Supplementary Material for: Sample Size Determination in Microarray Experiments for Class Comparison and Prognostic Classification

1 Sample size calculation for single-channel array

For simplicity, we assume samples coming from two classes. To simplify the presentation, we switch notation, so that $Y_{g1fs} = z_{fs}$ and $Y_{g2fs} = w_{fs}$. The normalized, background-adjusted log-intensities, z_{ij} and w_{kl}, are described by the model:

$$
\begin{aligned}
z_{ij} &= \mu_x + x_i + \epsilon_{ij}, i = 1...n, j = 1...m \\
x_i &\sim Normal(0, \tau_x^2) \\
\epsilon_{ij} &\sim Normal(0, \sigma^2) \\
w_{kl} &= \mu_y + y_k + \epsilon_{kl}, k = 1...n, l = 1...m \\
y_k &\sim Normal(0, \tau_y^2) \\
\epsilon_{kl} &\sim Normal(0, \sigma^2)
\end{aligned}
$$

where all the x_i, y_k, ϵ_{ij} and ϵ_{kl} are independent. We wish to test $\mu_x = \mu_y$ against the alternative $\mu_x \neq \mu_y$. The log-likelihood is

$$
ll = -\frac{n}{2} log \left(\left\| \widehat{\Sigma_x} \right\|^{-m/2} \right) -
$$

$$
\begin{aligned}
&\frac{1}{2}\sum_i (Z_i - \mu_x J_{m,1})^T \Sigma_x^{-1} (Z_i - \mu_x J_{m,1}) \\
&-\frac{n}{2} log \left(\left\| \widehat{\Sigma_y} \right\|^{-m/2} \right) \\
&-\frac{1}{2}\sum_i (W_i - \mu_y J_{m,1})^T \Sigma_y^{-1} (Z_i - \mu_y J_{m,1})
\end{aligned}
$$

where $J_{m,1}$ is a vector of 1's of length m, Z_i is the vector $(z_{i1}, ..., z_{im})^T$, and similarly W_i, and Σ_x and Σ_y are the compound symmetric covariance matrices corresponding to the Z's and W's.

Assuming that the variance parameters are known, then the covariance matrices are known and can be shown to be compound symmetric. Then the transformations

$$
\begin{aligned}
Z_i^* &= \Sigma_x^{-\frac{1}{2}} Z_i \\
W_i^* &= \Sigma_y^{-\frac{1}{2}} W_i
\end{aligned}
$$

result in covariance matrices of the form

$$
\begin{aligned}
cov(Z_i^*) &= I \\
cov(W_i^*) &= I
\end{aligned}
$$

where I is the identity matrix. Now if we can write the hypothesis test in a simple form in this transformed space, then approximate sample size calculations will be relatively simple functions of the variance parameters, which we can then transform back into the space of the original problem to compare with the sample size formulas for the fixed-effects model.

The following seven theorems comprise the formula derivation.

Theorem 1 *The covariance matrices of the Z_i and the W_k are compound symmetric*

Proof:

$$\begin{aligned}
var(z_{ij}) &= \tau_x^2 + \sigma^2 \\
cov(z_{i1}, z_{i2}) &= cov(\mu_x + x_i + \epsilon_{i1}, \mu_x + x_i + \epsilon_{i2}) \\
&= var(x_i) \\
&= \tau_x^2 \\
var(w_{kl}) &= \tau_y^2 + \sigma^2 \\
cov(w_{k1}, w_{k2}) &= \tau_y^2
\end{aligned}$$

By symmetry, what holds for $j = 1$ and $j = 2$ holds for all $j_1 \neq j_2$, and all i; and similarly for $l = 1$ and $l = 2$ holds for all $l_1 \neq l_2$, and all k. Therefore, the covariance matrices are

$$\begin{aligned}
cov(Z_i) &= I\sigma^2 + J_{m,m}\tau_x^2 \\
cov(W_k) &= I\sigma^2 + J_{m,m}\tau_y^2
\end{aligned}$$

which is compound symmetric. *Q.E.D.*

Theorem 2 *The inverse of a matrix of the form $I_m\alpha^2 + J_{m,m}\beta$, if it exists, is $I\frac{1}{\alpha^2} - J_{m,m}\frac{\beta}{\alpha^2(\alpha^2+m\beta)}$.*

