Biostatistics (2003), **4**, 3, *pp.* 465–477
Printed in Great Britain

The efficiency of pooling mRNA in microarray experiments

C. M. KENDZIORSKI[†]
Department of Biostatistics and Medical Informatics, University of Wisconsin, 6729 Medical Sciences Center, 1300 University Avenue, Madison, WI 53792, USA
kendzior@biostat.wisc.edu

Y. ZHANG
Department of Statistics, University of Wisconsin, Madison, WI 53792, USA

H. LAN, A. D. ATTIE
Department of Biochemistry, University of Wisconsin, Madison, WI 53792, USA

SUMMARY

In a microarray experiment, messenger RNA samples are oftentimes pooled across subjects out of necessity, or in an effort to reduce the effect of biological variation. A basic problem in such experiments is to estimate the nominal expression levels of a large number of genes. Pooling samples will affect expression estimation, but the exact effects are not yet known as the approach has not been systematically studied in this context. We consider how mRNA pooling affects expression estimates by assessing the finite-sample performance of different estimators for designs with and without pooling. Conditions under which it is advantageous to pool mRNA are defined; and general properties of estimates from both pooled and non-pooled designs are derived under these conditions. A formula is given for the total number of subjects and arrays required in a pooled experiment to obtain gene expression estimates and confidence intervals comparable to those obtained from the no-pooling case. The formula demonstrates that by pooling a perhaps increased number of subjects, one can decrease the number of arrays required in an experiment without a loss of precision. The assumptions that facilitate derivation of this formula are considered using data from a quantitative real-time PCR experiment. The calculations are not specific to one particular method of quantifying gene expression as they assume only that a single, normalized, estimate of expression is obtained for each gene. As such, the results should be generally applicable to a number of technologies provided sufficient pre-processing and normalization methods are available and applied.

Keywords: Experimental design; Gene expression; Microarrays; Pooled sample.

1. INTRODUCTION

Microarray technologies are now widely used to gain insight into the genetic basis of many complex biological processes. Although quite powerful, arrays are relatively expensive and, as a result, replication is done at a minimum. Reducing replication can adversely affect estimation of gene expression and

[†] To whom correspondence should be addressed

Biostatistics 4(3) © Oxford University Press; all rights reserved.

assessment of differential expression. To address these problems, mRNA samples are often pooled across subjects (Brown *et al.*, 2001; Jin *et al.*, 2001; Sotiriou *et al.*, 2001; Waring *et al.*, 2001; Ernard *et al.*, 2002; discussion in Churchill and Oliver, 2001). In addition to reducing the effect of biological variation, pooling is sometimes done out of necessity. In studies of tissues such as the hypothalamus or pancreatic islets, or in studies of small animals such as Drosophila, it can be difficult if not impossible to obtain a sufficient amount of mRNA from a single subject. In such cases, pooling can be useful. Whatever the reason for mRNA pooling, it is clear that doing so will affect data analysis and inference. The exact effects are not yet known as pooling has not been systematically studied in the context of microarray experiments.

Statistical questions related to pooling samples have received considerable attention in other areas of application. Such work began as early as the 1940s when WWII inductees were required to undergo blood tests for syphilitic antigens. Dorfman (1943) showed that the number of total tests could be reduced if blood samples were first pooled and tested, followed by individual retesting of all composite samples in positive pools only. The idea has been extended to a number of areas (for a review, see Gastwirth, 2000) including the detection of mutant alleles in a population (Amos *et al.*, 2000), the estimation of disease prevalence (Gastwirth and Hammick, 1989), and the estimation of joint allele frequencies and linkage disequilibrium (Pfeiffer *et al.*, 2002). In each of these cases, a general goal is to detect the presence of a characteristic, followed perhaps by estimation of population prevalence. The objective in a microarray experiment is different as gene expression quantification, as opposed to simply detection, is of interest.

In this paper, we consider the effects of mRNA pooling on estimates of gene expression. Quality of the estimates is assessed by the bias and variance along with the length of the resulting confidence interval. Estimates from both pooled and non-pooled designs are evaluated in Section 2. Conditions to ensure unbiased estimates from both designs are specified. In addition, it is shown that by pooling subjects, one can reduce the number of arrays required in an experiment while maintaining estimates and confidence intervals comparable to those obtained without pooling. A formula specifying the exact number of subjects and arrays required is given. The formula assumes that the variance components are known. This assumption is relaxed in Section 3 and implications of variance component estimation on the total number of subjects and arrays are considered. The assumptions that facilitate the calculations are discussed in the context of a quantitative real-time PCR (referred to hereinafter as RT-PCR) experiment in Section 4. The calculations throughout address the case where all subjects are sampled from the same population and thus share common nominal levels of gene expression. A brief note on relaxing this assumption is given in the Appendix.

2. EVALUATION OF ESTIMATES FROM POOLED AND NON-POOLED DESIGNS

A main goal of any microarray experiment is to estimate the nominal expression levels of a large number of genes. For m genes, we denote this nominal level by the m-vector θ. An experiment to estimate θ consists of extraction and labeling of mRNA from n_{s} subjects, hybridization to n_{a} arrays, followed by scanning and image processing. The technology that one uses dictates in large part the details of each of these steps. Our concern is not with a specific technology or image processing method, but with the actual measurements of gene expression, however obtained.

