BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 10 2007, pages 1217–1224
doi:10.1093/bioinformatics/btm081

Gene expression

Pooling mRNA in microarray experiments and its effect on power

Wuyan Zhang[1], Alicia Carriquiry[1,3,*], Dan Nettleton[1,3] and Jack C.M. Dekkers[2,3]

[1]Department of Statistics, [2]Department of Animal Science and [3]Center for Integrated Animal Genomics, Iowa State University, USA

Received on October 2, 2006; revised on February 26, 2007; accepted on February 28, 2007

Advance Access publication March 7, 2007

Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: Microarrays can simultaneously measure the expression levels of many genes and are widely applied to study complex biological problems at the genetic level. To contain costs, instead of obtaining a microarray on each individual, mRNA from several subjects can be first pooled and then measured with a single array. mRNA pooling is also necessary when there is not enough mRNA from each subject. Several studies have investigated the impact of pooling mRNA on inferences about gene expression, but have typically modeled the process of pooling as if it occurred in some transformed scale. This assumption is unrealistic.

Results: We propose modeling the gene expression levels in a pool as a weighted average of mRNA expression of all individuals in the pool on the original measurement scale, where the weights correspond to individual sample contributions to the pool. Based on these improved statistical models, we develop the appropriate F statistics to test for differentially expressed genes. We present formulae to calculate the power of various statistical tests under different strategies for pooling mRNA and compare resulting power estimates to those that would be obtained by following the approach proposed by Kendziorski *et al*. (2003). We find that the Kendziorski estimate tends to exceed true power and that the estimate we propose, while somewhat conservative, is less biased. We argue that it is possible to design a study that includes mRNA pooling at a significantly reduced cost but with little loss of information.

Contact: alicia@iastate.edu

1 INTRODUCTION

Microarray experiments are widely used to measure the expression levels of tens of thousands of genes simultaneously under different experimental conditions or during different time periods. One of the major interests in microarray experiments is to identify genes which are differentially expressed between conditions or time periods, and enable a deeper knowledge of complex biological problems at the genetic level. However, the unit cost of microarrays continues to be high; even for a moderate number of subjects, cost can be significant. One option to control costs is to pool the mRNA of a group of individuals and then run microarrays on the pools rather than on each individual. Pooling mRNA may also be required when there is not enough mRNA from each subject to hybridize individual microarrays.

The effect and efficiency of pooling mRNA in microarray experiments have been investigated by several researchers. Kendziorski *et al.* (2003) showed that pooling is most advantageous when biological variability (variability across subjects) in expression level is larger than technical variability (variability introduced in the experimental process). They also derived a formula for the total number of arrays and individuals required in an experiment involving mRNA pools to obtain gene expression estimates and confidence intervals comparable to those that would be obtained when analyzing individual arrays. They concluded that by increasing the total number of individuals in the experiment, it was possible to maintain precision of estimates by pooling, while decreasing the total number of arrays. Shih *et al.* (2004) also discussed the impact of pooling mRNA on the power of statistical tests. They derived an expression to carry out power calculations for a given number of arrays and individuals. Further, they used expression data obtained from mice to check the adequacy of the assumption that mRNA expression levels in the pool are close to the average expression levels of individuals in the pool. They showed that the assumption does not hold, especially when the signals are high.

Both Kendziorski *et al.* (2003) and Shih *et al.* (2004) derived their results on the transformed scale, that is, after the data were normalized and signal intensity was transformed. Thus, both studies assumed that the mRNA expression in the pool is the average expression of individual samples, and applied the assumption on the transformed scale. This assumption is, however, not realistic in a biological sense because in the laboratory, the mRNA is extracted from samples and then mixed to form an mRNA pool. Therefore, pooling occurs on the original scale and an assumption that holds on the transformed scale may not hold before transformation.

In this article, we address the issue of testing for differences in gene expression across treatments. We assume that the expression level in a pool is approximately equal to the average expression of individuals in the pool on the original scale. More precisely, we assume that the expression level in a pool is a weighted average expression of individual samples on the original scale, where weights correspond to the proportional contribution of each individual to the pool. By including the weights in the average, we account for the fact that in the

[*] To whom correspondence should be addressed.

process of combining individual mRNA samples, the mixing proportions may not be identical and thus, that the pool may contain more mRNA from some individuals than from others.

Under the assumptions above, we use a single gene as an example and derive expressions to calculate power under different treatment effect sizes, number of mRNA pools, number of individuals per mRNA pool and number of repeated measurements per pool. We wish to understand how much power is lost by pooling mRNA. We also wish to find efficient experimental designs for pooling mRNA samples, while keeping costs down.

2 SYSTEMS AND METHODS

2.1 Notation and model

Microarray gene expression measurements tend to be right skewed and thus not normally distributed. Therefore, data are usually transformed and normalized before statistical analysis. The most common transformation is the log transformation (Geller *et al.*, 2003). The transformed data can then be modeled as a linear function of treatment effects and one or more normally distributed random effects (Han *et al.*, 2004; Lu 2004; Shih *et al.*, 2004). The sources of variation in a microarray experiment are multiple and can be generally classified into two groups: biological variation and technical variation (Chen *et al.*, 2004; Kendziorski *et al.*, 2003). Biological variation is subject-to-subject variation in gene expression and is due to subject-specific genetic or environmental factors. Technical variation arises from the errors that can potentially be introduced at each of multiple steps in a microarray experiment. These include RNA sample preparation, microarray construction, hybridization and washing procedures, and signal detection methods. Here, we focus on the two major categories: biological variation and technical variation.