Proof:

$$\begin{aligned}
&(I_m\alpha^2 + J_{m,m}\beta) \cdot (I_m\frac{1}{\alpha^2} - J_{m,m}\frac{\beta}{\alpha^2(\alpha^2+m\beta)}) \\
&= I_m + J_{m,m}\frac{\beta}{\alpha^2} - J_{m,m}\frac{\beta}{\alpha^2+m\beta} - J_{m,m}\frac{m\beta^2}{\alpha^2(\alpha^2+m\beta)} \\
&= I_m + J_{m,m}\frac{\beta(\alpha^2+m\beta) - \beta\alpha^2 - m\beta^2}{\alpha^2(\alpha^2+m\beta)} \\
&= I_m
\end{aligned}$$

Q.E.D.

Theorem 3 *If* $\Sigma_z = cov(Z_i)$ *and* $\Sigma_w = cov(W_k)$, *then*

$$\begin{aligned}
\Sigma_z^{-1} &= I_m \frac{1}{\sigma^2} - J_{m,m} \frac{\tau_x^2}{\sigma^2(\sigma^2 + m\tau_x^2)} \\
\Sigma_w^{-1} &= I_m \frac{1}{\sigma^2} - J_{m,m} \frac{\tau_y^2}{\sigma^2(\sigma^2 + m\tau_y^2)}
\end{aligned}$$

if the inverses exist.

Proof: This follows directly from the previous two theorems. *Q.E.D.*

Theorem 4 *If* $\Sigma = I_m\alpha^2 + J_{m,m}\beta$, *then* $\Sigma^{\frac{1}{2}} = I_m\alpha + J_{m,m}\frac{\sqrt{m\beta+\alpha^2}-\alpha}{m}$

Proof:

$$\begin{aligned}
&\left(I_m\alpha + J_{m,m}\frac{\sqrt{m\beta+\alpha^2}-\alpha}{m}\right) \cdot \left(I_m\alpha + J_{m,m}\frac{\sqrt{m\beta+\alpha^2}-\alpha}{m}\right) \\
&= I_m\alpha^2 + 2\alpha J_{m,m}\frac{\sqrt{m\beta+\alpha^2}-\alpha}{m} + mJ_{m,m}\left[\frac{\sqrt{m\beta+\alpha^2}-\alpha}{m}\right]^2 \\
&= I_m\alpha^2 + J_{m,m}\frac{2\alpha\sqrt{m\beta+\alpha^2}-2\alpha^2}{m} + J_{m,m}\frac{m\beta+\alpha^2-2\alpha\sqrt{m\beta+\alpha^2}+\alpha^2}{m} \\
&= I_m\alpha^2 + J_{m,m}\beta
\end{aligned}$$

Q.E.D.

Theorem 5 *If the inverses exist, then* $\Sigma_z^{-\frac{1}{2}} = I\frac{1}{\sigma} + J_{m,m}\frac{1}{m}\left(\frac{1}{\sqrt{\sigma^2+m\tau_x^2}} - \frac{1}{\sqrt{\sigma^2}}\right)$ *and* $\Sigma_w^{-\frac{1}{2}} = I\frac{1}{\sigma} + J_{m,m}\frac{1}{m}\left(\frac{1}{\sqrt{\sigma^2+m\tau_y^2}} - \frac{1}{\sqrt{\sigma^2}}\right)$

Proof:

