BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 24 2005, pages 4378–4383
doi:10.1093/bioinformatics/bti717

Gene expression

Effect of pooling samples on the efficiency of comparative studies using microarrays

Shu-Dong Zhang* and Timothy W. Gant*

MRC Toxicology Unit, Hodgkin Building, Lancaster Road, University of Leicester, Leicester, UK

Received on July 27, 2005; revised on September 22, 2005; accepted on October 12, 2005
Advance Access publication October 18, 2005

ABSTRACT

Motivation: Many biomedical experiments are carried out by pooling individual biological samples. However, pooling samples can potentially hide biological variance and give false confidence concerning the data significance. In the context of microarray experiments for detecting differentially expressed genes, recent publications have addressed the problem of the efficiency of sample pooling, and some approximate formulas were provided for the power and sample size calculations. It is desirable to have exact formulas for these calculations and have the approximate results checked against the exact ones. We show that the difference between the approximate and the exact results can be large.

Results: In this study, we have characterized quantitatively the effect of pooling samples on the efficiency of microarray experiments for the detection of differential gene expression between two classes. We present exact formulas for calculating the power of microarray experimental designs involving sample pooling and technical replications. The formulas can be used to determine the total number of arrays and biological subjects required in an experiment to achieve the desired power at a given significance level. The conditions under which pooled design becomes preferable to non-pooled design can then be derived given the unit cost associated with a microarray and that with a biological subject. This paper thus serves to provide guidance on sample pooling and cost-effectiveness. The formulation in this paper is outlined in the context of performing microarray comparative studies, but its applicability is not limited to microarray experiments. It is also applicable to a wide range of biomedical comparative studies where sample pooling may be involved.

Availability: A Java Webstart application can be accessed at http://wads.le.ac.uk/htox/WadsSoftware/MrcStats/SCal4Poolings.jnlp

Contact: sdz1@le.ac.uk; twg1@le.ac.uk

1 INTRODUCTION

Pooling samples in biomedical studies has now become a frequent practice among many researchers. For example, >15% of the datasets deposited in the Gene Expression Omnibus Database involve pooled RNA samples (Kendziorski *et al*., 2005). The practice of pooling biological samples though is not a new phenomenon, as it can be traced back at least to 1940s (Dorfman, 1943) and has been used in different application areas (Gastwirth, 2000), e.g. for the detection of certain medical conditions and estimation of prevalence in a population. In the context of detecting differential gene expressions using microarrays, divergent views on the wisdom of pooling samples can be found in the literature (Agrawal *et al*., 2002; Affymetrix, 2004; Shih *et al*., 2004; Churchill and Oliver, 2001; Peng *et al*., 2003; Jolly *et al*., 2005). One of the arguments supporting the practice of pooling biological samples is that biological variation can be reduced by pooling RNA samples in microarray experiments(Churchill and Oliver, 2001). As more carefully described by Kendziorski *et al*. (2005), pooling can reduce the effects of biological variation, but not the biological variation itself. Another argument in support of pooling samples in microarray experiments is that it reduces financial cost. However, cost reduction is meaningful only if statistical equivalence between the pooled and the non-pooled experimental setups is maintained. Here we address this issue and present formulas to determine the conditions under which pooled and non-pooled designs are statistically equivalent.

To compare experimental designs with and without sample pooling the two designs must have something in common that can be measured, e.g. using the same or equivalent amount of resources or yielding the same level of detection power. Kendziorski *et al*. (2003) used the width of the 95% confidence interval for gene expression to compare different experimental designs with and without sample pooling. The criterion was that the narrower the confidence interval, the more accurate the results from the experimental design. In a comparative study where two groups of biological subjects are compared the common goal of the different experimental designs is to detect a change between the two groups with a given power at a given false positive rate, as adopted in Shih *et al*. (2004). We shall use the latter method to compare different designs. So in this work statistical equivalence means that the designs have the same statistical power at the same level of significance. Therefore the more appropriate experimental design will be the one which uses less resources to achieve this statistical equivalence.