The expression levels of tens of thousands of genes are measured simultaneously in a microarray experiment. For simplicity of notation, a single gene is considered in the following derivation and analysis. The true gene expression level of a gene on the *j*th individual in the *i*th treatment is denoted m_{ij} and can be modeled on the log scale as

$$\log(m_{ij}) = \mu_i + \epsilon_{ij}, \tag{1}$$

$i = 1, 2, \ldots, T$, $j = 1, 2, \ldots, N$. Here, T is the number of treatments (or conditions), N is the number of individuals per treatment, μ_i is the mean gene expression level for the *i*th treatment and ϵ_{ij} is biological error which is assumed to be independently, identically distributed as $N(0, \sigma_{\mathrm{b}}^2)$. We use σ_{b}^2 to denote the biological variance in gene expression. Then, the observed gene expression level o_{ij} in log scale can be modeled as

$$\log(o_{ij}) = \mu_i + \epsilon_{ij} + \xi_{ij} = \log(m_{ij}) + \xi_{ij}, \tag{2}$$

where ξ_{ij} is technical error which is also assumed to be independently, identically distributed as $N(0, \sigma_t^2)$. Biological and technical errors are assumed to be independent.

Suppose now that the mRNA from n subjects from the same treatment is combined to form a pool. We use m_{ij}^{p} to denote true expression for a gene in the *i*th treatment group and *j*th pool. We assume that the true expression level of a gene in the mRNA pool is a weighted average of the true expression levels of the gene in all individuals in the same mRNA pool (denoted $m_{ij1}, \ldots, m_{ijn}$) so that

$$m_{ij}^{\mathrm{p}} = \sum_{k=1}^{n} (w_{ijk} * m_{ijk}), \tag{3}$$

where

$$w_{ijk} = \frac{z_{ijk}}{z_{ij1} + z_{ij2} + \cdots + z_{ijn}},$$

$i = 1, 2, \ldots, T, j = 1, 2, \ldots, P, k = 1, 2, \ldots, n$. Here, P is the number of mRNA pools per treatment and n is the number of individuals per mRNA pool. Therefore, $P * n = N$, where N denotes the total number of individuals per treatment group in the experiment. When $n = 1$, the experiment involves no mRNA pooling (a microarray is made for each individual). The random unobservable weight w_{ijk} represents the relative contribution of individual k to pool j in treatment i, and the expectation of each weight is $\frac{1}{n}$. We write the weights as functions of the z_{ijk}, which denote the technical deviations from the ideal pool containing equal amounts of mRNA from each individual sample. We assume that the z_{ijk} are independently, identically distributed as $N(1, \sigma_{\mathrm{z}}^2)$, where σ_{z}^2 denotes the pooling technical variance which is assumed to be no larger than 0.2^2 so that the probability of negative weights will be negligible. If we denote the observed mRNA level in a pool by o_{ijl}^{p}, we can then model it on the log scale as:

$$\log(o_{ijl}^{\mathrm{p}}) = \log(m_{ij}^{\mathrm{p}}) + \xi_{ijl}, \tag{4}$$

where, ξ_{ijl} is technical error as defined earlier and $l = 1, 2, \ldots R$. Here, R is the number of replicated array measurements per pool and $R * P = A$, where A is the total number of arrays per treatment group. $P = A$ if each mRNA pool is measured only once. Note that model (4) for the transformed observed mRNA level in a pool is similar to model (2) formulated for observed mRNA in an individual array on the transformed scale.

2.2 Expectation and variance of $\log(m_{ij}^{\mathrm{p}})$

The distribution of $\log(m_{ij}^{\mathrm{p}})$ is analytically intractable, but simulations and goodness-of-fit testing show that it can be adequately approximated by a normal distribution (see additional discussion of this point in Section 4). We will use μ_i^{p} and $\sigma_{\mathrm{b}}^{\mathrm{p}2}$ to denote the mean and variance of this normal distribution. We can then write

$$\log(m_{ij}^{\mathrm{p}}) \sim \mu_i^{\mathrm{p}} + \tau_{ij} \tag{5}$$

and

$$\log(o_{ijl}^{\mathrm{p}}) \sim \mu_i^{\mathrm{p}} + \tau_{ij} + \xi_{ijl}, \tag{6}$$

where, τ_{ij} is assumed to be independent and identically distributed $N(0, \sigma_{\mathrm{b}}^{\mathrm{p}2})$.

We are interested in testing for differences in gene expression across treatments of the form $\mu_i - \mu_j$. In this subsection, we will derive expressions for μ_i^{p} and σ_{b}^{p2} and show that $\mu_i - \mu_j = \mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}}$, so that the tests of interest can be conducted using data from pools.

To derive an expression for $\mu_i^{\mathrm{p}} = E[\log(m_{ij}^{\mathrm{p}})]$, we expand $\log(m_{ij}^{\mathrm{p}})$ using a Taylor series to obtain

$$\log(m_{ij}^{\mathrm{p}}) = \log(\nu_i^{\mathrm{p}}) + \sum_{k=1}^{\infty} \frac{(-1)^{k-1}(m_{ij}^{\mathrm{p}} - \nu_i^{\mathrm{p}})^k}{k(\nu_i^{\mathrm{p}})^k}, \tag{7}$$

where $\nu_i^{\mathrm{p}} = E(m_{ij}^{\mathrm{p}})$. In the Appendix, we show that

$$\nu_i^{\mathrm{p}} = E(m_{ij}^{\mathrm{p}}) = e^{\mu_i + \sigma_{\mathrm{b}}^2/2} \tag{8}$$

and

$$\sigma_i^{\mathrm{p}2} = \mathrm{Var}(m_{ij}^{\mathrm{p}}) = \frac{1}{n}(e^{2\mu_i + 2\sigma_{\mathrm{b}}^2} - e^{2\mu_i + \sigma_{\mathrm{b}}^2})(1 + n^2\sigma_w^2), \tag{9}$$

where σ_w^2 is the variance of each weight w_{ijk}. Thus, using (7), (8) and (9); a second order approximation to μ_i^{p} is given by

$$\mu_i^{\mathrm{p}} = E[\log(m_{ij}^{\mathrm{p}})] \sim \mu_i + \frac{\sigma_{\mathrm{b}}^2}{2} - \frac{1}{2n}(e^{\sigma_{\mathrm{b}}^2} - 1)(1 + n^2\sigma_w^2). \tag{10}$$