Let

$$\begin{aligned}\alpha^2 &= \frac{1}{\sigma^2} \\ \beta &= \frac{-\tau_x^2}{\sigma^2(\sigma^2 + m\tau_x^2)}\end{aligned}$$

then

$$\begin{aligned}\alpha &= \frac{1}{\sigma} \\ \frac{1}{m}(\sqrt{\alpha^2 + m\beta} - \alpha) &= \frac{1}{m}\left(\sqrt{\frac{1}{\sigma^2} - \frac{m\tau_x^2}{\sigma^2(\sigma^2 + m\tau_x^2)}} - \frac{1}{\sigma}\right) \\ &= \frac{1}{m}\left(\frac{1}{\sqrt{\sigma^2 + m\tau_x^2}} - \frac{1}{\sqrt{\sigma^2}}\right)\end{aligned}$$

The proof is similar of $\Sigma_w^{-\frac{1}{2}}$. *Q.E.D.*

Theorem 6 *The distribution of Z_i^* and W_i^* are $Normal(\mu_x/(\sigma^2 + m\tau_x^2)J_{m,1}, I)$ and $Normal(\mu_y/(\sigma^2 + m\tau_y^2)J_{m,1}, I)$ respectively.*

Proof: Recall that $Z_i^* = \Sigma_x^{-\frac{1}{2}} Z_i$, and $W_i^* = \Sigma_y^{-\frac{1}{2}} W_i$.

$$\begin{aligned}E\left[\Sigma_x^{-\frac{1}{2}} Z\right] &= \Sigma_x^{-\frac{1}{2}} J_{m,1}\mu_x \\ &= J_{m,1}\frac{\mu_x}{\sigma} + J_{m,1}\mu_x\left[\frac{1}{\sqrt{\sigma^2 + m\tau_x^2}} - \frac{1}{\sigma}\right] \\ &= J_{m,1}\frac{\mu_x}{\sqrt{\sigma^2 + m\tau_x^2}}\end{aligned}$$

The mean vector for the W_i^* follows from a similar calculation. The covariance matrices in both cases are the identity matrix by construction. *Q.E.D.*

Theorem 7 *If $\tau = \tau_x = \tau_y$, then the sample size for achieving power $1 - \beta$ at distance δ for a two-sided test at level α of the null hypothesis $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x \neq \mu_y$ is*

$$n = 4\left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\frac{\sigma^2}{m} + \tau^2\right).$$

Proof: First, reduce by sufficiency.

$$\begin{aligned}
\bar{Z}_i^* &\sim Normal\left(\frac{\mu_x}{\sqrt{\sigma^2 + m\tau^2}}, \frac{1}{m}\right) \\
\bar{W}_k^* &\sim Normal\left(\frac{\mu_y}{\sqrt{\sigma^2 + m\tau^2}}, \frac{1}{m}\right) \\
\sqrt{\sigma^2 + m\tau^2}\bar{Z}_i^* &\sim Normal\left(\mu_x, \tau^2 + \frac{\sigma^2}{m}\right) \\
\sqrt{\sigma^2 + m\tau^2}\bar{W}_k^* &\sim Normal\left(\mu_y, \tau^2 + \frac{\sigma^2}{m}\right)
\end{aligned}$$

The required number of distinct samples n, assuming the variance parameters are known, is then

$$n = 4\left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau^2 + \frac{\sigma^2}{m}\right)$$

as desired. *Q.E.D.*

This sample size formula will only be approximate, because in reality the variance parameters are not known exactly and the distribution of the test statistic under the alternative hypothesis is non-central t, not normal.

2 Impact of correlation among genes on false-positive rate

Suppose that there are a total of G genes, and that an hypothesis test is carried out at the α significance level for each gene. For example, if there are two varieties of interest, numbered variety 1 and variety 2, then the test of the null hypothesis $H_0 : VG_{1g} = VG_{2g}$ versus the alternative $H_1 : VG_{1g} \neq VG_{2g}$ may be carried out for each gene g to identify those that are differentially expressed in the varieties.