The basic assumption underlying sample pooling is biological averaging; that the measure of interest taken on the pool of samples is equal to the average of the same measure taken on each of the individual samples which contributed to the pool. For example in the situation of a microarray experiment, if r individual samples contribute equally to a pool, and the concentrations of a gene's mRNA transcripts for the r samples are denoted by T_i with $i = 1,2,\ldots,r$ indexing the individual samples, the assumption of biological averaging says that the concentration of this gene's mRNA transcripts in the pool is $T = (1/r)\sum_{i=1}^{r} T_i$. However, for microarray experiment there is some debate on whether the basic assumption of pooling holds. Kendziorski *et al*. (2003, 2005)

*To whom correspondence should be addressed.

argue that there is limited support for this assumption. Here we do not seek to enter into this debate but rather take the assumption of biological averaging as valid, or at least approximately so, so that we are in a position to determine whether pooling samples is financially beneficial or not. The validity of biological averaging makes it possible (or easier) to derive a neat theoretical formulation. On a practical level though, the requirement for the validity of this assumption may not be as stringent as a theoretical formulation does. For instance, Kendziorski *et al*. (2005) show that even when biological averaging does not hold, pooling can be useful and inferences regarding differential gene expression are not adversely affected by pooling.

One situation where there is little alternative but to pool biological samples is where there is insufficient amount of RNA from each individual biological subject to perform single microarray hybridization. RNA amplification may be a possible way of obtaining more RNA, but may not be practically feasible when many individual biological subjects are involved as in the case of Jin *et al*. (2001). In such a circumstance, pooling samples is justified by the lack of alternative and will not be considered further here. Similarly we will not consider here the case where all the biological samples of the same group were pooled together, and multiple technical replicate measurements were carried out on the sample pool. This is sometimes seen in the literature (Muckenthaler *et al*., 2003), but such an experimental design leaves no degree of freedom to estimate the biological variance. Thus valid inferences about the differences between the two populations of biological subjects under study cannot be made. Here we only consider situations other than the above two and where pooling may reduce the overall costs of the experiments.

2 A GENERAL FORMALISM

For every comparative study, there is at least one measurable quantity which is the quantity of interest. The goal of the study is to deduce from the data collected whether there is any difference between the means of the two populations. As measuring all the biological subjects in two populations is rarely possible in most situations representatives from a population are randomly selected and measurements are made on these. These are then taken to infer the properties of the population.

Let X be the measurable quantity that is being determined in the experiment, e.g. the expression level of a gene. In the case of one-channel microarray, X could denote the logarithm (most commonly base 2 is used) of fluorescence intensity or the logarithm of the fluorescence ratio in the case of two-channel microarray. Let x_i^{c}denote the value of X for an individual subject i in the control population (c), and x_j^{t} that of the individual subject j in the treatment population (t). We assume that x_i^{c}s for all individuals in the control population are independent normally distributed with a mean μ_{c} and a variance σ_{c}^2, denoted by $x_i^{\mathrm{c}} \sim N\left(\mu_{\mathrm{c}}, \sigma_{\mathrm{c}}^2\right)$ for all i. Similarly, $x_j^{\mathrm{t}} \sim N\left(\mu_{\mathrm{t}}, \sigma_{\mathrm{t}}^2\right)$ for all j.

2.1 A general experimental setup

For a general experimental setup individual subjects from both populations are randomly selected and tissue samples collected from each. Tissue sample pools are made by pooling a given number r of randomly selected tissue samples (of the same population) together. Note that to make n pools we need to have selected nr individual subjects from the population. m measurements are then made on each pool of tissue samples. So m is the number of technical replications of measurement on each pool. Notice that by introducing two parameters r and m a general and flexible experimental setup has been created. For instance, if we set $r = 1$, the experiment would be equivalent to no pooling of tissue samples. And if we set $m = 1$ there is no technical replication. Under the basic assumption of biological averaging, the result of pooling r tissue samples in equal proportions together is that the value of X for the pool is the average of those subjects which formed this pool,

$$\tilde{x} = \frac{1}{r} \sum_{i=1}^{r} x_i. \tag{1}$$

It follows that $\bar{x} \sim N\left(\mu_{\mathrm{c}}, \sigma_{\mathrm{c}}^2/r\right)$ for a pool from the control population, or $\bar{x} \sim N\left(\mu_{\mathrm{t}}, \sigma_{\mathrm{t}}^2/r\right)$ for a pool from the treated population. Note that in this paper we shall only discuss pooling samples with equal individual contributions. While pools formed by unequal contributions from individual samples are possible, such pooled experimental design is generally less effective than the equal pooling, as already shown by Peng *et al*. (2003) with their simulated results.