Also note that, from (9), the kth term of the summation in (7) may be written as $(-1)^{k-1}(m_{ij}^{\mathrm{p}}e^{-\mu_i-\sigma_{\mathrm{b}}^2/2}-1)^k/k$, where

$$m_{ij}^{\mathrm{p}}e^{-\mu_i-\sigma_{\mathrm{b}}^2/2}=\sum_{k'=1}^{n}w_{ijk'}m_{ijk'}e^{-\mu_i-\sigma_{\mathrm{b}}^2/2}$$

by definition of m_{ij}^{p}. It is easy to show that $m_{ijk'}e^{-\mu_i-\sigma_{\mathrm{b}}^2/2}$ has a log normal distribution that depends only on σ_{b}^2 and is free of μ_i. Thus, for any two treatments i and j,

$$\mu_i-\mu_j=\log(\nu_i^{\mathrm{p}})-\log(\nu_j^{\mathrm{p}})=\mu_i^{\mathrm{p}}-\mu_j^{\mathrm{p}} \tag{11}$$

using (7) and (8).

To obtain an approximation for σ_b^{p2}, we use the Delta method to obtain

$$\begin{aligned}\sigma_{\mathrm{b}}^{\mathrm{p}2}&=\mathrm{Var}[\log(m_{ij}^{\mathrm{p}})]\sim\mathrm{Var}(m_{ij}^{\mathrm{p}})/\{E(m_{ij}^{\mathrm{p}})\}^2\\&=\frac{1}{n}(e^{\sigma_{\mathrm{b}}^2}-1)(1+n\sigma_w^2),\end{aligned} \tag{12}$$

where the last equality follows from (8) and (9). In the Appendix, we show that $\sigma_w^2\sim\frac{n-1}{n^3}\sigma_z^2$. Combining with (12), we have

$$\sigma_{\mathrm{b}}^{\mathrm{p}2}=\mathrm{Var}[\log(m_{ij}^{\mathrm{p}})]\ \sim\frac{1}{n}(e^{\sigma_{\mathrm{b}}^2}-1)\left(1+\frac{n-1}{n^2}\sigma_{\mathrm{z}}^2\right). \tag{13}$$

We now note that the random effect τ_{ij} in model (6) is an error term attributable to biological variation in expression level and to the additional variability that is introduced when pooling mRNA samples.

2.3 Power in a design that includes pooling mRNA

One interesting finding is that $\mu_i^{\mathrm{p}}-\mu_j^{\mathrm{p}}=\mu_i-\mu_j$ (see Expression 11). Therefore, the hypothesis for testing $\mu_1^{\mathrm{p}}=\mu_2^{\mathrm{p}}=\cdots=\mu_T^{\mathrm{p}}$ in the design that includes pooling is equivalent to the hypothesis for testing $\mu_1=\mu_2=\cdots=\mu_T$ in a design that involves individual microarrays. The corresponding F-test statistic for the design with pooling is given by

$$F=\frac{T*(P-1)\sum_{i=1}^{T}P(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2}{(T-1)\sum_{i=1}^{T}\sum_{j=1}^{P}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot})^2}, \tag{14}$$

where $y_{ijl}=\log(o_{ijl})$.

The null hypothesis of no treatment differences is rejected at significance level α if the F statistic is larger than $F_{T-1,\,T*(P-1),\,\alpha}$, where $F_{df1,\,df2,\,\alpha}$ is the $(1-\alpha)*100$th percentile of a central F distribution with $df1, df2$ degrees of freedom.

If the type I error is controlled at level α, power of the test is given by

$$1-\beta=Pr(F_{T-1,\,T*(P-1)}(\delta^2)>F_{T-1,\,T*(P-1),\,\alpha}), \tag{15}$$

with non-centrality parameter δ^2, where

$$\delta^2=\frac{P\sum_{i=1}^{T}(\mu_i^{\mathrm{p}}-\overline{\mu^{\mathrm{p}}})^2}{\sigma_b^{\mathrm{p}2}+\frac{1}{R}\sigma_t^2}\sim\frac{P\sum_{i=1}^{T}(\mu_i^{\mathrm{p}}-\overline{\mu^{\mathrm{p}}})^2}{\frac{1}{n}(e^{\sigma_b^2}-1)\left(1+\frac{n-1}{n^2}\sigma_z^2\right)+\frac{\sigma_t^2}{R}}, \tag{16}$$

with P and R as defined earlier and $\overline{\mu^{\mathrm{p}}}$ equal to the mean of μ_i^{p} across treatments.

For a more general test of hypothesis for a linear combination of the means $C\mu^{\mathrm{p}}=d$, where $\mu^{\mathrm{p}}=(\mu_1^{\mathrm{p}},\mu_2^{\mathrm{p}},\ \ldots,\mu_T^{\mathrm{p}})^t$ with C a known matrix of constants with full rank r, power is calculated as

$$1-\beta=p(F_{r,\,T*(P-1)}(\delta^2)>F_{r,\,T*(P-1),\,\alpha}), \tag{17}$$

where the non-centrality parameter is given by

$$\delta^2=\frac{P}{\sigma_{\mathrm{b}}^{\mathrm{p}2}+\frac{1}{R}\sigma_t^2}(C\mu^{\mathrm{p}}-d)^t(CC^t)^{-1}(C\mu^{\mathrm{p}}-d). \tag{18}$$

As in (16), we can approximate this non-centrality parameter by replacing $\sigma_{\mathrm{b}}^{\mathrm{p}2}$ with its approximation in (13).

