Supplement to Statistical Design of Reverse Dye Microarrays

Dobbin, K.*, Shih, J. and Simon, R.

National Cancer Institute
Biometric Research Branch
6130 Executive Blvd., MSC 7434
Bethesda, MD 20892
USA

Email: dobbinke@mail.nih.gov
Phone: 301-451-6244

*To whom correspondence should be addressed.

1 Discussion of Paired Samples ANOVA Elements

G_g Represents the average expression level for the gene g in the population.

GA_{ga} Represents variation in the spots for a particular gene on different arrays.

GD_{gd} Represents the gene-specific dye effect.

GV_{gv} Represents differences in average expression for a particular gene between the two varieties (normal and cancerous).

GP_{gp} Represents variation in gene expression of this particular gene in normal tissue between different individuals.

GVP_{gvp} Represents variation in the effect of the cancer on the expression of this particular gene in different individuals.

For an individual spot on a particular array, this model postulates that the observed background-adjusted, normalized log-intensity is a result of additive effects of the amount of RNA in the sample, the size and quality of the spot, the dye effects, and random error. Included in the random error are inhomogeneities in the RNA sample and technical issues in the measurement, extraction, and reverse-transcription and labelling reactions.[1]

We assume that the cancer does not have exactly the same affect on the expression level of a gene for all individuals in the population. The GVP_{gvp} terms represent this variation (on average, for this individual or subgroup of individuals). In order to estimate terms of interest, it is necessary to assume this effect is normal with some variance τ_g^2, or to introduce a set of constraint equations which these terms satisfy. We generally favor the former approach. The error term is assumed normally distributed with mean zero and variance σ_g^2.

For each gene, an ANOVA model is fit. The term of interest is the contrast between the expression level for "cancer" and for "normal," represented by $GV_{g1} - GV_{g0}$. The purpose of the GP and GVP terms are not to estimate the effects, which are generally not of interest, but to account for participant effects when there are repeated samples for several participants. Including the GVP term in the ANOVA estimation may reduce the residual error variance by eliminating the contribution from τ_g^2, but on the other hand will also result in a loss of degrees of freedom for error. More importantly, including GVP in the model will produce less efficient estimates of the $GV_{g1} - GV_{g0}$ term of interest (see Further Discussion section below).

The analysis of variance table for these data is given in Table A of the supplement. We present three ANVOA tables because there are often very few or no degrees of freedom for estimating the

[1] Our model assumes a single RNA extraction for each sample.

sample-specific effects GP and GVP. In the left column is the ANOVA when GP and GVP are included in the model; the middle column excludes GVP; and the right column excludes both GP and GVP. We derive results for all three analyses. In fact, a single design appears optimal for all cases. We think it a reasonable plan to lump the GVP effects in with error (middle column) in order to improve power and efficiency for the term of interest.

When $k = 0$, the GP effects are no longer estimable and should be removed from the model (rightmost column of supplement Table A). Because the GP effects cancel out of the within-array contrasts on which the GV estimator is based, no assumption about the distribution of these effects is implied when the term is removed from the model.

2 Results for Paired Samples

Consider first the case with GVP effects in the model. In Appendix A, the minimum variance linear unbiased estimator of $GV_{g1} - GV_{g0}$ under the usual model constraints is derived, and the variance of the estimator is shown to be

$$var(\widehat{GV}_{g1} - \widehat{GV}_{g0}) = \frac{2\sigma_g^2}{m}.$$

For fixed m, this function is a constant in k.

Now we turn to the case without GVP effects in the model. In Appendix B, Lagrange multipliers are used to derive the minimum variance linear unbiased estimator of $GV_{g1} - GV_{g0}$ under the usual model constraints, and the variance of the estimator is shown to be

$$var(\widehat{GV}_{g1} - \widehat{GV}_{g0}) = \frac{(\tau_g^2 + \sigma_g^2)(\tau_g^2 + 2\sigma_g^2)}{(m-k)\tau_g^2 + m\sigma_g^2}.$$

For a fixed number of arrays, m, the k that minimizes the variance is $k = 0$. This equation is also valid for $k = 0$, in which case, GP terms are not estimated explicitly. Note that if $k > 0$ and neither GVP nor GP is in the model, then the variance will be greater than given by this equation, so that again $k = 0$ is optimal. When $k = 0$, we don't need to estimate τ_g^2 explicitly for inference, but only the sum $\tau_g^2 + 2\sigma_g^2$, which is the residual variance of the log-ratios.

