we want to make sure that only a small number in these selected genes are false positives (discoveries). A brief survey of some of the other methods is in the DISCUSSION.

METHODS

cDNA microarrays are printed on glass microscope slides using established protocols (15). The analysis of microarray data can be separated into several stages. The first stage is to eliminate spots with very weak signal, which we define as the signal that is indistinguishable from the background noise. Yang et al. (17) demonstrated that any signal less than 3 standard deviations of the background noise should be discarded. Due to the variation in printing material on the cDNA arrays, the expression is more accurately measured by the ratio of sample intensity vs. intensity from a common reference (5, 12). To derive the ratio, the intensities from both channels have to be normalized. Although there are a number of methods to normalize the two channels, Yang et al. (17) determined that there was actually a very small difference among the different normalization methods (10) when spots with sufficient signal are chosen. In this study, we assume that the analysis will be done after flagging for weak spots, log-transformation, and normalization. Given these assumptions, the useful data is now $Y_{ijk}$, where $i$ is the treatment index, $j$ is the subject index, $t$ is the time index, and $k$ is the replicate index. A few assumptions have to be made in order to obtain the analytic results for sample size determination. The most general model in this study is

$$y_{itk} = \mu_i + \delta_{it} + \epsilon_{itk}$$

where $\mu_i$ is the mean expression level under the $i$th treatment at time $t$, $\delta_{it}$ represents the subject variation, and $\epsilon_{itk}$ is the experimental variation. Throughout this paper, $\mu_i$ is a fixed constant (or fixed effect), and $\delta_{it}$ is a random sample from subjects under treatment $i$, or it is a random effect. Some of the indices can also be eliminated when they are not relevant. For example, when the data is not measured over time, the index $t$ is eliminated, and when there is only one treatment, the index $i$ is omitted.

Adjustment of the threshold significance level. As in the usual hypothesis testing, a significance level $\alpha$ is assigned. When there is only one hypothesis, this guarantees that the chance of incorrectly rejecting the null hypothesis is $\alpha$. When many hypotheses are tested simultaneously, $\alpha$ cannot be considered as the significance level for rejecting all the true null hypotheses. The level for each individual test has to be adjusted. One option for adjusting the test is to use the Bonferroni inequality, i.e., to set every test at the new significant level $\alpha/n_h$, where $n_h$ is the number of hypotheses tested simultaneously. We use the notation $n_e$ because this number is the number of genes in the array. However, when $n_e$ is very large, the Bonferroni adjustment is
too conservative because it is meant to control family-wise error rate (1), i.e., to control the probability of one type I error among thousands of hypotheses. In microarray experiments, it is important to find the best compromise between maximizing the power of detection and minimizing false-positive identification. Very often, although not always, additional follow-up research such as database search and real-time PCR will be done to further eliminate false-positive results (6). Thus the FDR defined as the expected proportion of wrong discovery (1) fits this purpose better. At the sample size determination stage, we need to know approximately the minimum number of genes that will be accepted under a given selection rule. An experimenter may know, from past experience, that at least $n_0$ genes will be accepted, then the adjusted significance level for each gene is

$$
\alpha' = \alpha \max (1, n_0/n_i) \quad (2)
$$

to guarantee an $\alpha$ FDR. The derivation is given in the APPENDIX I. The value $n_0$ has to be determined by experience or a pilot experiment. In our experience, there are always a considerable number of genes that will show different expressions. If there is no way to determine $n_0$, then the experimenter may let $n_0 = 0$ and determine the most conservative (largest) sample size. Sometimes it may be necessary to do a pilot study to determine $n_0$ and other variance components in the model. For our laboratory, it is common practice to do a pilot study to get an idea of these parameters and then determine the sample size for the required power.

At the data analysis stage, other methods such as permutation tests (4, 16) may be used, but their power are not as easy to compute at the design stage.