For notational simplicity, we have assumed balance in the number of pools per treatment (P), the number of individuals per pool (n) and the number of arrays per pool (R). The F statistic and power formulas presented in subsection can be generalized in a straightforward manner to account for differing numbers of pools per treatment. Differing numbers of individuals per pool and/or differing numbers of arrays per pool, however, create problems with the standard assumption of homogeneity of variance. We have shown that the variance of $\bar{y}_{ij\cdot}$ is non-trivial function of the number of individuals per pool (n), the biological variance (σ_{b}^2), variance associated with varying contributions of samples to each pool (σ_z^2), additional technical variance (σ_{t}^2), and the number of arrays per pool (R). If n and/or R are not constant for all pools, the variance of $\bar{y}_{ij\cdot}$ will not be the same for all values of i and j, and the tests and power calculations presented in this subsection will not be valid. Furthermore, if n and/or R vary across pools, the resulting heterogeneity of variance cannot be easily dealt with by weighting the observations because optimal weights will depend on unknown variance components σ_{b}^2, σ_{t}^2, and σ_z^2. Thus, varying numbers of individuals per pool and/or varying numbers of arrays per pool should be avoided.

3 DISCUSSION

3.1 Comparing estimates of power

The power approximations presented in the previous section are based on an assumption of normality and Delta-method approximations. To estimate the impact of these approximations, we simulated data and calculated power numerically and using the analytical expressions derived in Section 2.3. We also compared power from simulation with power calculated analytically under the Kendziorski model (Kendziorski *et al.*, 2003).

We simulated individual data under two different scenarios. For the first scenario, we fixed the number of treatment groups at three ($T=3$) and the number of individuals per treatment group at 100 ($N=100$). The mean expression difference between adjacent treatment groups on the log scale ($\mu_1-\mu_2=\mu_2-\mu_3$) was assumed to be 0.2, 0.3, 0.4 or 0.5. Biological and technical variances were fixed at 0.75 and 0.25 ($\sigma_{\mathrm{b}}^2=0.75,\sigma_{\mathrm{t}}^2=0.25$), respectively. Pooled data under our model were simulated as a weighted average of five or three individuals (weight variation was $\sigma_{\mathrm{z}}^2=0.05^2$) on the original scale. Therefore, $n=5$ and we considered 20 pools per treatment ($P=20$). For the second scenario, we simulated less individuals and less pools with $N=15$, $n=3$ and $P=5$ while keeping all the other parameters the same. For each scenario, a one-way ANOVA model was fitted to the simulated pooled data to test whether the mean expression level was different across treatment groups. We compared the power of the tests at $\alpha=0.05$. Results are presented in Table 1. Power calculated by simulation was based on 10 000 replicates of the experiment. The entries in the column labeled 'Analytical power' were calculated under two different models: the proposed model (Expression 6) and the Kendziorski model (Kendziorski *et al.*, 2003).

In both scenarios, the predicted power as computed using the approach proposed by Kendziorski *et al.* (2003) appears to be overly optimistic in that it consistently exceeds power calculated from simulation. This may be because their approach does not account for the additional variance introduced in the pooling step and because they assume that

Table 1. Power of the test for treatment difference computed numerically by simulation and analytically by the proposed model and the Kendziorski model

Mean expression difference	Power calculated by simulation	Analytical power	
		Proposed model	Kendziorski model
$N=100$, $n=5$, $P=20$			
0.2	0.383	0.341	0.396
0.3	0.684	0.669	0.747
0.4	0.909	0.904	0.947
0.5	0.991	0.985	0.994
$N=15$, $n=3$, $P=5$			
0.2	0.089	0.090	0.101
0.3	0.159	0.145	0.169
0.4	0.252	0.226	0.274
0.5	0.379	0.334	0.405

mRNA samples can be pooled on the log scale directly. If we set σ_z^2 equal to zero while keeping all the other parameters unchanged as in the first scenario and calculate power again by simulation, we find that the power estimates are 0.383, 0.682, 0.906 and 0.987, which are very close to the power when $\sigma_z^2 = 0.05^2$. Therefore, the additional variance introduced in the pooling step does not affect power much, and the assumption that pooling mRNA occurs on the log scale is the main factor causing the overestimation in Kendziorski model. On the other hand, the power computed using the analytical expression in Section 2.3 is conservative because our estimate for the variance is conservative. Therefore, true power is at least high as our predicted power.

3.2 The effect of repeated measurements on power

For a given set of experimental conditions, biological, technical and weight variation in the pooled data are often fixed. Therefore, the power of the test for a given set of conditions depends on the number of pools, the number of repeated measurements per pool and the number of individuals per pool. Consider the following example: suppose that there are three treatment groups ($T=3$) and 100 individuals per treatment ($N=100$), and let the mean expression difference between any two adjacent treatment groups on the log scale be 0.5 ($\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.5$), which represents a 1.65-fold difference on the original scale. Suppose that total variation is equal to 1, biological variance is three times as large as technical variance ($\sigma_{\mathrm{b}}^2 = 0.75$ and $\sigma_{\mathrm{t}}^2 = 0.25$), and technical SD in the pooling step is 5% of the standardized mean ($\sigma_z^2 = 0.05^2$). Then, for a fixed number of arrays per treatment ($A = 5, 10, 15, 20$), the effect of obtaining repeated measurements on each pool on power is shown in Figure 1. We computed power analytically using Expressions (15) and (16) with any R and P values that match the equation $R * P = A$. Note, however, that R and P will always have integer values in an actual experiment. Power decreases as the number of repeated measurements per sample increases for fixed numbers of individuals and arrays. Therefore, when the number of subjects is fixed and the number of arrays is limited, a more efficient strategy is to create multiple pools and measure each once rather than to create fewer pools and measure each multiple times. This is consistent with findings in Kendziorski *et al.* (2003). In the remainder, we assume that each pool is measured once ($R = 1, P = A$).