Define the a collection of indicator variables d_g such that, if the hypothesis test for gene g is significant, then $d_g = 1$, otherwise, $d_g = 0$. The number of significant hypothesis tests is then $\sum_g d_g$. The expected number of significant tests is then by linearity of expectation

$$\begin{aligned} E\left[\sum_g d_g\right] &= \sum_g E\left[d_g\right] \\ &= G\alpha \end{aligned}$$

where G is the number of genes. Recall that linearity of expectation holds even when the variables are not independent.

While the expectation of the number of false positives will remain unaffected by the correlation among the genes, the distribution of the number of false positives will be affected by the correlation structure.

3 Sample size comparisons for technical replicates in a reference design

Fix the total number of arrays available at n. We compare a design that uses no technical replicates to a design that uses r technical replicates for each sample.

$$\begin{aligned} y_{gadvf} &= G_g + GA_{ga} + GD_{gd} + GV_{gv} + (GF)_{gf(v)} + \epsilon_{gadvf} \\ GF_{gf} &\sim N(0, \kappa^2) \end{aligned}$$

$$\epsilon \sim N(0, \lambda^2)$$

Under this model, the variance of the log-ratios in a reference design is $\kappa^2 + 2\lambda^2$. The variance of an average of r technical replicate log-ratios is $\kappa^2 + \frac{2\lambda^2}{r}$.

Technical replicates can be handled by averaging the replicate measurement log-ratios, then using the averages as if they were the original log-ratios. If there are r replicates per sample, this results in n/r total averages, one for each non-reference sample. Hence the efficiency of the class mean estimate for a technical replicate design is

$$\frac{\kappa^2 + \frac{2\lambda^2}{r}}{n/r} = \frac{r\kappa^2 + 2\lambda^2}{n}.$$

For example, suppose n_1 arrays are required to achieve a specified power or efficiency when no technical replicates are used. Then let n_r be the number of arrays required to achieve that same power or efficiency when r technical replicates per sample are performed. Then the two required sample sizes are related by the equation:

$$\begin{aligned} \frac{n_r}{n_1} &= \frac{r\kappa^2 + 2\lambda^2}{\kappa^2 + 2\lambda^2} \\ &= \frac{r(\kappa/\lambda)^2 + 2}{(\kappa/\lambda)^2 + 2} \end{aligned}$$

For instance, we have observed $(\kappa/\lambda)^2$ with median approximately 7.3 in one dataset. Hence, if $n_1 = 20$ arrays are required with a single replicate on each array, it would require $n_2 = 36$ arrays and 18 samples with two technical replicates per sample, and $n_3 = 51$ arrays and 17 samples with three technical replicates per sample, to achieve the same power and efficiency.

4 Sample size comparison formula for a paired samples design with technical replicates

The analysis of variance model for paired data requires a conceptual shift in the parameter interpretations (reference to our dye swap paper and the appendix thereto). In this paper one sees that several models can be fit to paired data, and the most appropriate one will depend on the level and extent of the replication.

For the model we think most likely to be appropriate in practice, the variance of the contrast of interest for a design which is balanced with respect to the dyes, and no technical replicates is

$$\frac{\tau_g^2 + 2\sigma_g^2}{n_{balanced}}$$

where τ_g^2 is the variance in the effect of the cancer on the expression level of the gene g in the population of interest, and σ_g^2 is the error term associated with gene g. The variance for a model with one technical replicate or one dye swap technical replicate for each array is

$$\frac{2\tau_g^2 + 2\sigma_g^2}{n_{dyeswap}}.$$

5 Sample size for pooling

For the reference design, recall the model for non-pooled dual-label data was

$$log\left(Y_{gadvf}^{unpooled}\right) \quad = \quad G_g + GA_{ga} + GD_{gd} + GV_{gv} + (GF)_{gf(v)} + \epsilon_{gadvf},$$