Single time point comparison between two treatment groups. For the case of a comparing between treatments at one time point, one can drop the $t$ index so Eq. 1 becomes

$$
y_{ij} = \mu_i + \delta_j + \epsilon_{ij} \quad (3)
$$

where $i = 1, 2; j = 1, \ldots, n_c; k = 1, \ldots, m$; $\delta_j$ is the between subjects variation, and $\epsilon_{ij}$ is experimental variation. When $n > 1$, replicates are used to reduce the intermicroarray variation under the same biological condition. We assume that both of $\delta_j$ and $\epsilon_{ij}$ are normally distributed with zero mean and standard deviation (SD) $\sigma_\delta$ and $\sigma_\epsilon$, respectively.

The hypotheses are

$$
H_0: \mu_1 = \mu_2 \text{ versus } H_1: \mu_1 \neq \mu_2 \quad (4)
$$

for each gene. After the data has been collected, the mean and variance for each gene is defined by

$$
y_i = \Sigma y_{ij}/(n_c m), \quad \hat{\sigma}_i^2 = \Sigma (y_{ij} - y_i)^2/(n_1 + n_2 - 2) \quad (5)$$

and the decision rule is to accept $H_1$ at significance level $\alpha'$ when

$$
t = \left[ y_1 - y_2 \right]/\left[ \hat{\sigma}_1 \left( 1/n_1 + 1/n_2 \right)^{1/2} \right] \geq t_{\alpha'} \quad (6)
$$

where $\alpha'$ is defined by Eq. 2, $\nu = n_1 + n_2 - 2$, and $t_{\alpha'}$ is the upper $\alpha'$ quantile of the $t$ distribution with $\nu$ degrees of freedom. Other types of tests such as unequal variance $t$-test or a Mann-Whitney test can also be used, but the results should be very similar (18). To determine the sample size, we need to specify the power $(1 - \beta)$ at a predetermined level of significant gene expression difference $|\mu_1 - \mu_2|$. In traditional statistics, what we need is the value

$$
\Delta = |\mu_1 - \mu_2|/\sigma_\epsilon \quad (8)
$$

and the variance ratio

$$
\rho = \sigma_\delta^2/\sigma_\epsilon^2 \quad (9)
$$

It is often difficult to determine what $\Delta$ constitutes a significant expression difference. We believe that the overlap concept can make the interpretation of the alternative easier. When the alternative hypothesis is true, we have $|\mu_1 - \mu_2| > 0$. Naturally, any expression closer to $\mu_1$ will be classified to treatment 1 and vice versa. Let $w$ be the proportion of subjects with expression apparently belonging to the other group

$$
w = 1 - \Phi(\Delta/2) \quad (10)
$$

where $\Phi(x)$ is the standard normal distribution function. Although $\Delta$ and $w$ are equivalent and can be easily converted from one to another, it is our opinion that the idea of overlapping between groups has an easier interpretation. A historical survey and recent developments in using overlapping for sample size determination have been published (3, 8, 14). Hwang et al. (9) has also used the same idea in microarray analysis. To determine the sample size based on Eqs. 7 and 11, we use the noncentral F distribution (see APPENDIX II).

Serial time points with one treatment group. This experimental design will measure gene changes under one treatment over time, while the gene expression under no treatment condition is known. Let the time points be 1, 2, . . ., $T$. Since there is only one treatment, the index $i$ can be dropped, and Eq. 1 becomes

$$
y_{ijk} = \mu_i + \delta_j + \epsilon_{ijk} \quad (12)
$$

where $\delta_j$ represents the subject variation. Under the null hypothesis, when there is no treatment, we should have

$$
H_0: \mu_1 = \mu_2 = \cdots = \mu_T \quad (13)
$$

i.e., gene expression is not affected by the treatment. The test for this hypothesis would be a two-way analysis of variance with time and subject as two discrete factors with the subjects being a random effect. The alternative is that $H_0$ is not true, i.e., gene expression has been changed. To determine the sample size, we need to specify the pattern of $\{\mu_t\}$ under $H_1$. Moreover, we need to know how this pattern varies among subjects. We use a simple linear expression by letting