When we take a measurement on a pool p, the measured value is

$$y_{p,k} = \tilde{x}_p + \epsilon_k, \tag{2}$$

where p indexes pools, k indexes measurements and ϵ_k is a random error term assumed to be independently and normally distributed as $\epsilon_k \sim N\left(0, \sigma_\epsilon^2\right)$. Hereafter σ_ϵ^2 will be referred to as the technical variance, σ_c^2 the biological variance for the control population and σ_{t}^2 the biological variance for the treatment population.

The output of the experiment are the measurements on the two sets of pools. For the control group, we have $y_{p,k}^{\mathrm{c}}$ for $p = 1, \ldots, n_{\mathrm{c}}$ and $k = 1, \ldots, m$. And for the treatment group, we have $y_{p,k}^{\mathrm{t}}$ for $p = 1, \ldots, n_{\mathrm{t}}$ and $k = 1, \ldots, m$. Here n_{c} and n_{t} are the numbers of pools prepared for the control and treatment populations, respectively. Our task is to infer population properties from these measured data. In particular, we want to know whether there is any difference between the two population means μ_{c} and μ_{t}. It can be shown that

$$\bar{Y}^{\mathrm{c}} = \frac{1}{mn_{\mathrm{c}}} \sum_{p=1}^{n_{\mathrm{c}}} \sum_{k=1}^{m} y_{p,k}^{\mathrm{c}} \tag{3}$$

is an unbiased estimator of μ_{c}, with a variance

$$\frac{1}{n_{\mathrm{c}}} \left(\frac{\sigma_{\mathrm{c}}^2}{r} + \frac{\sigma_\epsilon^2}{m} \right), \tag{4}$$

and similarly,

$$\bar{Y}^{\mathrm{t}} = \frac{1}{mn_{\mathrm{t}}} \sum_{p=1}^{n_{\mathrm{t}}} \sum_{k=1}^{m} y_{p,k}^{\mathrm{t}} \tag{5}$$

is an unbiased estimator of μ_{t}, with a variance

$$\frac{1}{n_{\mathrm{t}}} \left(\frac{\sigma_{\mathrm{t}}^2}{r} + \frac{\sigma_\epsilon^2}{m} \right). \tag{6}$$

If we make an additional assumption that the variances for the two populations of biological subjects are the same, i.e. $\sigma_{\mathrm{c}}^2 = \sigma_{\mathrm{t}}^2 = \sigma^2$, then the difference between Equation (5) and (3), $D = \bar{Y}^{\mathrm{t}} - \bar{Y}^{\mathrm{c}}$, is an

unbiased estimator of $\mu = \mu_t - \mu_c$ with a variance

$$\sigma_D^2 = \left(\frac{1}{n_c} + \frac{1}{n_t}\right)\left(\frac{\sigma^2}{r} + \frac{\sigma_\epsilon^2}{m}\right). \tag{7}$$

The factor $\left(\sigma^2/r + \sigma_\epsilon^2/m\right)$ in Equation (7) can be estimated without bias by

$$\begin{aligned} s_p^2 = &\frac{1}{n_c + n_t - 2}\sum_{p=1}^{n_c}\left(\frac{1}{m}\sum_{k=1}^{m} y_{p,k}^c - \bar{Y}^c\right)^2 \\ &+ \frac{1}{n_c + n_t - 2}\sum_{p=1}^{n_t}\left(\frac{1}{m}\sum_{k=1}^{m} y_{p,k}^t - \bar{Y}^t\right)^2. \end{aligned} \tag{8}$$

It is then clear that

$$t = \frac{(\bar{Y}^t - \bar{Y}^c) - (\mu_t - \mu_c)}{s_p\sqrt{1/n_c + 1/n_t}} \tag{9}$$

follows the Student's t-distribution with $n_c + n_t - 2$ degrees of freedom. In detecting a differential gene expression, we want to test the null hypothesis $\mu_c = \mu_t$ against an alternative hypothesis $\mu_c \neq \mu_t$. So our test statistic is

$$t_0 = \frac{(\bar{Y}^t - \bar{Y}^c)}{s_p\sqrt{1/n_c + 1/n_t}}, \tag{10}$$

and there are no unknowns in Equation (10). Note that t_0 can be seen as a generalized two-sample t-test statistic, which reduces to the statistic of the traditional two-sample t-test with equal variance when we set the parameters $r = 1$ (no pooling of tissue samples) and $m = 1$ (no technical replication of measurements). Shih *et al*. (2004), arrived at two separate statistics, one for non-pooled design, the other for pooled design. The t_0 defined by Equation (10) is in more general form, setting $r = 1$ and $m = 1$ in Equation (10) recovers Shih *et al*.'s statistic for non-pooled design; while setting $r > 1$ and $m = 1$ recovers Shih *et al*.'s statistic for pooled design. Note that m does not need to equal 1. Here by incorporating two additional parameters r and m, the statistic t_0 can deal with situations where there are pooled tissue samples and multiple technical replications.