We assume here that sufficient data pre-processing has been done to remove artifacts within the array and across a set of arrays. In this case, the gene expression measurements for gene g denoted by $x_{g,1}, x_{g,2}, \dots, x_{g,n_{\mathrm{a}}}$ are considered independent and identically distributed samples from a distribution parametrized by mean θ_g. The processed measurements are assumed to be affected primarily by two sources of variation: subject-to-subject and array-to-array variability (referred to hereinafter as simply biological variability and technical variability, respectively). In spite of fluctuations, the average $\bar{x}_g = \frac{1}{n_{\mathrm{a}}} \sum_{k=1}^{n_{\mathrm{a}}} x_{g,k}$ estimates θ_g.

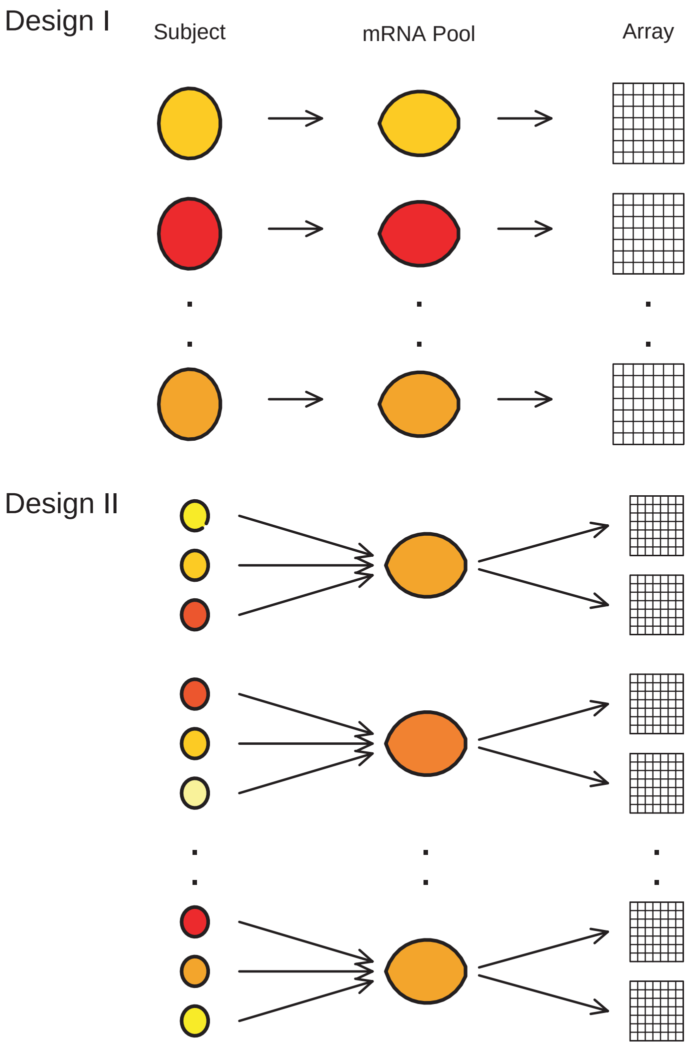

Fig. 1. Schematic diagram of designs I and II for a given number of subjects and arrays (numbers chosen only for illustration purposes). Design I requires that an mRNA sample from each subject is probed with one array. In design II, mRNA samples from different subjects are first pooled together; then, replicate samples are drawn from the pool and hybridized onto a set of arrays. This process is repeated some number of times (equal to the number of mRNA pools, n_{p}). Note that for design II, the number of subjects that make up an mRNA pool, (here, $r_{\mathrm{s}} = 3$), need not equal the number of arrays that probe that pool, (here, $r_{\mathrm{a}} = 2$).

By studying the effects of these different sources of variation, we can determine values of n_{s} and n_{a} that provide for optimal estimates. Three things considered in assessing the quality of $\bar{x}_g$ are the bias and the variance of the estimate along with the length of the resulting confidence interval for θ_g.

The problem of estimating gene expression could be addressed by implementing any one of many potential experimental designs. We consider two designs below. The gene-specific subscript will be dropped as each gene is considered individually. Gene-specific dependencies are discussed in Section 5. The first design (design I) requires that an mRNA sample from each subject is probed with one array

(Figure 1, upper panel). Due to variability among subjects and measurement error inherent to the array, any observed x_i is defined as

$$x_i = \theta + \epsilon_i + \xi_i \tag{2.1}$$

$i = 1, 2, \ldots, n$, where $n = n_{\mathrm{s}} = n_{\mathrm{a}}$. Here, ϵ_i represents biological variation among subjects and ξ_i represents measurement error (technical variation). We assume that $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$ and $\xi_i \sim N\left(0, \sigma_\xi^2\right)$; and furthermore that biological and technical errors are independent.

For the second design, design II, mRNA samples from r_{s} different subjects are first pooled together; then, r_{a} replicate samples are drawn from the mRNA pool and hybridized onto a set of arrays. This is repeated some number of times denoted by n_{p} to represent the number of distinct mRNA pools (Figure 1, lower panel). In design I, $r_{\mathrm{s}} = r_{\mathrm{a}} = 1$ and $n_{\mathrm{p}} = n_{\mathrm{s}} = n_{\mathrm{a}}$; note that unlike design I, here the number of subjects that contribute mRNA to a pool ($r_{\mathrm{s}} \geqslant 1$) need not equal the number of arrays used to probe that pool ($r_{\mathrm{a}} \geqslant 1$) and thus the total number of subjects ($n_{\mathrm{s}} = r_{\mathrm{s}} \cdot n_{\mathrm{p}}$) need not equal the total number of arrays ($n_{\mathrm{a}} = r_{\mathrm{a}} \cdot n_{\mathrm{p}}$). For this design,

$$x_{i,j} = \theta + \epsilon_i{}' + \xi_{i,j} \tag{2.2}$$

where $i = 1, 2, \ldots, n_{\mathrm{p}}$ and $j = 1, 2, \ldots, r_{\mathrm{a}}$. Here, $\epsilon_i{}'$ represents pool-to-pool variability and again we assume Gaussian errors and independence between biological and technical error. If r_{s} is very large, biological variability might be negligible. However, for a moderate number of subjects, one should account for variation among the mRNA pools (Churchill and Oliver, 2001). Assuming that the mRNAs average out across the pool, we would expect the variability of $\epsilon_i{}'$ to be reduced to $\sigma_\epsilon^2/r_{\mathrm{s}}$.