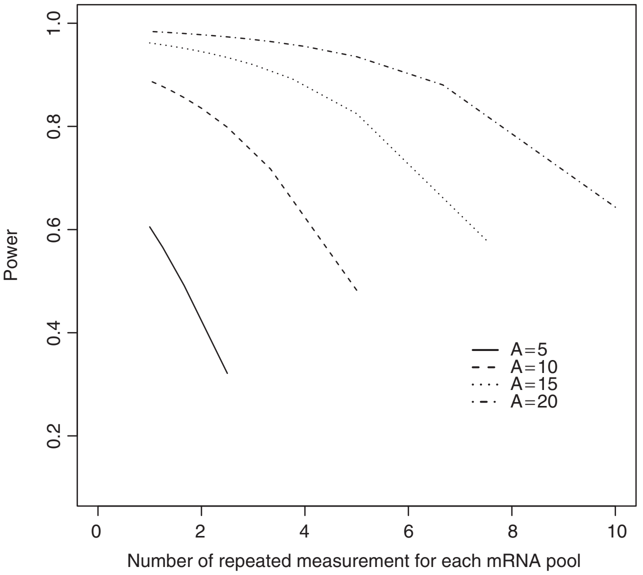

Fig. 1. The effect of repeated measurement on power for different total numbers of arrays per treatment: $T = 3, N = 100, \mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.5$, $\sigma_{\mathrm{b}}^2 = 0.75, \sigma_{\mathrm{t}}^2 = 0.25, \sigma_z^2 = 0.05^2$ and $A = 5, 10, 15, 20$.

3.3 The effect of the number of mRNA pools on power

Figure 2 shows power that is computed using Expressions (15) and (16) when different numbers of pools are created under various mean expression differences between adjacent treatments ($\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.2, 0.3, 0.4, 0.5$). For a fixed number of individuals, the power of the test based on individual samples is always higher than when samples are pooled, as would be expected. Power increases as the number of pools increases, and it is maximized when $P=N$, i.e. when we microarray each individual. The rate at which power increases with mean expression difference is relatively high when the number of pools is small, but relatively low when the number of pools is relatively large. When the number of pools is large enough (30 or higher, approximately), we observe no further increase in the power. For example, under $\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.4$, power increased by 0.2, 0.05 and 0.005 when P increased from 10 to 20, from 20 to 30 and from 50 to 60. The almost flat trend is especially obvious when the mean expression difference is larger ($\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.4, 0.5$). The slow or almost flat trend in the power curve makes it possible to find a pooling design with power that approaches the power that can be achieved with individual arrays and at the same time control costs. For example, when n changes from 1 to 2 (individual arrays versus pools of two individuals per sample), power dropped from 0.9999 to 0.9993, from 0.994 to 0.982 and from 0.91 to 0.85 when $\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.5, 0.4, 0.3$. The higher the power of tests based on individual samples, the higher the number of individuals that can be pooled together without significant

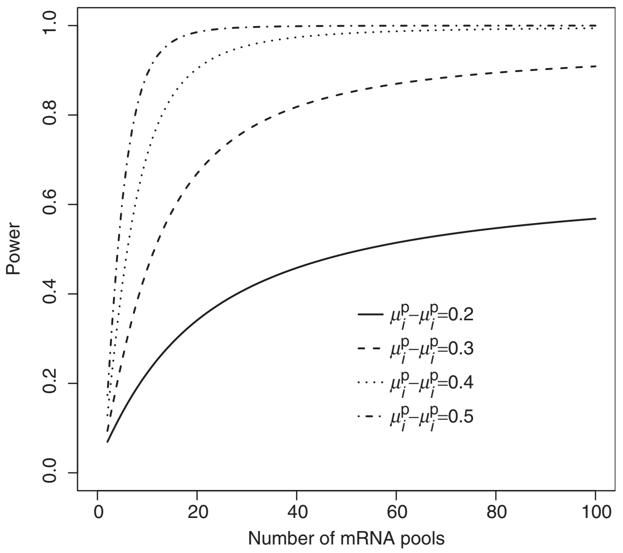

Fig. 2. Relationship between number of pools and power for different treatment effect sizes $\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.2, 0.3, 0.4, 0.5$ and for $T = 3$, $N = 100$, $\sigma_{\mathrm{b}}^2 = 0.75, \sigma_{\mathrm{t}}^2 = 0.25, \sigma_z^2 = 0.05^2$.

loss of information. For example, when $\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.5$, a design that involves forming $P = 10$ pools with $n = 10$ individuals each has 90% of the power of the design that involves no sample pooling, and yet the cost of arrays is only 10% of the cost of arraying every individual.

3.4 The effect of biological, technical and weight variability on power

From the results presented in Section 2.3, we know that power depends on σ_{b}^2 and σ_{t}^2. The effect of the ratio of biological to technical variance on power is shown in Figure 3. As would be expected, power in designs that involve pooling samples increases as the technical variance gets smaller relative to the biological variance. For example, when the mean expression difference is 0.5 and the design includes 10 pools of 10 samples each, power increases from 0.72 when $\sigma_{\mathrm{b}}^2 = \sigma_{\mathrm{t}}^2$ to 0.92 when $\sigma_{\mathrm{b}}^2 = 4\sigma_{\mathrm{t}}^2$.

The additional technical variation introduced in the pooling step does not appear to affect power much (Fig. 4), even if the pooling technical variance is rather high ($\sigma_z^2 = 0.2^2$). This is because in the denominator of Expression (16), σ_z^2 is very small compared to $e^{\sigma_{\mathrm{b}}^2} - 1$ and σ_{t}^2. Also the effect of pooling technical variation is further decreased by the factor $\frac{n-1}{n^2}$. Therefore, the additional technical variation introduced in the pooling step is not a major factor to consider in power calculation.