2.2 Criteria of significance

As with any statistical test we need to specify a threshold P-value P_{th} to claim significant results in the test. When all the other parameters are given, setting P_{th} is equivalent to setting a threshold, say $|\xi|$, for the statistics t_0 defined in Equation (10). With this threshold t-value, our criteria for claiming a significant test is as follows: If $t_0 > |\xi|$, we declare that $\mu_t - \mu_c > 0$; if $t_0 < -|\xi|$, it is claimed as $\mu_t - \mu_c < 0$. So the rate at which false positive claims are made is

$$\begin{aligned} P_{th} &= \int_{-\infty}^{-|\xi|} \rho_{n_c+n_t-2}(t_0)\,dt_0 + \int_{|\xi|}^{\infty} \rho_{n_c+n_t-2}(t_0)\,dt_0 \\ &= 2\int_{-\infty}^{-|\xi|} \rho_{n_c+n_t-2}(t_0)\,dt_0 = 2T_{n_c+n_t-2}(-|\xi|), \end{aligned} \tag{11}$$

where $\rho_{n_c+n_t-2}(.)$ is the probability density function (PDF) of the Student's t-distribution with $n_c + n_t - 2$ degrees of freedom, and $T_{n_c+n_t-2}(.)$ is the corresponding cumulative probability distribution function (CDF). It is therefore apparent that the threshold t-value $|\xi|$ can be obtained by solving the equation $2T_{n_c+n_t-2}(-|\xi|) = P_{th}$ with a given false positive rate P_{th}.

3 POWER FUNCTION

In Zhang and Gant (2004) we presented a power function for a new statistical t-test (hereafter referred to as 'two-labelling t-test') in the context of using two-colour microarrays to detect differential gene expression. Following similar steps we can derive the power function for the generalized two-sample t-test presented in this paper, which reads

$$S = \int_0^{\infty} p_{n_c+n_t-2}(Y)\Phi\left[\frac{-|\xi|\sqrt{Y}}{\sqrt{n_c + n_t - 2}} + \frac{|\mu|}{\sigma_D}\right]dY, \tag{12}$$

where $p_{n_c+n_t-2}(Y)$ is the PDF for the χ^2-distribution with $n_c + n_t - 2$ degrees of freedom and $\Phi(.)$ is the CDF for the standard normal distribution. The rate S at which a true difference between μ_t and μ_c can be successfully detected is a function of n_c, n_t, $|\mu|/\sigma_D$, and $|\xi|$. With σ_D given by the square root of Equation (7), and $|\xi|$ determined by solving Equation (11) at a given false positive rate P_{th}, S is, eventually, a function of P_{th}, n_c, n_t and $|\mu|/\sigma_D$.

A few points are worth noting here.

(1) The two-labelling t-test presented in Zhang and Gant (2004) was designed to deal with systematic labelling biases generated during microarray experimentation. The t-test presented in this paper, however, assumes no systematic data biases. In the case of two-colour microarrays this requires a common reference design. In such an experimental design the labelling biases cancel themselves out in the calculation of the test statistic.

(2) In Zhang and Gant (2004), the biological variances of the two populations under comparison do not have to be the same, i.e. we did not assume $\sigma_c^2 = \sigma_t^2$. For the t-test in this paper, we have made an additional assumption that $\sigma_c^2 = \sigma_t^2$. Relaxing this requirement was possible, as in the case of the traditional two-sample t-test with unequal variance (Brownlee, 1965), but an exact power function could not be readily obtained.

(3) The exact power function obtained in this paper allows evaluation of the effects of pooling biological samples and the effects of taking multiple technical measurements, thus giving researchers quantitative guidance on the practice of pooling samples.

(4) By setting the parameters $r = 1$ and $m = 1$, an exact power function is provided for the traditional two-sample t-test with equal variance.

4 RESULTS

We have implemented the computation of the power function S of Equation (12) as a Java application, which can be accessed at the URL given in the abstract. Here we apply this to microarray comparative studies for finding differentially expressed genes and investigate the effect of pooling RNA samples in the experiments. We also compare our exact results with some approximate results presented by other authors (Shih *et al*., 2004) to demonstrate why an exact formula is desirable.