We evaluate designs I and II by considering the finite-sample properties of $\bar{x}$. In particular, we consider the bias and variance, as well as the lengths of the associated confidence intervals. For both designs, $E[\bar{x}_{..}] = \theta$;

$$\sigma^2_{\bar{x},(1)} = \frac{1}{n_{\mathrm{p}1}}\left(\sigma_\epsilon^2 + \sigma_\xi^2\right) \quad \text{and} \quad \sigma^2_{\bar{x},(2)} = \frac{1}{n_{\mathrm{p}2}}\left(\frac{\sigma_\epsilon^2}{r_{\mathrm{s}2}} + \frac{\sigma_\xi^2}{r_{\mathrm{a}2}}\right) \tag{2.3}$$

where $\sigma^2_{\bar{x},(1)}$ and $\sigma^2_{\bar{x},(2)}$ denote the variance of $\bar{x}$ in designs I and II, respectively. In each case, the estimator is unbiased and the variance decreases as the number of arrays increases. The precision of the estimate in design I depends only on the variance components and the number of replicate pools (since $r_{\mathrm{s}1} = r_{\mathrm{a}1} = 1$); for design II the precision also depends on the number of mRNA samples that are pooled and the number of arrays used to probe a given pool. As shown below, this fact can be used to obtain an estimate of gene expression that is as precise as that obtained from design I, but using fewer arrays.

Consider the squared length of the $100(1-\alpha)\%$ confidence interval (CI) for θ given by $4z^2_{\alpha/2}\left(\sigma^2_{\bar{x}}\right)$ where z_α is the αth quantile for a standard Normal distribution. A comparison of the confidence intervals for designs I and II (with squared lengths denoted respectively by l_1^2 and l_2^2) is obtained by considering the ratio $R = l_1^2/l_2^2$. Letting $n_{\mathrm{s}1}$ ($n_{\mathrm{s}2}$) and $n_{\mathrm{a}1}$ ($n_{\mathrm{a}2}$) denote the total number of subjects and arrays in design I (II), it can be shown that $R = 1$ when

$$n_{\mathrm{s}2} = n_{\mathrm{s}1}\left[\frac{\lambda}{K(\lambda+1) - \frac{n_{\mathrm{a}1}}{n_{\mathrm{a}2}}}\right] \tag{2.4}$$

where $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\xi^2}$ and K is the ratio of the critical values associated with designs I and II (here, $K = z^2_{\alpha/2}/z^2_{\alpha/2} = 1$). Equation (2.4) shows that by increasing the number of subjects in design II, the number of arrays can be decreased, without changing the squared length of the confidence interval.

As an example, consider an experiment with the mRNA from 15 individuals probed using 15 arrays (design I with $n_{\mathrm{s}1} = n_{\mathrm{a}1} = n_{\mathrm{p}1} = 15$) and suppose that the biological variability is four times as large as the technical variability ($\sigma_\epsilon^2 = 4$ and $\sigma_\xi^2 = 1$ which gives $\lambda = 4$). In this case, the variance of the estimate is 0.333 (by equation (2.3)) and the squared length of the confidence interval is 5.122 ($\alpha = 0.05$). The number of subjects and arrays required in a pooled experiment to obtain a comparable confidence interval is given by equation (2.4). Since $n_{\mathrm{s}2}$ as defined by equation (2.4) might not be integer valued, the values given are considered lower bounds on the total number of subjects. If the total number of arrays in the pooled experiment is reduced to 12, equation (2.4) indicates that the mRNA from 16 subjects is required to obtain an interval comparable to that obtained without pooling. If the number of arrays is reduced to 10, the mRNA from at least 18 subjects ($n_{\mathrm{s}2} = 17.14$) is required.

As these calculations report the total number of subjects and arrays in a given design, they give no information about the exact way in which to allocate the totals to pools. For the case discussed above (with 18 subjects and 10 arrays), one could construct three pools each containing the mRNA from six subjects and probe two of the pools using three arrays and one of the pools using four arrays. Alternatively, the mRNA from nine subjects could be combined to give a total of two pools, each of which is probed using five arrays. Each of these designs gives an equally precise estimate of θ. Another allocation which gives the same precision is combining the mRNA from all subjects into one pool and probing that one pool with all arrays ($n_{\mathrm{p}} = 1$). With this last design, all information regarding biological variability is lost and thus σ_ϵ^2 is not estimable (recall here that we are assuming known variance components). This is not useful in practice; thus, when determining the optimal allocation of the total number of subjects and arrays, the ability to estimate the variance components, σ_ϵ^2 and σ_ξ^2, must be considered.