$$
\mu_i = \gamma t_i, \quad \text{and } \delta_j = \delta_j \quad (13)
$$

where $\gamma = 0$ under $H_0$ ($\gamma \neq 0$ under $H_1$), and the $\delta_j$ terms represent the variation of the subjects response, which are assumed to be independent and identically distributed normal random variables with 0 mean and variance $\sigma_\delta^2$ ($\sim N(0, \sigma_\delta^2)$) and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$. The reason for this alternative model is that with a linear time effect, the slope usually changes with the subjects. However, the linear relationship is not crucial, because when time is considered discrete, Eq. 13 can be a good approximation to a quadratic change. For example, the pattern in Fig. 1, left, can be expressed linearly when time is reordered. We have tried to use specific alternatives such as a quadratic model (which covers the linear model) to increase the power of the test. However, the increase in power is very limited. One could also make the model more general by adding correlations between time points. But during the design stage, this correlation is difficult to ascertain. The random effect, $\delta_j$, in the slope should be enough to represent the intra-subject correlation.

Again, we will use the overlap concept to specify the alternative. In the model in Eq. 13, suppose the alternative is $\gamma > 0$. (The case $\gamma < 0$ works almost identically.) We let $w$ be the proportion of subjects (in the population) that show the opposite ($\gamma + \delta_j < 0$) mRNA expression. The power can be computed by simulation using SAS PROC MIXED, i.e., we simulate the data a large number of times under the alternative and use the proportion of the times that $H_1$ is accepted as
reordered from the times on the left.

cant, then 22 subjects per treatment group is considered signifi-
ificant or less overlap between genes from the two treatments.

Serial time points with two treatment groups. This experiment will
measure the gene expression changes over time from two treatment
groups or from one treatment and one control. The model is now the
full model in Eq. 1 with the null hypothesis

\[ H_0: \mu_{ij1} = \mu_{ij2}; \mu_{ij2} = \mu_{ij3}; \ldots; \mu_{ijT} = \mu_{ijT} \]

and the general alternative that \( H_0 \) is not true. This model is based
on two factors, time and treatment, and subjects are nested in the
treatments. The experiment is to test the treatment effect or the
treatment-time interaction. We will not consider the treatment-time
interaction in the design stage. To specify an alternative for the sample
size determination, more parameters are involved. We follow the
linear approximation in Eq. 13 with

\[ \mu_{ijt} = \mu_{ij} + \gamma t \quad \text{and} \quad \delta_{ijt} = \delta_{ijT} \]

where the \( \delta \)'s terms share the same assumptions as those in Eq. 13.
Again we can define the alternative by using the overlap idea with \( w \)
being equal to the proportion of subjects that have the opposite slope
(\( \gamma + \delta_{ij} < 0 \)) in mRNA expression when it belongs to the treatment
group with \( \gamma > 0 \). This model can be expanded to include other
alternatives such as the change of intercept. All of these modifica-
tions can be easily embedded into the SAS program mentioned in
APPENDIX III.

The details of power computation are again given in APPENDIX II.

RESULTS

For a single time point. Figure 2 presents the sample size
requirements for a microarray experiment with 5,000 genes
(\( n_g \)) assayed with an FDR of 0.05 (\( \alpha \)) and power = 0.85 when
\( n_0 = 0, 5, 20, \) and 50. The sample sizes for the two treatments
are assumed to be equal (i.e., \( n_1 = n_2 = n \)) and the number of
replicates equals 1 (\( m \)). The variance ratio (\( r \)) defined in Eq. 9
varies from 0 to 1. For example, if \( n_g = 5,000, r = 0.25, \) and
a 15% or less overlap between genes from the two treatments
is considered significant, then 22 subjects per treatment group
is required to reach a power of 0.85 (\( n_0 = 0 \)), but only 15
subjects if we expect at least 50 genes will show significance.
However, Fig. 2 is used only as a demonstration of the change
of sample size under various conditions. The software given in
APPENDIX III will produce the required sample size under any
significance level, power, and other parameter values.

For serial time points. Although Fig. 2 covers some useful
experimental cases, it is more difficult to provide examples of
general interest in serial time point measurements due to the
large number of parameters. The power depends not only on
the biological/experimental variations as well as the number of
time points (\( T \)), but also the time unit of \( T \). For example, for a
fixed slope \( \gamma \), the relative size of \( \gamma T \) to the experimental
standard deviation plays a crucial role. In other words, seven
points in 1 day will be different from seven points in a week
with the same growth rate \( \gamma \) in Eqs. 13 and 14. Thus the
variance ratio (\( r \)) alone is not sufficient to determine the sample
size.