Samples of mRNA from individuals are sometimes pooled in microarray experiments, either because the biological material available from each individual is not sufficient to array or to keep costs down. It is to be expected that statistical tests to detect differences in mean gene expression levels across treatments will be affected when they are based on pools of mRNA samples rather than on individual samples, since some information is bound to be lost. In particular, the power of F-tests in the usual ANOVA models is expected to decrease when the experimental design involves pooling of individual samples.

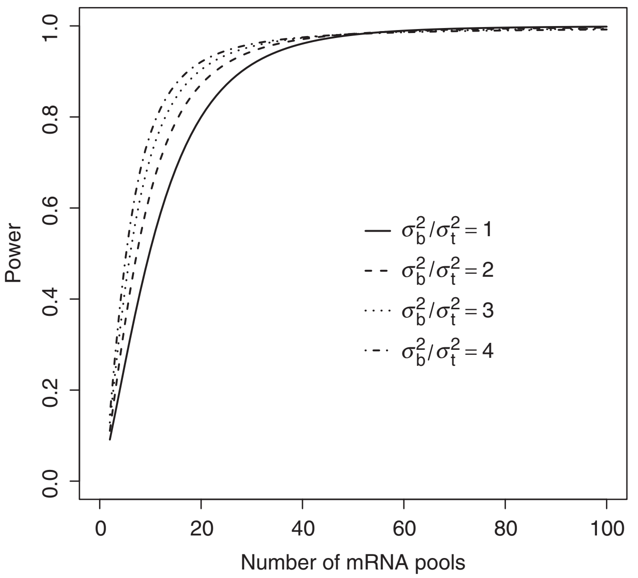

Fig. 3. Relationship between number of pools and power for different ratios of biological to technical variance $\sigma_{\mathrm{b}}^2/\sigma_{\mathrm{t}}^2 = 1, 2, 3, 4$ and for $T = 3, N = 100, \mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.4$ and $\sigma_z^2 = 0.05^2$.

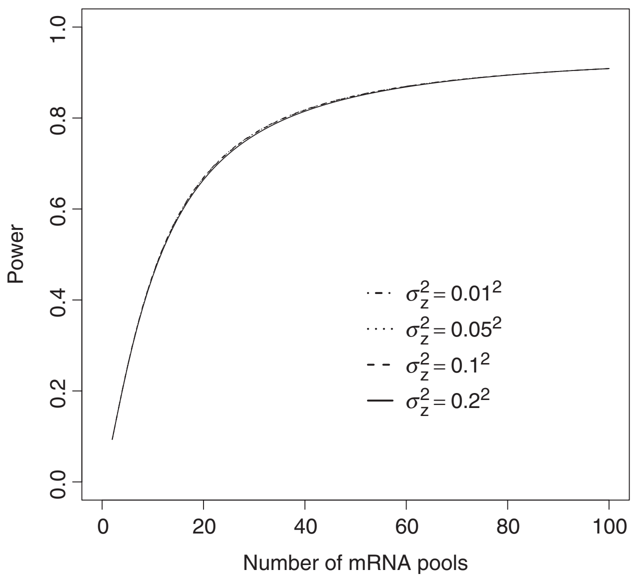

Fig. 4. Relationship between number of pools and power for different pooling technical variance $\sigma_z^2 = 0.01^2, 0.05^2, 0.1^2, 0.2^2$ and for $T = 3$, $N = 100, \mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = 0.3, \sigma_{\mathrm{b}}^2 = 0.75, \sigma_{\mathrm{t}}^2 = 0.25$.

Several authors have investigated the statistical properties of F-tests based on pooled mRNA samples (Kendziorski *et al.*, 2003 and Shih *et al.*, 2004). One limitation in these studies is that the statistical models adopted imply that the mRNA samples are pooled on the log scale, which is unrealistic. We investigated the power of F-tests in ANOVA models when mRNA samples are pooled, but extended the models so that the pooling process is carried out on the original scale. In our formulation, mRNA pools are weighted averages of individual mRNA samples and consider the measurement error that is introduced when pooling potentially different amounts of mRNA from individuals into a pool. We argue that when pooling is assumed to occur on the log scale, the power of the tests is overestimated and propose an approach to calculate power under the more realistic scenario of pooling on the original scale.

It is not possible to derive an analytical expression for the distribution of pooled gene expression on the log scale. Therefore, we assume that gene expression on the log scale is normally distributed. To check this assumption, we conducted simulation studies and found that, at least for the range common to microarray data, the normality assumption appears to be reasonable. Our focus is on deriving expressions to calculate the power of F-tests to detect mean gene expression difference across treatments in designs that involve pooling. Because the F-test is robust to modest departures from normality (Mendes and Pala, 2004), we anticipate that assuming a normal model for the gene log-expression values will not have a noticeable effect on our results. We show that the power estimated using the approach we propose here is conservative in that it tends to slightly underestimate true power; therefore, true power is at least as high as the estimates resulting from implementing the method we propose.

As might be expected, the power of the tests depends not only on the size of the treatment effect but also on the total number of individuals and pools, the number of pools per treatment, the number of replicated measurements obtained for each pool and the magnitude of biological and technical variability. For the technical variability, we distinguish the usual variance introduced in the various steps of microarray experiments and the variability that is introduced during the pooling process, resulting from the possibly differential contributions of individual samples to the pool.

We used simulated gene expression data to compare the power of F-tests that can be achieved when analyzing individual mRNA samples and under various pooling designs. We computed power analytically and also via simulation, and compared results to those that would be obtained by implementing the approach proposed by Kendziorski *et al.* (2003). We found that given a fixed number of individuals and arrays, power tends to be higher when a larger number of pools is measured once than when replicate measurements are obtained on a smaller number of pools. This holds for all values of the biological, technical and pooling variabilities considered in our study. Not surprisingly, we also found that power of tests based on individual samples is always higher than power based on pooled samples. For large enough effect sizes, however, it is possible to design an experiment that involves pooling mRNA samples that almost achieves the power that would be obtained when arraying individual samples, but at a fraction of the cost. Thus, our results suggest that under some conditions, pooling mRNA samples in microarray experiments can be a good strategy if cost is a consideration.