3. Determination of the total number of subjects, arrays, and pools

The variance components σ_ϵ^2 and σ_ξ^2 are not known in practice, and as a result the variance of $\bar{x}_{..}$ is estimated by

$$W = \left(\frac{\hat{\sigma}_\epsilon^2}{n_{\mathrm{s}}} + \frac{\hat{\sigma}_\xi^2}{n_{\mathrm{a}}}\right)$$
$$= \frac{1}{n_{\mathrm{p}}(n_{\mathrm{p}}-1)} \sum_{i=1}^{n_{\mathrm{p}}} (\bar{x}_{i\cdot} - \bar{x}_{..})^2.$$

In this case, the squared length of the confidence interval for θ is a random quantity given by

$$l^2 = 4\left(t_{n_{\mathrm{p}}-1,\alpha/2}\right)^2 (W)$$

where $t_{\nu,\alpha}$ is the αth quantile for the Student t distribution with ν degrees of freedom.

Note that for a fixed number of subjects and arrays, the squared length of the interval is minimized when the number of pools is maximized. This happens when the number of arrays used to probe a pool, r_{a}, is 1 (n_{a}=n_{p}). This holds by definition for design I and we impose this constraint throughout for design II. A comparison of the lengths for designs I and II is obtained by the ratio

$$\frac{l_1^2}{l_2^2} = \frac{t_1^2 \hat{\sigma}_{\bar{x},(1)}^2}{t_2^2 \hat{\sigma}_{\bar{x},(2)}^2}$$

where t_1 and t_2 denote the critical t values for designs I and II respectively.

To compare designs I and II as in Section 2, we could identify the total numbers of subjects, arrays, and pools in each design such that $R_2 = \frac{E[l_1^2]}{E[l_2^2]} = 1$. *Expected* squared lengths are now considered since

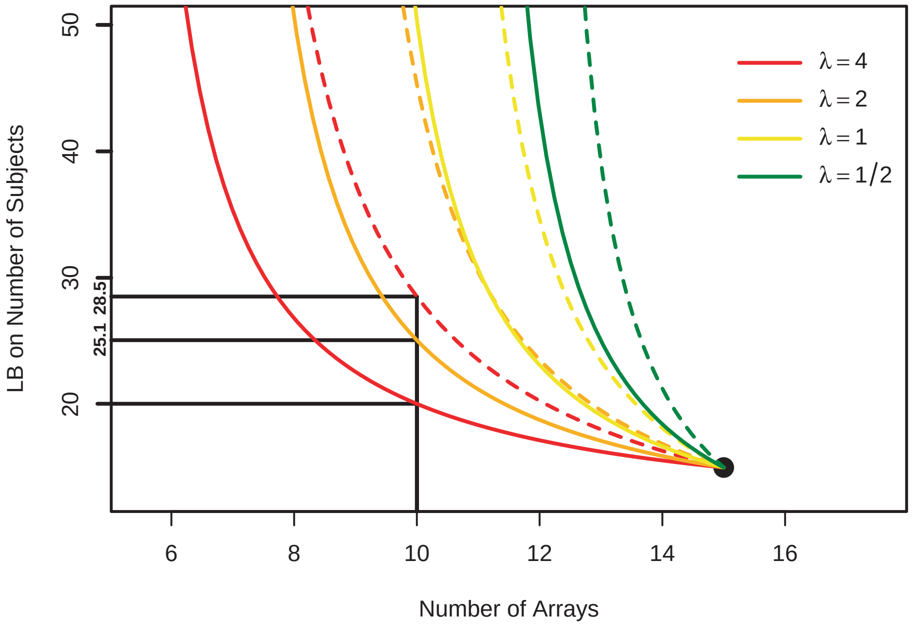

Fig. 2. The solid lines give the total number of arrays and lower bound (LB) on the total number of subjects required in design II to obtain a 95% confidence interval with expected squared length equal to that obtained using design I with 15 subjects and 15 arrays, for varying lambda ($\lambda = \sigma_\epsilon^2/\sigma_\xi^2$). For $\lambda = 4$, design I gives $E[l^2] = 6.133$ and $\text{var}(\bar{x}) = 0.333$ (see Table 1). For design II, the numbers required are given by the leftmost solid line shown. Thus, if 10 arrays are run, the mRNA from at least 21 subjects ($n_{s2} = 20.04$) is required. For $\lambda = 2$, if 10 arrays are run, mRNA from at least 26 subjects ($n_{s2} = 25.07$) is required. The dashed lines give the totals required to obtain equivalent variances on the squared lengths.

the variance components must be estimated. W is an unbiased estimator and so $R_2 = 1$ when equation (2.4) is satisfied for $K = t_1^2/t_2^2$. Thus, the totals determined by equation (2.4) give designs resulting in confidence intervals with equivalent expected squared lengths. As discussed below, variance component estimation has a small effect on the total numbers of subjects and arrays.

Consider the example in Section 2 ($n_{s1} = n_{a1} = n_{p1} = 15$; $\lambda = 4$) and suppose that the variance components are not known. If the total number of arrays in design II is reduced to 12, equation (2.4) with $K = t_1^2/t_2^2$ indicates that the mRNA from at least 18 subjects ($n_{s2} = 17.15$) is required. For 10 arrays, at least 21 subjects ($n_{s2} = 20.04$) are required. Recall that when the variance components are assumed known, a design with 12 (10) arrays requires 16 (18) subjects (see Section 2). There is a slight increase in the number of subjects required here as the variance components are no longer assumed to be known.

Figure 2 gives all possible combinations of the total numbers of subjects and arrays required in design II so that $R_2 = 1$ when $n_{s1} = n_{a1} = 15$. Results are shown for varying λ and indicate that pooling is most advantageous when λ is large. This makes sense since it is the effect of biological variability that is reduced by pooling. Table 1 gives a number of designs for which $R_2 = 1$ for other values of n_{s1} and n_{a1}. As shown, the expected squared lengths of the confidence intervals are the same for each design pair; the precision of the estimator $\bar{x}$ is slightly higher in the pooled design (since $R_2 = 1 \Rightarrow \sigma_{\bar{x},(1)}^2 > \sigma_{\bar{x},(2)}^2$ when $n_{a2} < n_{a1}$).