Figure 3 shows one example on the sample size required for
detecting mRNA expression changes over time. We used the
data from Yang et al. (17). In that study, we found a \( \sigma_r \approx 0.40
\)
for log2-transformed ratios between the sample and reference.
If one assumes \( n_g = 5,000, \sigma_r = 0.5, \) and 5% or less overlap
between the two conditions (effect size = 3.29), and a twofold
change in expression at the end of the time series (\( \gamma T = 1 \)),
then the sample size curves for the one-sample and two-sample
designs can be obtained. The Bonferroni correction (\( n_0 = 0 \))
and expected 1% of genes to be selected (\( n_0 = 50 \)) for replicate
sizes \( m = 1, 2, \) and 4 are shown. The FDR is set at 0.05, and
the power is set at 0.80.

Note that in this design, the end time point is fixed. Thus an
increase in the number of time points can merely add points
between the predefined start and end points. Since the highest
difference happens at the end point, there is only a small
increase in power by observing more than four time points. It
is also quite intuitive that \( T \) can compensate for the number of

\[ n_0 = 0 \]

\[ n_0 = 5 \]

\[ n_0 = 20 \]

\[ n_0 = 50 \]
replicates (m), because both of them increase the information for the treatment effect when the treatment is effective.

**DISCUSSION**

This paper provides a reasonable way to estimate the sample size for microarray experiments. The analysis methods are chosen to be simple in order to make the unknown parameters as few as possible. More sophisticated methods may be used after the data has been collected, but based on our experience, different analysis methods usually produce similar results for the prominent gene effects.

The computer software is easy to comprehend. Thus a user can easily modify the programs when other tests are more appropriate. For example, if treatment-time interactions are a main concern, then this term can be easily added to the SAS procedure. The alternative under other models, if known, can also be easily generated using the simulation program with simple modifications.

The sample size determination discussed in this paper depends on experience and the validity of certain assumptions. Of course, there is a certain possibility of failure that an experimenter must face. Even to determine the sample size for the simplest two-sample t-test, we must assume that the two treatments have the same variance. There is no guarantee on this before the experiment is done. Similarly, in most microarray experiments, a large \( n_0 \) can be expected. If the result shows the assumed \( n_0 \) is too large, then we must admit our misjudgment and may choose to do more experiments to increase the power. This strategy is usually more economical than choosing a large sample size by always assuming the worst scenario.

Zien at al. (18) have presented a method similar to ours. In addition to the t-test, they also mentioned the Mann-Whitney nonparametric test and a model with additive noise. However, they did not address serial time points nor the use of the FDR in the analysis. Their results are based on a simulation which is not as accurate as the SAS output based on a noncentral F distribution. Black and Doerge (2) also used the t-test for the power calculation, but they considered only the number of replications with a Bonferroni correction. Subject level variation and FDR were not considered. Pan et al. (13) discussed sample size based on the mixture model. However, under the null hypothesis when all the means are zero, the mixture model is difficult to identify. Thus it may not be worth the effort to go through such a complex model building procedure at the design stage. The paper by Hwang et al. (9) tried to determine the sample size for finding the best discriminant function for the data similar to those given in Ref. 7. But the goal of finding the discriminant function is different from identifying genes with different expressions. The use of two highly correlated genes adds little discriminant power, but both of them are important to be identified.

**APPENDIX I**

**Adjustment of the Significant Level Based on the False Discovery Rate**

In single hypothesis testing, the significance level \( \alpha \) can be also interpreted as the false discovery rate, FDR, i.e.,

\[ \alpha = \Pr[\text{Claim difference (discovery)} | \text{No difference}] \]

In multiple hypotheses testing, if all the null hypotheses are true, then the Bonferroni correction is also the same as the FDR. However, if at least \( n_0 > 0 \) genes are expected to be accepted, then

\[ \text{FDR} = E[ \frac{\text{GenesWronglyAccepted}}{\text{AlltheGenesAccepted}} ] \]