Table 1. For the first scenario described in the text, the detection power of designs with different levels of pooling

n_c	n_t	S	r	N_s	N_m
5	5	0.8175	1	10	10
5	6	0.9026	1	11	11
6	5	0.9026	1	11	11
5	7	0.9488	1	12	12
7	5	0.9488	1	12	12
6	6	**0.9553**	1	**12**	**12**
6	7	0.9796	1	13	13
7	6	0.9796	1	13	13
3	3	0.3012	2	12	6
3	4	0.5555	2	14	7
4	3	0.5555	2	14	7
3	5	0.7602	2	16	8
5	3	0.7602	2	16	8
4	4	0.7937	2	16	8
4	5	0.9196	2	18	9
5	4	0.9196	2	18	9
5	5	**0.9771**	2	**20**	**10**
3	3	0.4060	3	18	6
3	4	0.6962	3	21	7
4	3	0.6962	3	21	7
3	5	0.8774	3	24	8
5	3	0.8774	3	24	8
4	4	0.9008	3	24	8
4	5	**0.9745**	3	**27**	**9**
5	4	0.9745	3	27	9
5	5	0.9957	3	30	10
2	2	0.0444	5	20	4
2	3	0.1930	5	25	5
3	2	0.1930	5	25	5
2	4	0.4732	5	30	6
4	2	0.4732	5	30	6
3	3	0.5324	5	30	6
3	4	0.8262	5	35	7
4	3	0.8262	5	35	7
4	4	**0.9657**	5	**40**	**8**
2	2	0.0643	18	72	4
2	3	0.2994	18	90	5
3	2	0.2994	18	90	5
2	4	0.6718	18	108	6
4	2	0.6718	18	108	6
3	3	0.7309	18	108	6
3	4	**0.9515**	18	**126**	**7**
4	3	0.9515	18	126	7
4	4	0.9969	18	144	8

The $r = 1$ panel represents that of non-pooled design. Other parameter values are $\sigma^2 = 0.05$, $\sigma_\epsilon^2 = 0.0125$, $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$ and $m = 1$. $N_s = r(n_c + n_t)$ is the total number of biological subjects required, and $N_m = m(n_c + n_t)$ is the total number of measurements (microarrays) needed, counting both the control and the treatment populations. The preset targets are false positive rate being controlled at $P_{th} = 0.001$, to detect 2-fold differential expression ($\mu = 1$) with power no less than 0.95. The minimum number of biological subjects (N_s) and microarrays (N_m) that meet the preset targets are highlighted with bold fonts.

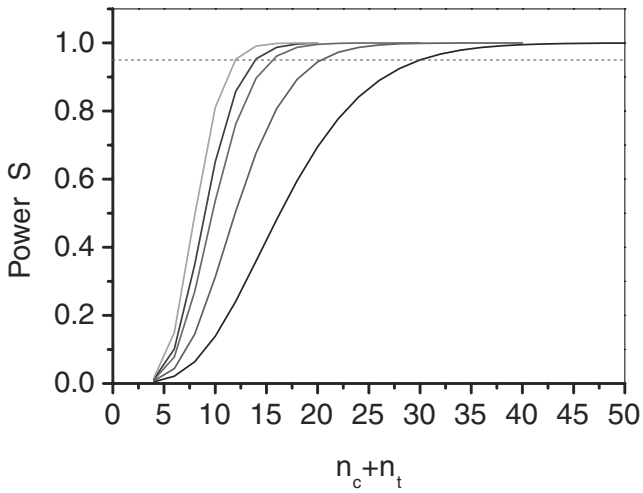

Fig. 1. The power S as a function of the total number of pools $n_c + n_t$. The parameters used are for the second scenario $\sigma^2 = 0.2$, $\sigma_\epsilon^2 = 0.05$, $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$, $P_{th} = 0.001$, $\mu = 1$ and $m = 1$. The five solid curves correspond to different levels of pooling, from right to left, $r = 1$, $r = 2$, $r = 4$, $r = 6$ and $r = 15$, respectively. The dashed line indicates the 95% power, the intersections of which with the power curves specify the total numbers of pools (assuming $n_c = n_t$) needed to achieve the target power. The total number of biological subjects and the total number of arrays can then be calculated simply by $N_s = r(n_c + n_t)$, and $N_m = m(n_c + n_t)$, respectively.