Equation (2.4) does not consider the variance of the squared lengths, which might be important to

Table 1. *Pairs of designs for which $R_2 = 1$ for the case where the biological variability is four times as large as the technical variability* ($\lambda = 4$). *The fourth column gives the expected squared length of the 95% confidence interval using the given total number of subjects and arrays. The last column gives the variance of the estimates* (*calculated from equation* (2.3)). *Since* n_{s2} *as defined by equation* (2.4) *might not be integer valued, the values given are considered lower bounds* (*LB*) *on the total number of subjects. In practice, the total number of subjects would be increased from this lower bound, resulting in an estimate with slightly lower variance and shorter confidence interval.*

Design	LB on total number of subjects (n_s)	Total numbers of arrays (n_a)	$E[l^2]$	var($\bar{x}$)
I	5.0	5	30.835	1.000
II	7.826	4	30.835	0.761
I	10.0	10	10.235	0.500
II	18.137	6	10.235	0.387
I	15.000	15	6.133	0.333
II	20.036	10	6.133	0.300
I	20.000	20	4.381	0.250
II	23.336	15	4.381	0.238
I	25.000	25	3.408	0.200
II	27.687	20	3.408	0.194

ensure reproducibility of the experiment. Consideration of $RV = \frac{\text{var}[l_1^2]}{\text{var}[l_2^2]}$ shows that $RV = 1$ when

$$n_{s2} = n_{s1}\left[\frac{\lambda}{K\,(\lambda+1)\,\frac{\sqrt{n_{a2}-1}}{\sqrt{n_{a1}-1}} - \frac{n_{a1}}{n_{a2}}}\right]. \tag{3.1}$$

The totals for some selected designs are shown in Figure 2 (dashed lines). Note that $RV = 1$ implies $R_2 \geqslant 1$ (when $n_{a2} \leqslant n_{a1}$) and so totals determined by equation (3.1) ensure for the pooled design that both the expectation and variance of the squared lengths of the confidence intervals are no bigger than those from design I; and the gene expression estimates are more precise. Recall that we have imposed the constraint $r_a = 1$ since in this case, the expected squared length of the confidence interval is minimized. We note here that under this constraint, var[l^2] is also minimized.

3.1 *Cost analysis*

A comparison of designs is more meaningful when cost is taken into consideration. The cost of one experiment is approximately $p_s \cdot n_s + p_a \cdot n_a$, where p_s and p_a represent the prices of one subject and one array, respectively. This formula is approximate as it does not account directly for economies of scale. For a typical array experiment, we estimate that $p_s = \$50.00$ and $p_a = \$700.00$ where p_s includes one animal and labor; p_a includes one array, reagents, and labor. These prices are consistent with our experience using Affymetrix chips in mouse experiments. We did adjust the unit prices in an attempt to account approximately for economies of scale.

Consider some selected designs. Design I, using 15 subjects and 15 arrays, costs $11 250. As arrays are still quite expensive relative to subjects, implementation of design II can result in substantial decrease in total cost. Design II with 26 subjects and 10 arrays costs $8300; 21 subjects and 10 arrays costs $8050. Of course, design II should only be used if properties of the gene expression estimate and associated

Table 2. *Gene expression measurements were quantified by RT-PCR for six genes in five mice using designs I and II. Variability across the measurements from design I consists of biological and technical variability* ($\hat{\sigma}^2_{\epsilon,\mathrm{RT}} + \hat{\sigma}^2_{\xi,\mathrm{RT}}$)*; the variability among RT-PCR measurements from the pooled mRNA samples* (*design II*) *consists only of technical variability* ($\hat{\sigma}^2_{\xi,\mathrm{RT}}$)*. The subscript RT is used to emphasize that the estimates are obtained using RT-PCR, and not microarray, data.*

Gene	$\hat{\sigma}^2_{\epsilon,RT} + \hat{\sigma}^2_{\xi,\mathrm{RT}}$	$\hat{\sigma}^2_{\xi,\mathrm{RT}}$	$\hat{\lambda}_{\mathrm{RT}}$
1	0.0361	0.0036	9.03
2	0.1521	0.0144	9.56
3	0.0841	0.0100	7.41
4	0.2401	0.0324	6.41
5	0.4489	0.0961	3.67
6	0.2209	0.0324	5.82

confidence interval are maintained. The particular pooling designs that result in comparable properties are determined by the biological and technical variability as quantified by λ. Thus, for these calculations to be used in practice, an estimate of λ is required. Preliminary experiments to estimate λ in an RT-PCR experiment are discussed below.

4. DATA

A quantitative RT-PCR experiment was done to quantify the mRNA abundance of six genes in the livers of five mice, using designs I and II. RT-PCR is a relatively inexpensive method that quantifies mRNA abundance. It requires significantly less total RNA and has a wider range of detection than a microarray experiment. The disadvantage is that expression is quantified one gene at a time. In this approach, total RNA was isolated from five mice using RNAzol reagent (Tel-Test, Inc.). For design I, RT-PCR measurements were obtained for each of the six genes using total RNA samples from the five individual mice. For design II, the five individual total RNA samples were pooled and RT-PCR measurements were obtained for each of the six genes using pooled RNA samples. Five replicate measurements were made using the pooled RNA for each of the six genes.