If we pool k samples, and make the simplifying assumption that each sample is equally represented in the pool and that the effect of pooling is additive on the log scale, then for pool p (for example), which is made up of samples $1...k$, we will have

$$log\left(Y_{gadvp}^{pooled}\right) \;=\; G_{g}^{\{2\}} + GA_{ga}^{\{2\}} + GD_{gd}^{\{2\}} + GV_{gv}^{\{2\}} + \frac{1}{k}\sum_{f=1}^{k}(GF)_{gf(v)}^{\{2\}} + \epsilon_{gadvp}^{\{2\}}.$$

Hence, what will change in the model is GF terms. Noting that $var\left[\frac{1}{k}\sum_{f=1}^{k}(GF)_{gf(v)}\right] = \frac{\tau^2}{k}$, we see that a reduction in the biological variation associated with the pooled sample is the only substantive change in the model. Therefore, sample size calculations will only be affected though this change.

6 A bijective function relating the terms in a single-label model to the terms in a dual-label model

Recall that the single-label model is

$$log\left(Y_{gvf}^{\{1\}}\right) \;=\; G_{g}^{\{1\}} + GV_{gv}^{\{1\}} + (GF)_{gf(v)}^{\{1\}} + \epsilon_{gvf}^{\{1\}}.$$

Also recall that the dual-label model is

$$log\left(Y_{gadvf}^{\{2\}}\right) \;=\; G_{g}^{\{2\}} + GA_{ga}^{\{2\}} + GD_{gd}^{\{2\}} + GV_{gv}^{\{2\}} + (GF)_{gf(v)}^{\{2\}} + \epsilon_{gadvf}^{\{2\}},$$

which, in terms of the log-ratio for a single array containing, say, sample 1 from variety 1, is

$$\log\left(Y_{gadv1}^{\{2\}}/Y_{gadv0}^{\{2\}}\right) \;=\; GD_{g1}^{\{2\}} - GD_{g2}^{\{2\}} + GV_{g1}^{\{2\}} - GV_{g0}^{\{2\}} + (GF)_{g1(v)}^{\{2\}} - (GF)_{g0(v)}^{\{2\}} + \epsilon_{gadv1}^{\{2\}} - \epsilon_{gadv0}^{\{2\}}$$

All log-ratios in a dual-label reference design will have the $GD_{g1} - GD_{g2}$ term. This represents the difference between the dyes with respect to dye incorporation.

Since the single-dye array has only one dye, the impact of this dye is contained in the G_g term of that model.

Note that the Bijective functions

$$\begin{array}{rcl} G_g^{\{1\}} & \Leftrightarrow & GD_{g1}^{\{2\}} - GD_{g2}^{\{2\}} \\ GV_{gv}^{\{1\}} & \Leftrightarrow & GV_{g1}^{\{2\}} - GV_{g0}^{\{2\}} \\ GF_{gf}^{\{1\}} & \Leftrightarrow & GF_{g1}^{\{2\}} - GF_{g0}^{\{2\}} \\ \epsilon_{gadvf}^{\{1\}} & \Leftrightarrow & \epsilon_{gadv1}^{\{1\}} - \epsilon_{gadv0}^{\{2\}} \end{array}$$

form a bijective map from the single-dye probability model to the dual-dye probability model.

Finally, recall that the correlation structure among the log-ratios in the reference design was shown to have no impact on the statistical inference procedures we considered.

Therefore, we have shown that the two models are equivalent.