4.1 Comparison with approximate results

Based on their approximate formulas, Shih *et al*. (2004) considered two scenarios to compare the number of biological subjects and number of microarrays in the non-pooled and pooled designs. Here we give exact results for the two scenarios to show the difference to the approximate results. In the first scenario, we consider that the common biological variance of the two populations is $\sigma^2 = 0.05$, and the technical variance $\sigma_\epsilon^2 = 0.0125$, which gives the biological-to-technical variance ratio $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$. The preset target of the experiment in this scenario is that the false positive rate being controlled at $P_{th} = 0.001$ and the power being no less than $S = 0.95$ to detect a 2-fold differential gene expression, which corresponds to $\mu = 1$ with base 2 logarithm (Shih *et al*., 2004). In Table 1, we present results for different pooling parameter r. It can be seen from the first panel of this table that in order to hit the preset target, the non-pooled design ($r = 1$) requires at least 12 biological subjects divided evenly to the two populations, i.e. 6 from each of the two populations. Having seven subjects from one population and five subjects from the other is insufficient to achieve the target of 95% detection power. The effects of other levels of pooling on the detection power are also shown in Table 1. The minimum number of biological subjects (N_s) and microarrays (N_m) that meet the preset targets is highlighted with bold fonts. It is clear that as the level of pooling is increased (with increasing r), the number of microarrays N_m can be reduced, but the number of biological subjects N_s has to be increased. For example, in order to reduce the number of arrays from 12 (Table 1, first panel) to 8 (Table 1, fourth panel), the number of biological subjects to form the pools must be increased from 12 to 40.

For the second scenario we consider the case $\sigma^2 = 0.2$, $\sigma_\epsilon^2 = 0.05$, which gives $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$. Again the preset targets are to detect a true differential expression $\mu = 1$ with no less than 95% power while the false positive rate is set at $P_{th} = 0.001$. Using these parameters, the power S as a function of $n_c + n_t$ is plotted in Figure 1 for different levels of sample pooling. For the

Table 2. Comparison of our exact results and the approximate results of Shih *et al*. (2004)

N_m	N_s (Exact)	N_s (Approx)	Conditions
11		11	
12	12		
10	20	13	$C_m \geq 4C_s$
9	27	16	$C_m \geq 7C_s$
8	40	21	$C_m \geq 13C_s$
7	126	37	$C_m \geq 86C_s$
30	30		
21	42		$C_m \geq 1.33C_s$
22		35	
16	64	50	$C_m \geq 4.4C_s$
14	84	64	$C_m \geq 10C_s$
12	180	104	$C_m \geq 48C_s$

The upper panel of the table is for the first scenario, where $\sigma^2 = 0.05$, $\sigma_\epsilon^2 = 0.0125$, $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$. The lower panel is for the second scenario, where $\sigma^2 = 0.2$, $\sigma_\epsilon^2 = 0.05$, and $\lambda = \sigma^2/\sigma_\epsilon^2 = 4$. The targets of both scenarios are that the false positive rate $P_{th} = 0.001$ and the power no less than $S = 0.95$. The last column in each panel gives the cost conditions when pooling samples become beneficial relative to a lower level of pooling shown in this table.

non-pooled design ($r = 1$), $N_s = 30$ total biological subjects and $N_m = 30$ arrays are required to hit the preset targets. Similar to the first scenario, as the level of pooling is increased, the number of arrays N_m is reduced while the number of subjects increased to meet the preset targets.

In Table 2, we summarize our exact results and the approximate results of Shih *et al*. (2004). It can be seen that the difference between the two can be very large, indicating the need for exact results. For example, in the first scenario when $N_m = 8$ the approximate result of Shih *et al*. (2004) predicts that a minimum of 21 biological subjects are required. In practice 24 subjects are required as 24 is the minimum number >21 and divisible by 8. However this experiment setup (24 subjects forming 8 pools, 8 microarrays) will only give a detection power of 90%. To meet the target power of 95%, 40 biological subjects are actually required by our exact result. If an experiment with $N_m = 7$ microarrays is planned, Shih *et al*. predicts that 37 subjects are required, but in fact 126 subjects must be used to achieve the target. Generally, the approximate formulas of Shih *et al*. (2004) are too optimistic in assessing the benefits of pooling samples and reducing the number of microarrays, because they underestimate the number of biological subjects required.