For each gene and each RNA sample (1 μg, from an individual or the pool), first strand cDNA was synthesized using SuperScript II Reverse Transcriptase (Gibco BRL) primed with a mixture of oligo-dT and random hexamers. Real-time PCR was performed using a GeneAMP 5700 Sequence Detection System (Applied Biosystems). The reaction was carried out in a 25 μl volume in 1 x SYBR Green PCR Core Reagents (Applied Biosystems) containing cDNA template from 10 ng of total RNA and 6 pmol each primer. For each gene, the cDNA samples were grided onto an optical 96-well reaction plate in an alternating and duplicating manner. The average measurements from at least two repeats were taken. The expression level of 18S RNA was used as a normalization control. For each reaction, we determined the cycle at which the abundance of accumulated PCR product crosses a specific threshold: the threshold cycle (C_{T}). The difference in average C_{T} values between the 18S RNA and a specific gene was calculated for each individual (Δ C_{T}). This difference is comparable to the log-transformed, normalized mRNA abundance.

The variability across the RT-PCR measurements from design I is made up of biological and technical variability; the variability between the measurements from design II consists only of experimental variability. The estimates are shown in Table 2. For each of these genes, $\hat{\lambda}_{\mathrm{RT}}$ is over 3. The subscript is used to emphasize that the estimates are obtained using RT-PCR, and not microarray, data. RT-PCR

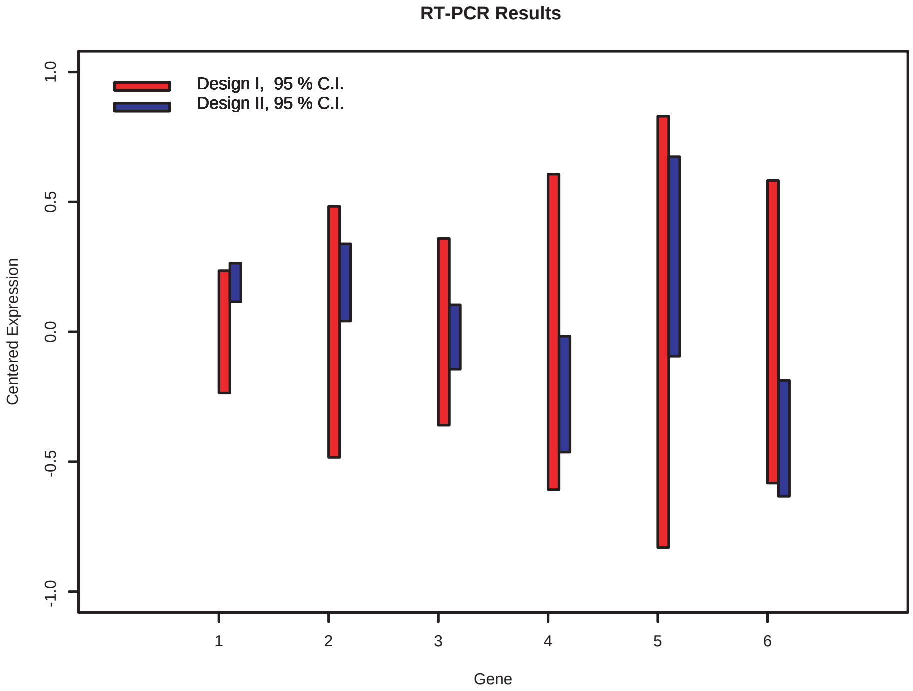

Fig. 3. RT-PCR results from six genes using designs I and II. The mRNA from five mice was obtained. For design I, gene expression was quantified via RT-PCR using the individual mRNA samples. The 95% confidence interval (CI) across the five samples is shown in light grey. For design II, the five mRNA samples were pooled and five replicate measurements were obtained from the pool. The 95% CI is shown in dark grey. Each CI is centered by the average of RT-PCR measurements from the five individual samples for each gene.

most likely has a lower technical variance ($\sigma^2_{\xi,\mathrm{RT}}$) than technologies that quantify gene expression on a large scale (e.g. oligonucleotide or spotted arrays). As such, we expect estimates of $\lambda_{\mathrm{RT}} = \sigma^2_{\epsilon,\mathrm{RT}}/\sigma^2_{\xi,\mathrm{RT}}$ to be higher in general than estimates of λ obtained using microarray technologies.

In addition to estimating λ_{RT}, these data can be used to check the assumption that pooling mRNAs from individuals does not introduce a bias (see Section 2). This seems to be the case. For each of the genes, Figure 3 shows the 95% confidence intervals for the measurements from individual (light grey) and pooled (dark grey) samples. The intervals are centered by the average of the values from the individual mRNA samples. As shown, the measurements obtained from the pooled mRNA samples are close to the average of the individual measurements, with reduced variability.

5. Discussion

Experimental design issues particular to microarray experiments have received recent attention (Kerr and Churchill, 2001a, 2001b; Black and Doerge, 2002; Yang and Speed, 2002), but the question of pooling mRNA across subjects has not yet been addressed statistically. We have considered the effects of pooling mRNA on the estimation of gene expression and have shown that under certain conditions, pooling can be advantageous in terms of cost and efficiency. The biggest advantage occurs when the biological variability is large relative to the technical variability. A formula is given which defines the total number of subjects and arrays required in a pooled experiment to achieve a confidence interval with expected squared length equal to that obtained in the no-pooling case. Totals prescribed by this formula ensure that gene expression estimates are more precise in the pooled experiment. Several assumptions are made to facilitate derivation of such a formula.