7 Table showing performance of FDR estimate

π	α	$\mathbf{1-\beta}$	**MC FDR: Mean(SD)**	$\widehat{\mathbf{E}}[\mathbf{FDR}]$
.005	.001	.95	.166(.045)	.17
.005	.001	.90	.175(.047)	.18
.005	.001	.80	.188(.049)	.20
.05	.001	.95	.020(.0058)	.02
.05	.001	.90	.021(.0072)	.02
.05	.001	.80	.022(.0064)	.02
.20	.001	.95	.0042(.0014)	.004
.20	.001	.90	.0045(.0014)	.004
.20	.001	.80	.0051(.0016)	.005
.005	.01	.95	.673(.023)	.68
.005	.01	.90	.679(.025)	.69
.005	.01	.80	.705(.025)	.71
.05	.01	.95	.165(.014)	.17
.05	.01	.90	.168(.013)	.17
.05	.01	.80	.189(.017)	.19
.20	.01	.95	.040(.0050)	.04
.20	.01	.90	.041(.0045)	.04
.20	.01	.80	.047(.0048)	.05
.005	.005	.95	.512(.040)	.51
.005	.005	.90	.510(.034)	.53
.005	.005	.80	.549(.039)	.55
.05	.005	.95	.090(.011)	.09
.05	.005	.90	.094(.012)	.10
.05	.005	.80	.101(.012)	.11
.20	.005	.95	.020(.0035)	.02
.20	.005	.90	.021(.0031)	.02
.20	.005	.80	.023(.0034)	.02

Table 1: Monte Carlo (MC) simulation of the false discovery rate (FDR). Each line represents 100 MC simulations. $\widehat{\mathbf{E}}[\mathbf{FDR}]$ is the estimated FDR using the estimation formula in the text. α is the significance level, $1-\beta$ is the power, and π is the true number of differentially expressed genes.

8 Review of other microarray sample size papers

Most of the papers addressing sample size issues have focused on determining an adequate sample size for comparing individual specimens, without making statistical inference to larger populations. For instance:

- Lee et al. (2000) and Black and Doerge (2002) discuss methods for determining the number of replicate spots needed on an array when the goal of the experiment is to compare two individual labeled cDNA samples;
- Pan et al. (2002) discuss computational algorithms to determine the number of (technical) replicate arrays required for nonparametric analyses when the goal is to compare two individual RNA samples adequately;
- Wolfinger et al. (2001) present power plots for comparing individual samples in a mixed effects analysis of variance model;
- Lee and Whitmore (2002) present a Bayesian approach to sample size and power for comparing individual specimens.

Other investigators have suggested methods for sample size determination for linear discrimination analysis (Hwang et al., 2002), where the goal is to test the global hypothesis that the classes do not differ with regard to expression of any genes. We have presented some sample size formulas for statistical inference in microarray studies for particular experimental situations (Simon et al., 2002; Dobbin et al., 2003a; Dobbin et al., 2003b), where the goal is to make statistical inference to populations of specimens. Zien et al. (2003) used a model similar to the one we proposed (Dobbin and Simon, 2002) to develop a simulation program that provides estimates of sensitivity and specificity in response to a given set of user inputs, one of which is number of samples in each group.

9 Design choice discussions

A simple reference design places sub-samples from a single reference RNA sample on each array, tagged with the same dye. Suppose we have a simple reference design experiment comparing two groups. In general, if the goal is to compare two groups, we recommend a balanced block design, in which no reference RNA is used and each pair of classes appears together the same number of times over the arrays (Scheffé, p. 161). But there are reasons one might prefer to use a reference design. Because spot effects, which are generally fairly large, are confounded with gene-specific sample effects (i.e., biological variation), comparing samples on different arrays directly becomes problematic in a balanced block design. This confounding has a number of implications: 1) cluster analysis of samples may be impossible unless one ignores the spot effects or introduces additional model assumptions, neither of which seems advisable (Dobbin and Simon, 2002); 2) similarly, it may be impossible to construct a class predictor out of a subset of genes; or, 3) to develop a multi-gene prognostic marker, because of this confounding; 4) normalizing data gathered at different times or in different experiments will also be very problematic with a balanced block design. Another problem with a balanced block design is that if there is more than one way to classify the samples, then designing the experiment so that each potential classification can be made efficiently may be difficult or impossible. In sum, if one is considering an analysis that goes beyond just identifying differentially expressed genes, then a reference design will often be preferable. Hence, a sample size formula for comparing classes in a reference design may be of some interest.