Moreover

\[ \text{FDR} = E \left[ \frac{\text{GenesWronglyAccepted}}{\text{AlltheGenesAccepted}} \right] \leq E \left[ \frac{\text{GenesWronglyAccepted}}{n_0} \right] = \frac{\alpha' n'}{n_0} \leq \frac{\alpha n_0}{n_0} = \alpha \]

where \( \alpha' \) is the adjusted significance level used in each test and \( n' \) is the number of genes with no expression differences. When \( n_0 = 0 \), it means that there may be no gene accepted. If this is true, then there is no false discovery. However, \( n_0 \) is only the expected number; there may be genes accepted after the experiment. Since the smallest number is 1, \( \alpha' \) is the Bonferroni correction. This gives the max(1, \( n_0 \)) in Eq. 2.
APPENDIX II

Data Analysis and Power Computation

Two sample one time point per subject. The data analysis is based on Eq. 7. For a given \( w \), \( \Delta = 2\Phi^{-1}(1-w) \), and the power is

\[
p = 1 - F]^{1}(\xi, \lambda^2)
\]

where \( F]^{1}(\xi, \lambda^2) \) represents the F distribution of 1, \( v = n_1 + n_2 - 2 \) degrees of freedom with noncentrality parameter \( \lambda^2 \) evaluated at the threshold \( \xi = F]^{1}(\nu, \alpha') \), which is the upper \( \alpha' \) quantile of the F distribution with 1, \( v = n_1 + n_2 - 2 \) degrees of freedom, and

\[
\lambda^2 = \frac{\Delta^2}{(1 + r/m)(1/n_1 + 1/n_2)}
\]

with \( r \) defined by Eq. 9. Since \( r \) and \( m \) are confounded, we may let \( m = 1 \) in the computer program by changing \( r \).

One sample serial time points experiment. The analysis is based on the mixed model with time as a fixed effect and the subject as a random effect. The SAS procedure is given in the software.

For the power simulation, data are generated by

\[
y_{ijt} = (\gamma_i + \delta_{jt}) + \epsilon_{ijt}, \quad j = 1, \ldots, n; \quad t = 1, \ldots, T \quad (15)
\]

where \( k \) in Eq. 12 is embedded into \( y_{ijt} \) and \( \epsilon_{ijt} \) so they are averages over \( m \) values, and \( \delta_{jt} \) is a normal random deviate with 0 mean and variance \( \sigma_{\delta}^2 \), \( \epsilon_{ijt} \) are standard normal deviates, and

\[
\gamma_i = -\sigma_{\gamma} \Phi^{-1}(w)
\]

The generated data is then tested by the mixed model procedure for the significance of the time effect. The proportion of the rejected samples is the estimated power of this test. The mixed model takes care of the random effect from subjects. Users may refer to a SAS example in Ref. 11 (p. 90) to see how SAS program is written.

Two sample serial time points experiment. This analysis is based on a mixed model with time and treatment as fixed effects and the subject as a random effect.

Without loss of generality, we may let \( m = 1 \) and \( \mu_{1t} = \mu_{2t} = 0 \). For a given \( T \) and \( w \), data are generated by

\[
y_{ijt} = (\gamma_i + \delta_{jt})t + \epsilon_{ijt}, \quad i = 1, 2; \quad j = 1, \ldots, n; \quad t = 1, \ldots, T \quad (16)
\]

where \( \delta_{jt} \) and \( \epsilon_{ijt} \) are generated by the same procedure in Eq. 15 with \( \gamma_1 = 0 \) and

\[
\gamma_2 = -2\sigma_{\gamma} \Phi^{-1}(w)
\]

The power is again estimated the same way as the one-treatment case.

APPENDIX III

Software

The software is located at http://web.stat.ufl.edu/~yang/microarray for the microarray sample size determination.

Program one_time.sas. This subroutine contains the SAS program for one time point study based on Eq. 3.

Program sequent_1.sas. This subroutine contains the SAS program for one sample serial time points measurements based on Eqs. 12 and 13. Program sequent_2.sas. This subroutine contains the SAS program for one sample serial time points measurements based on Eqs. 1 and 14.

All the input instructions are given by the comments in the program. Examples corresponding to the sample sizes in Fig. 2 and Fig. 3 are in the input (*.sas) and the output (*.lst) programs.

REFERENCES