4.2 Cost analysis

Depending on the material costs involved in the biological subjects and microarrays, the conditions where pooling samples becomes beneficial may be different from laboratory to laboratory. Here we show examples to determine these conditions. Denoting the cost associated with each biological subject as C_s (including materials and labour, etc.) and the cost associated with a microarray as C_m, the total costs for an experiment in microarray comparative study is $C_T = N_sC_s + N_mC_m$. Taking the first scenario as an example, the total cost of a non-pooled design to achieve our preset targets is

$$C_T(r = 1) = 12C_s + 12C_m,$$

and the total cost for pooled design with $r = 2$ is

$$C_T(r = 2) = 20C_s + 10C_m.$$

Therefore in order that the pooled design with $r = 2$ is beneficial we must have

$$C_T(r = 2) \leq C_T(r = 1), \tag{13}$$

which requires that $C_m \geq 4C_s$. Put another way, only when the cost associated with one microarray C_m is more than four times the cost of a subject C_s, does the pooling design with $r = 2$ become preferable to the non-pooled design. Similarly a higher level of pooling with $r = 3$ becomes preferable to $r = 2$ only when $C_m \geq 7C_s$. Furthermore the conditions for increasing the level of pooling from $r = 3$ to $r = 5$ are $C_m \geq 13C_s$, and so on. Table 2 gives these conditions for further levels of pooling.

For the first scenario using the actual cost figures given in Shih *et al*. (2004) where $C_s = \$230$ and $C_m = \$300$, it can be seen that none of the pooling conditions is met. Therefore for this laboratory pooling samples is not recommended. However, if we use the cost figures of Kendziorski *et al*. (2003) where $C_s = \$50$ and $C_m = \$700$, an optimal design is a pooled design with $r = 5$.

For the second scenario, it is a similar story. The cost figures of Shih *et al*. (2004) ($C_s = \$230$ and $C_m = \$300$) give $C_m = 1.30C_s$, which does not satisfy any of the pooling conditions. So again the non-pooled design with $N_m = 30$ and $N_s = 30$ is recommended. On the other hand, the cost figures of Kendziorski *et al*. (2003) ($C_s = \$50$ and $C_m = \$700$) give $C_m = 14C_s$ which satisfies all the pooling conditions in the lower panel of Table 2 except the last row. So in Kendziorski *et al*.'s laboratory the pooled design with $N_m = 14$ and $N_s = 84$ would be recommended.

5 DISCUSSION

We have in this paper presented exact formulas for calculating the power of microarray experimental design with different levels of pooling. These formulas can be used to determine the conditions of statistical equivalence between different pooling setups. As in Kendziorski *et al*. (2003) and Shih *et al*. (2004), the calculations presented in this paper are for an individual gene, so the statistical equivalence for different designs of pooling can be determined with regard to one particular gene. However, microarray monitors thousands of genes simultaneously, and the biological and technical variances vary from gene to gene, therefore no single result of statistical equivalence between pooled and non-pooled designs applies equally to all genes on the array. So in practice how would the formulations in this work be used? One possible way, as suggested by Kendziorski *et al*. (2003), is to specify the distributions of σ^2 and σ_ϵ and calculate the total number of subjects and arrays that maximize the average power across the array. In theory, if the biological variances and technical variances were known for all genes on the array, an equivalence condition between pooled and non-pooled designs could be determined for each gene individually. The overall (or say, average) equivalence condition between pooled and non-pooled designs could be obtained, e.g., by some form of averaging operation over all genes. An alternative and probably a more practical way is to use representative values of σ^2 and σ_ϵ. We therefore propose that parameters for 'typical gene' be used as inputs for the power and sample size calculations. A typical gene is a gene whose biological and technical variances have the most

probable values among the genes, i.e. the mode of the distribution for biological and technical variances of genes. Alternatively, the median or mean variances across genes could be used as representative values Shih *et al*. (2004).