The calculations assume that mRNA abundance in individual samples averages out when pooled. This makes sense intuitively, and seems to be the case for the RT-PCR data presented here. A second assumption is that there is a linear decrease in biological variability following pooling. This assumption could not be checked using the data presented here and remains an open problem. We note that if $\epsilon_i{}' \sim N\left(0, \sigma_\epsilon^2/r_s{}^\eta\right)$, where $\eta \neq 1$, equation (2.4) will depend on η which will need to be estimated. A more critical assumption is that sufficient normalization and data processing has taken place to correct for background and remove systematic sources of variation, leaving two primary sources of variability: biological and technical. A number of background correction and normalization methods are available (for a review, see Schuchhardt *et al.*, 2000 or Yang *et al.*, 2002). The particular methods used will be in large part technology dependent; the details of this are not considered here.

A final assumption is that following normalization and perhaps transformation, the nominal level of gene expression is estimated by the average of the expression measurements. This might not be the case and distributions of measurements should be checked on a case by case basis. In cases of severe outliers, alternative estimators (e.g. median) could be considered within this framework.

In addition to the expected size of the confidence interval, one might be concerned with the variance since this impacts experimental reproducibility. For example, the list of genes identified as differentially expressed will change the least across experiments when the variance in the confidence interval lengths is minimized. We have determined the total number of subjects and arrays required in the pooled design to ensure that the variability among confidence interval lengths is no more than that from design I. The totals are conservative in that they guarantee smaller expected squared lengths and more precise expression estimates. In practice, we recommend designing experiments with a total number of subjects in the pooled design somewhere in between the totals prescribed by equations (2.4) and (3.1) (see Figure 2).

As shown in Figure 2, the totals depend on the biological and technical variability $\left(\lambda = \frac{\sigma_\epsilon^2}{\sigma_\xi^2}\right)$ and indicate that the advantages of pooling increase with increasing λ. Most likely, this ratio is gene dependent; and the best, representative, value of λ to use in a set of experiments measuring multiple genes is not yet known. A conservative approach would be to use a lower bound for λ thus ensuring efficient estimates across the array. Alternatively, one could specify the distribution governing λ and calculate the total numbers of subjects and arrays that maximize average efficiency across the array. Doing so gives a smaller number of subjects and arrays than simply calculating estimates based on maximum variability, at the possible price of loss in efficiency for some genes. This remains an open question.

Microarray data is required to estimate λ for a large number of genes and address such questions. Preliminary results using RT-PCR data are consistent with the conjectures that the ratio is gene dependent and biological variability is larger than technical variability. However, the technical variability in the RT-PCR assay is probably smaller than that in most microarray technologies and, as a result, estimates of λ_{RT} reported here are likely higher than would be observed in a microarray experiment. A study by Pritchard *et al.* (2001), using spotted microarray data, suggests that this is the case. In that work, estimates of λ were obtained for three tissues in mouse; λ was greater than one for approximately 70% of the genes.

Provided a representative value of λ can be determined, the total number of subjects and arrays can be specified and mRNA pools constructed. A single array should be used to probe each pool, and allocation of the total number of subjects to pools should be made. The calculations assume a balanced design and thus imply that an equal number of subjects be allocated to each pool. For the case considered in Section 3 with $\lambda = 4$, if $n_{\mathrm{a}2} = 10$, then $n_{\mathrm{s}2} = 20.04$ and so 'at least' 21 subjects are required. For the design to be balanced, this implies 30 subjects are required. Calculations (see Appendix) indicate that an unbalanced design can be used provided the allocation is made as close to balanced as possible. For example, with 10 arrays and 21 subjects, one could potentially construct nine mRNA pools with the mRNA from one subject and combine mRNA samples from the remaining 12 subjects into the last pool. This is not optimal. Instead, one should construct nine pools consisting of the mRNA from two subjects and one pool with the mRNA from three subjects (see Appendix for further detail).

Another major consideration in evaluation of pooled designs is that of contamination. If there is some proportion of subjects with a systematically altered level of expression that is not detected during data pre-processing, estimates of expression will be biased with increased variance (see Appendix). Much more work is required to address the problem of contamination. A simple, though not optimal, solution would be to remove pools that appear to be outliers relative to others. In cases where obtaining sufficient mRNA is not at issue, a better solution is to consider a design which allows for the mRNA from one individual to contribute to more than one pool. Such a design should allow not only for identification, but also quantification, of the bias introduced by contaminated samples. Then, estimates from pooled samples could be adjusted accordingly resulting in more accurate estimates of the nominal level of gene expression.

In summary, investigators are faced with a number of difficult questions when designing a microarray experiment. One of them is whether or not to pool mRNA across subjects. Intuitively, pooling is advantageous if the level of biological variation is high compared to technical variation on the array. We provide conditions supporting this intuition. We also quantify how to take advantage of pooling when considering experimental cost. As discussed, a number of questions remain open, but consideration of such issues should lead to more efficiently designed microarray experiments.

Appendix

In the unbalanced case, the number of subjects contributing mRNA to a pool and the number of arrays used to probe a pool could vary from pool to pool. Then,

$$x_{i,j} = \theta + {\epsilon_i}' + \xi_{i,j}$$

where ${\epsilon_i}'$ has mean zero and variance $\frac{\sigma_\epsilon^2}{rs_i}$; $\xi_{i,j}$ has mean zero and variance σ_ξ^2. Then

$$\sigma^2{}_{\bar{x}} = \left(\frac{\left(\sum_{i=1}^{n_\mathrm{p}} \frac{ra_i^2}{rs_i}\right) \sum_{i=1}^{n_\mathrm{p}} rs_i}{\left(\sum_{i=1}^{n_\mathrm{p}} ra_i\right)^2} \right) \frac{\sigma_\epsilon^2}{n_\mathrm{s}} + \frac{\sigma_\xi^2}{n_\mathrm{a}} \tag{A.1}$$