One of the important features of our model is that it attempts to mimic the pooling process as it happens in the lab. We assume that the mRNA pool is a weighted average of expression levels from individual mRNA samples, where the weights are random variables with mean $1/n$. Because the log is a non-linear transformation, the log of this weighted average will be different from a weighted average of log-transformed individual expression levels. Based on (10), we have

$$\mu_i^{\mathrm{p}} - \mu_i \sim \frac{\sigma_{\mathrm{b}}^2}{2} - \frac{1}{2n}(e^{\sigma_{\mathrm{b}}^2} - 1)(1 + n\sigma_w^2).$$

Simple calculus shows that this function of σ_b^2 is positive and increasing on the interval $(0, -\log(1/n + \sigma_w^2)]$. When $n \geq 2$, $-\log(1/n + \sigma_w^2)$ is typically larger than σ_{b}^2 in microarray experiments. Thus, the difference between μ_i^{p} and μ_i is expected to grow larger with σ_{b}^2. Shih *et al.* (2004) assumed in their work that $\mu_i^{\mathrm{p}} - \mu_i = 0$ and then tested this assumption using data collected in a microarray experiment on mice. They found that the number of genes with significantly different expression levels across different treatments was higher than that would have been expected by chance; this effect was even stronger when expression levels were high. Also, Kendziorski *et al.* (2005) confirmed further that the pools and the average of individuals were not always in agreement for certain genes and suggested that modeling the pooling process on the transformed scale could be a possible reason. These results can be explained well under our model. Since we show that $\mu_i^{\mathrm{p}} - \mu_i > 0$, the 95% confidence intervals for the difference between mean expression level in the pool and in individual arrays are not centered at zero. Further, since the difference between the two means can be approximated by a positive, increasing function of the biological variance, and the biological variance tends to be positively associated with gene expression level, we expect that the shifting of the confidence intervals will be more pronounced when the signal is stronger. In addition, confidence intervals that account for the added variance introduced in the pooling process are somewhat wider. According to our model and to the results obtained by simulation, the proportion of genes that fall outside the 95% confidence intervals discussed by Shih *et al.* (2004) is 0.077, 0.096, 0.101 and 0.151 when the biological variance is 0.2, 0.4, 0.6 or 0.8 respectively, and the technical variance is held constant at 0.25. These unexpectedly high proportions can be explained under our model, which accommodates pooling on the original (rather than on the log) scale.

One other interesting finding is that after log transformation and assuming of normality, the expected mean expression difference in a design that involves pooling is the same as in a design without pooling, i.e. $\mu_i^{\mathrm{p}} - \mu_j^{\mathrm{p}} = \mu_i - \mu_j$. Thus, a test for the hypothesis that $\mu_i^{\mathrm{p}} = \mu_j^{\mathrm{p}}$ is equivalent to the test $\mu_i = \mu_j$. This property might not hold under other transformations, however.

We have focused on power calculation under designs that pool or do not pool mRNA when testing expression differences for a single gene. In microarray experiments, tests involve tens of thousands of genes and the biological variation may differ from gene to gene. Therefore, pool designs required to reach a certain power may be different between genes due to differences in biological variation across genes. Thus, finding a single efficient design for pooling mRNA which results in the desired power for all the genes in the microarray experiment might be a challenge.

Jung (2005) proposed a microarray sample size calculation method for the two-group comparison problem that allows researchers to determine the sample size necessary to identify approximately r_1 differentially expressed genes while controlling the false discovery rate (FDR) at a desired level f. This basic concept is easily extended to the T-treatment case and designs that involve pooling. Borrowing notation and concepts from Jung (2005), we have that the individual significance level necessary for identification of approximately r_1 differentially expressed genes while controlling FDR at level f is

$$\alpha^* = \frac{fr_1}{(1-f)m_0}, \tag{19}$$

where, m_0 is the number of non-differentially expressed genes among all genes tested. Given α^*, we seek values of n and P such that

$$\sum_{g \in M_1} \xi_g(\alpha^*) \geq r_1, \tag{20}$$

where, M_1 denotes the set of indices of differentially expressed genes and $\xi_g(\alpha^*)$ denotes the power of the α^* level test for gene g that can be approximated using (15) and (16). Based on the results of our simulation study in Section 3.1, the use of our power approximation in (22) should suggest sample sizes that are at least as large as necessary to identify the desired number of differentially expressed genes. On the other hand, we would expect the method of Kendziorski *et al.* (2003) to recommend sample sizes that are smaller than those actually needed to achieve the specified level of discovery.

As in any power and sample size calculation, users are required to provide the values of unknown parameters like m_0 and—separately for each gene—the variance and mean parameters necessary for computing (15) and (16). While it is conceivable to estimate such parameters from pilot microarray experiments, a discussion of such strategies is beyond the scope of this article. As a starting point in the absence of pilot data, researchers may wish to assume that effect sizes and variance components are identical for all differentially expressed genes, in which case (22) can be simplified as discussed by Jung (2005).

Whether to pool individuals and how to pool them to minimize the loss of information are important issues in microarray experiments. For a fixed total number of individuals and arrays, a design that includes mRNA pools always leads to smaller power than a design in which each array corresponds to an individual. Under some conditions, however, the loss of power is small, and it is possible to find a low-cost design which almost achieves the power that can be obtained when arraying each individual.

ACKNOWLEDGEMENT

This study was funded by the Center for Integrated Animal Genomics at Iowa State University.

Conflict of Interest: none declared.

REFERENCES

Chen,J. *et al.* (2004) Analysis of variance components in gene expression data. *Bioinformatics*, **20**, 1436–1446.

Geller,S. *et al.* (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**, 1817–1823.

Han,E. *et al.* (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *J. Gerontol. Biol. Sci.*, **4**, 306–315.