An issue associated with microarray experiments is the problem of multiple inferences, where a separate null hypothesis is being tested for each gene. Given thousands of null hypotheses being tested simultaneously, the customary significance level $\alpha = 0.05$ for declaring positive tests will surely give too many false positives. For example, if among a total number $N = 10\,000$ of genes being tested, $N_0 = 4000$ are truly null genes (genes that are non-differentially expressed between the two classes), the expected number of false positive results would be $4000 \times 0.05 = 200$, which may be too many to be acceptable. Thus a smaller threshold *P*-value for declaring differentially expressed genes should be used. Effectively controlling false positives in a multiple-testing situation such as microarray experiments is an area which has drawn much attention in recent years due to the wider application of microarray technology. As discussed in our previous work in Zhang and Gant (2004), generally speaking, all different multiple-testing adjustment methods eventually amount to effectively setting a threshold *P*-value and then rejecting all the null hypotheses with *P*-value below this threshold. The classical Bonferroni multiple-testing procedure, which controls family-wise error rate at α by setting the threshold $P_{\text{th}} = \alpha/N$, is generally regarded as being too conservative in the microarray context. The FDR (false discovery rate) idea, initially due to Benjamini and Hochberg (1995) in dealing with the multiple-testing problem, has now been widely accepted as appropriate to the microarray situation. Recently, Efron (2004) extended the FDR idea by defining FDR, a local version FDR. When planning microarray experiments in terms of power and sample size calculation, the FDR of Benjamini and Hochberg (1995) is more appropriate and convenient to use. There are now in literature a few slightly different variants of the definition of FDR (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003; Grant *et al*., 2005), but in essence it is defined as the proportion of false positives among all positive tests declared. To provide an interface between FDR and the formulation in the previous sections, here we show that there is a simple correspondence between controlling FDR and specifying the traditional type I error rate and power. Suppose that there are a total number N of genes being monitored by microarray, so there will be N hypotheses being tested, one for each gene. Suppose that a fraction π_0 of the N genes are true null genes, i.e. genes that are non-differentially expressed between the two classes. Given the type I error rate P_{th}, the expected number of false positive tests is $P_{\text{th}}N\pi_0$; Given the power S, the expected number of non-null genes (truly differentially expressed genes) that are declared positive is $SN(1 - \pi_0)$. So the FDR achieved by this setting is

$$\text{FDR} = \frac{P_{\text{th}}N\pi_0}{P_{\text{th}}N\pi_0 + SN(1-\pi_0)} = \frac{P_{\text{th}}\pi_0}{P_{\text{th}}\pi_0 + S(1-\pi_0)}. \qquad (14)$$

Here π_0 is an important parameter in controlling FDR, for which several different methods of estimating this parameter have been proposed (Pounds and Morris, 2003; Storey and Tibshirani, 2003; Zhang and Gant, 2004). Especially the method we presented in Zhang and Gant (2004) is an accurate yet computationally much simpler algorithm than the one proposed by Storey and Tibshirani (2003). With the interface Equation (14), FDR can be readily presented and incorporated into the calculations.

ACKNOWLEDGEMENTS

We wish to acknowledge the support of the microarray team of the MRC Toxicology Unit particularly Reginald Davies, JinLi Luo and Joan Riley. We also thank the two anonymous reviewers for their helpful and constructive comments.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2004), Sample pooling for microarray analysis. *Technical note*, Affymetrix, San Diego.

Agrawal,D. *et al*. (2002) Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J. Natl Cancer Inst*., **94**, 513–521.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Brownlee,K.A. (1965) *Statistical Theory and Methodology in Science and Engineering*. John Wiley and Sons, Inc., New York.

Churchill,G.A. and Oliver,B. (2001) Sex, flies and microarrays. *Nat. Genet*., **29**, 355–356.

Dorfman,R. (1943) The detection of defective members of large populations. *Ann. Math. Stat*., **14**, 436–440.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc*., **99**, 96–104.

Gastwirth,J.L. (2000) The efficiency of pooling in the detection of rare mutations. *Am. J. Hum. Genet*., **67**, 1036–1039.

Grant,G.R. *et al*. (2005) A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics*, **21**, 2684–2690.

Jin,W. *et al*. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet*., **29**, 389–395.

Jolly,R.A. *et al*. (2005) Pooling samples within microarray studies: a comparative analysis of rat liver transcription response to prototypical toxicants. *Physiol Genomics*, **22**, 346–355.

Kendziorski,C. *et al*. (2005) On the utility of pooling biological samples in microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 4252–4257.

Kendziorski,C.M. *et al*. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**, 465–477.

Muckenthaler,M. *et al*. (2003) Regulatory defects in liver and intestine implicate abnormal hepcidin and Cybrd1 expression in mouse hemochromatosis. *Nat. Geneti*., **34**, 102–107.

Peng,X. *et al*. (2003) Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics*, **4**, 26.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics*, **19**, 1236–1242.

Shih,J.H. *et al*. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*, **20**, 3318–3325.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.

Zhang,S.-D. and Gant,T.W. (2004) A statistical framework for the design of microarray experiments and effective detection of differential gene expression. *Bioinformatics*, **20**, 2821–2828.