is minimized when $\frac{ra_i}{rs_i} = \frac{ra_j}{rs_j}$ for $i \neq j$. Consider the example discussed in Section 3 where $\lambda = 4$, $n_{\mathrm{a}2} = 10$, and $n_{\mathrm{s}2} = 20.04$. Table 1 shows (and equation (A.1) can be used to verify) that $E[l^2] = 6.133$ and $\sigma_{\bar{x}}^2 = 0.300$. For the unbalanced case with $n_{\mathrm{a}2} = 10$ and $n_{\mathrm{s}2} = 21$, there are a number of options for allocation of the subjects to pools. Equation (A.1) indicates that the designs should be as balanced as possible. If one constructs nine mRNA pools with the mRNA from one subject and combines mRNA samples from the remaining 12 subjects into the last pool, this gives $E[l^2] = 9.484$ and $\sigma_{\bar{x}}^2 = 0.463$. Each is inflated from the design without pooling. Instead, a design with nine pools consisting of the mRNA from two subjects and one pool with the mRNA from three subjects results in $E[l^2] = 6.004$ and $\sigma_{\bar{x}}^2 = 0.293$, an improvement over the no-pooling case.

The results considered assume that following appropriate background correction, normalization, and perhaps transformation, mRNA levels from subjects are samples from a distribution with common mean,

θ. This of course might not be the case. There could be some proportion of subjects, denoted by p, with a level of expression systematically altered by an amount η that is not detected during data pre-processing. In this case, equation (2.2) becomes

$$x_{i,j} = \theta + \epsilon_i{}' + \delta_i{}' + \xi_{i,j}$$

where

$$\delta_i = \begin{cases} \eta & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p. \end{cases}$$

Here, δ_i represents contamination of individual i and $\delta_i{}'$ represents the average amount of contamination across the pool.

For both designs, the estimators are biased and the variance is increased from the non-contamination case: $E[\bar{x}_{..}] = \theta + \eta p$ and

$$\sigma^2{}_{\bar{x},(1)} = \frac{1}{n_{\mathrm{p1}}}\left(\sigma_\epsilon^2 + \eta^2 p(1-p) + \sigma_\xi^2\right) \quad \text{and} \quad \sigma^2{}_{\bar{x},(2)} = \frac{1}{n_{\mathrm{p2}}}\left(\frac{\sigma_\epsilon^2 + \eta^2 p(1-p)}{r_{\mathrm{s2}}} + \frac{\sigma_\xi^2}{r_{\mathrm{a2}}}\right).$$

ACKNOWLEDGEMENTS

The authors thank Michael A. Newton and Joseph L. Gastwirth for critical reviews of an earlier draft and two anonymous referees for providing helpful comments.

REFERENCES

AMOS, C. I., FRAZIER, M. L. AND WANG, W. (2000). DNA pooling in mutation detection with reference to sequence analysis. *American Journal of Human Genetics* **66**, 1689–1692.

BLACK, M. A. AND DOERGE, R. W. (2002). Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**, 1609–1616.

BROWN, K. M., MACDONALD, T., COGEN, P. H., CHEN, Y. W. AND PETERSON, K., *et al.* (2001). Identification of expression changes of prognostic and therapeutic value in metastasizing medulloblastoma. *Nature Genetics* **27**, 89.

CHURCHILL, G. A. AND OLIVER, B. (2001). Sex, flies and microarrays. *Nature Genetics* **29**, 355–356.

DORFMAN, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436–440.

ERNARD, W., KHAITOVICH, P., KLOSE, J., ZOLLNER, S. AND HEISSIG, F., *et al.* (2002). Intra-and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343.

GASTWIRTH, J. L. AND HAMMICK, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *Journal of Statistical Planning and Inference* **22**, 15–27.

GASTWIRTH, J. L. (2000). The efficiency of pooling in the detection of rare mutations. *American Journal of Human Genetics* **67**, 1036–1039.

JIN, W., RILEY, R. M., WOLFINGER, R. D., WHITE, K. P., PASSADOR-GURGEL, G. AND GIBSON, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389–395.

KERR, M. K. AND CHURCHILL, G. A. (2001a). Statistical design and analysis of gene expression microarray data. *Genetical Research* **77**, 123–128.

KERR, M. K. AND CHURCHILL, G. A. (2001b). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

PHEIFFER, R. M., RUTTER, J. L., GAIL, M. H., STRUEWING, J. AND GASTWIRTH, J. L. (2002). Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genetic Epidemiology* **22**, 94–102.

PRITCHARD, C. C., HSU, L., DELROW, J. AND NELSON, P. S. (2001). Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences* **98**, 13266–13271.

SCHUCHHARDT, J., BEULE, D., MALIK, A., WOLSKI, E., EICKHOFF, H., LEHRACH, H. AND HERZEL, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Researc* **28**, e47.

SOTIRIOU, C., CHAND, K., PETERSEN, D., JAZAERI, A. A. AND LIU, E. T. (2001). Core biopsy versus surgical tumor specimens for microarray analysis of gene expression profiles. *Nature Genetics* **27**, 88–89.

WARING, J. F., JOLLY, R. A., CIURLIONIS, R., LUM, P. Y. AND PRAESTGAARD, J. T., *et al.* (2001). Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicology and Applied Pharmacology* **175**, 28–42.

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. AND SPEED, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, 1–10.

YANG, Y. H. AND SPEED, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Genetics Reviews* **3**, 579–588.

[*Received October 21, 2002; revised January 11, 2003; accepted for publication January 16, 2003*]