Jung,S.-H. (2005) Sample size for FDR control in microarray data analysis. *Bioinformatics*, **21**, 3097–3104.

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

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

Lu,C. (2004) Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays. *BMC Bioinformatics*, **5**, 103–108.

Mendes,M. and Pala,L. (2004) Evaluation of four tests when normality and homogeneity of variance assumptions are violated. *J. App. Sci.*, **4**, 38–42.

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

APPENDIX

Expression (8) is derived as follows:

$$\begin{aligned} \nu_i^{\mathrm{p}} &= E(m_{ij}^{\mathrm{p}}) \\ &= E[\sum_{k=1}^{n}(w_{ijk} \times m_{ijk})] \\ &= \sum_{k=1}^{n} E(w_{ijk}) \times E(m_{ijk}) \\ &= e^{\mu_i + \frac{\sigma_{\mathrm{b}}^2}{2}} \sum_{k=1}^{n} E(w_{ijk}) \\ &= e^{\mu_i + \frac{\sigma_{\mathrm{b}}^2}{2}}. \end{aligned}$$

To simplify notation in the derivation of (9), we suppress the i and j subscripts on w_{ijk} and m_{ijk}. Using this simplification, we have

$$\begin{aligned} \sigma_i^{\mathrm{p}2} &= \mathrm{Var}\left[m_{ij}^{\mathrm{p}}\right] = \mathrm{Var}\left[\sum_{k=1}^{n}(w_k m_k)\right] \\ &= E\left[\left(\sum_{k=1}^{n} w_k m_k\right)^2\right] - \left[E\left(\sum_{k=1}^{n} w_k m_k\right)\right]^2 \\ &= E\left[\sum_{k=1}^{n}(w_k m_k)^2\right] + 2\sum_{k=1}^{n}\sum_{l>k}^{n} E(w_k m_k w_l m_l) \\ &\quad - \left[\sum_{k=1}^{n} E(w_k)E(m_k)\right]^2 \\ &= \sum_{k=1}^{n}\left[E(w_k)^2 E(m_k)^2\right] + 2e^{2\mu_i + \sigma_{\mathrm{b}}^2}\sum_{k=1}^{n}\sum_{l>k}^{n} E(w_k w_l) \\ &\quad - \left[e^{\mu_i + \frac{\sigma_{\mathrm{b}}^2}{2}}\sum_{k=1}^{n} E(w_k)\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= e^{2\mu_i+2\sigma_{\mathrm{b}}^2}\sum_{k=1}^{n}E(w_k)^2 \\
&\quad + e^{2\mu_i+\sigma_{\mathrm{b}}^2}E\left[\left(\sum_{k=1}^{n}w_k\right)^2-\sum_{k=1}^{n}w_k^2\right] \\
&\quad - e^{2\mu_i+\sigma_{\mathrm{b}}^2} \\
&= ne^{2\mu_i+2\sigma_{\mathrm{b}}^2}E(w_k^2)+e^{2\mu_i+\sigma_{\mathrm{b}}^2}\left[1-nE(w_k^2)\right]-e^{2\mu_i+\sigma_{\mathrm{b}}^2} \\
&= nE(w_k)^2\left(e^{2\mu_i+2\sigma_{\mathrm{b}}^2}-e^{2\mu_i+\sigma_{\mathrm{b}}^2}\right) \\
&= n\left[(Ew_k)^2+\sigma_w^2\right]\left(e^{2\mu_i+2\sigma_{\mathrm{b}}^2}-e^{2\mu_i+\sigma_{\mathrm{b}}^2}\right) \\
&= n\left(\frac{1}{n^2}+\sigma_w^2\right)\left(e^{2\mu_i+2\sigma_{\mathrm{b}}^2}-e^{2\mu_i+\sigma_{\mathrm{b}}^2}\right) \\
&= \left(\frac{1}{n}+n\sigma_w^2\right)\left(e^{2\mu_i+2\sigma_{\mathrm{b}}^2}-e^{2\mu_i+\sigma_{\mathrm{b}}^2}\right).
\end{aligned}
$$

To show that $\sigma_w^2 \sim \frac{n-1}{n^3}\sigma_z^2$, we use the multivariate Delta method. Recall that $\mathrm{Var}[g(x)] \sim [g'(\mu_x)]^t\Sigma_x g'(\mu_x)$ where, x is a random vector with mean μ_x and variance-covariance matrix Σ_x and $g'(\mu_x)$ denotes the vector of partial derivatives of g evaluated at μ_x. If we let $x=(z_{ij1},\ldots,z_{ijn})^t$ and choose $g(x)=x_1/\sum_{k=1}^{n}x_k$, then $g(x)=w_{ij1}$, $\mu_x=(1,\ldots,1)^t$, $\Sigma_x=\sigma_z^2 I_n$ (where I_n is the $n \times n$ identity matrix), and

$$
\frac{\partial g}{\partial x_\ell}=\begin{cases}\dfrac{\sum_{k=1}^{n}z_{ijk}-z_{ij1}}{\left(\sum_{k=1}^{n}z_{ijk}\right)^2} & for\ell=1,\\[2ex] \dfrac{-z_{ij\ell}}{\left(\sum_{k=1}^{n}z_{ijk}\right)^2} & for\ell\neq 1.\end{cases}
$$

Thus,

$$
\begin{aligned}
\sigma_w^2 &= \mathrm{Var}(w_{ij1}) \sim [g'(\mu_x)]^t\Sigma_x g'(\mu_x) \\
&= \left\{\left(\frac{n-1}{n^2}\right)^2+(n-1)\left(\frac{-1}{n^2}\right)^2\right\}\sigma_{\mathrm{z}}^2 \\
&= \frac{(n-1)^2+n-1}{n^4}\sigma_z^2=\frac{(n-1)}{n^3}\sigma_{\mathrm{z}}^2.
\end{aligned}
$$