In general, we (Dobbin et al. 2003), and others (Liang et al., 2003), don't recommend using dye-swaps (i.e., running the same samples on two different arrays, one with the dye labeling reversed) in a reference design unless a goal is comparison of experimental samples to the reference. For comparisons of classes of non-reference samples, dye-swap arrays do not remove any bias from the comparisons and generally represent replication at the wrong level, and hence result in a loss of efficiency for the comparisons. On the other hand, some technical replicates may be informative to assure that the assay is reproducible.

The motivation for having more than one label in a spotted microarray system is that variability in log intensity corresponding to spots on different arrays is greater than variability within a spot, from one labeled sample to the other. This is because of variability in spot size for printed arrays and because the labeled sample is not distributed uniformly across the surface of the array. This situation has an analogue in agricultural experiments, where often the variability between different tracts of land, due to location, weather, soil conditions, etc., is greater than the variability within a tract of land. The tracts of land, which are analogous to the spots on the microarray, are referred to as blocks, and the resulting experimental design is called a block design. One essentially "blocks out" the effect of the spots/tracts in such an experiment by basing all inference on within-spot/tract comparisons. When comparing several classes with this type of block experiment, it is a well-established fact in the agricultural model that a balanced block design, in which samples from each pair of varieties appear together on an array the same number of times, produces the greatest efficiency and power. The model for microarray data is somewhat different than the agricultural analogue, but we have shown that it is still true that the balanced block design is the most efficient and powerful, and can be a significant improvement in these respects over other designs (Dobbin and Simon, 2002; Dobbin et al., 2003).

The goal of a paired design is typically to identify genes expressed differently within each pair, e.g., genes whose expression level is affected by the treatment, or the presence of the tumor. For paired samples and dual-label microarrays, putting one member from each pair together on each array is a natural design that is also most efficient. In this type of design, each individual serves as their own control, which can effectively eliminate much biological variation from the comparison of interest, increasing precision. We have recommended balancing the pairs with respect to the dyes in order to eliminate the dye bias from the comparisons; that is, if the sample pairs consist of normal and tumor tissue from the same individual, then half the normal tissue would be tagged with Cy3 and the other half with Cy5. Researchers have also run each sample pair twice, once with each dye labeling, to eliminate the dye bias (Boer et al., 2001; Lossos et al., 2002). Such a design is sometimes called a complete dye-swap design.

Kendziorski et al. (2003) study the efficiency of pooling as a function of the ratio of the biological variation to the technical measurement error variation, similar to our investigation into the effects of this ratio on experimental design efficiency (Dobbin and Simon, 2002).

There are practical limits to the applicability of the sample size formula for pooled samples. For instance, it is only correct under the assumption that each sample is equally represented in the pool and the pool is homogeneous; but as the number of samples increases, then intuitively the validity of these assumptions becomes more questionable. And, in general, the more samples are pooled together the more things that can go wrong. For instance, there is a potential quality control issue because one must be sure that all the contributing RNA to the pool is of good quality, since problems with a particular RNA sample may not be detectable after pooling, and this could lead to erroneous inference. Also, if one is combining a large number of samples, then the different RNA may interact in unexpected ways that would result in non-additive effects on the expression level of individual genes.

Sometimes investigators with a limited number of samples and arrays ask whether, given a fixed number of arrays, pooling some of the samples together and replicating the pooled samples over the arrays may be preferable to assigning each sample to a different array. For instance, if one has 24 samples and plans to run 24 arrays, one can either run one sample per array, or pool pairs of samples together and run each pool twice (technical replication), or pool triples of samples together and run each triple three times, etc. In fact, the one sample per array strategy is best, because one gains no efficiency by pooling together samples rather than assigning each to its own array. But one will lose degrees of freedom for error, and therefore power for detecting differential expression.

In conclusion, if one is not forced to pool for technical reasons (e.g., insufficient RNA from each individual is available for the microarray), then it appears that pooling will probably rarely be a good idea. The tradeoff between the reduction in arrays required under a pooled design will rarely be worth the increase in the number of samples required, particularly when one factors in the potential loss of robustness and power that comes with the pooled design.