3 Sample Size for Optimal Paired Samples Design (no reference)

Setting $k = 0$ (and removing GVP and GP from the model) results in

$$var\left(\widehat{GV}_{g1} - \widehat{GV}_{g0}\right) = \frac{\tau_g^2 + 2\sigma_g^2}{m}.$$

Consider the null hypothesis $H_0 : GV_{g1} = GV_{g0}$. Let α be the size of the test, i.e., the probability of a type I error; and $1-\beta$ be the desired power of the test against the specific alternative $|GV_{g1} - GV_{g0}| \geq \delta$. Then the sample size formula is:

$$m_0 = \left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_g^2 + 2\sigma_g^2\right).$$

Here $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the Gaussian distribution. A sample size greater than or equal to m_0 will ensure power of $1-\beta$ against the specific alternative. The quantity $\left(\tau_g^2 + 2\sigma_g^2\right)$ is the population variation in the differences between tumor and normal expression in individuals (i.e., between the two channels on the slides), and some estimate of the variance can be used in the calculation.

4 Sample Size for Optimal Unpaired Samples Design (no reference)

Setting $k = 0$ results in $var\left(\widehat{GV}_{g1} - \widehat{GV}_{g2}\right) = \frac{\tau_{g1}^2 + \tau_{g2}^2 + 2\sigma_g^2}{m}$. Let α be the size of the test, i.e., the probability of a type I error; and $1-\beta$ be the desired power of the test against the specific alternative $|GV_{g1} - GV_{g2}| \geq \delta$. Then the usual computation produces the sample size formula

$$m_0 = \left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_{g1}^2 + \tau_{g2}^2 + 2\sigma_g^2\right).$$

A sample size greater than or equal to m_0 will ensure power of $1-\beta$ against the specific alternative. The quantity $\left(\tau_{g1}^2 + \tau_{g2}^2 + 2\sigma_g^2\right)$ is the population variation in the log-ratios, and some estimate of the variance can be used in the calculation.

5 Sample Size for Optimal Reference Design When Comparison With Reference is Primary Goal

For a size α test of equality of mean expression, the sample size required to assure power $1-\beta$ against the alternative that the means differ by δ is:

$$m_0 = \left[\frac{z_{\alpha/2} + z_\beta}{\delta}\right]^2 \left(\tau_g^2 + 2\sigma_g^2\right).$$

Here $\tau_g^2 + 2\sigma_g^2$ is the variance in the log ratios, and some estimate of the variance can be used in the calculation.

6 Further Discussion

There are some drawbacks to running half the samples forward and the other half backward. Analysis software is required that performs the analysis of variance to estimate the gene-specific dye biases and adjust the variety by gene effects accordingly. This adjustment would usually not be necessary if all samples were run forward, or all run both forward and backward and the average of the log-ratios were entered. Also, if one wishes to run a cluster analysis on a set of paired samples, then one will need to adjust for dye bias before doing the cluster analysis; otherwise, dye bias may create erroneous clusters. We have noticed with one set of paired samples that there was some separation between the forward arrays and backward arrays in the multi-dimensional scaling plot when no dye adjustment was made. Adjusting for gene-specific dye bias using the model-based methods discussed in this paper appeared to correct this problem. Finally, if one has only a small number of samples available, and enough RNA to run each sample several times, then running multiple arrays for each sample will always be preferable to running a single array for each sample.

In the unpaired case, we have represented the distribution of the log expression level for each gene among individuals of the same variety (or phenotype) as normally distributed. In the paired case, we have made a similar distributional assumption about the differences between gene expression in the pairs. In both cases, each gene is allowed to have it's own level of variation within the population, represented by the τ_g^2 parameters, and we have made no assumption about the correlation structure among genes. We believe normality on the log scale within phenotype to be a reasonable assumption that will be approximately true for most genes, but one that may be violated for some genes. But without some assumption on the distribution the designs cannot be compared. It seems better to base design decisions on assumptions which are likely to be approximately true for most genes than allow concern over a few potential rogue genes to undermine the microarray design process.

One might question our characterization of dye bias. We have represented gene specific dye effects by the addition of a GD interaction term to the linear model. The GD interaction is the same for all arrays and for all samples. We assume that this type of dye effect is caused by specific characteristics of the RNA transcripts which facilitate or impede the incorporation of one of the dyes (i.e., dUTP fluorescently labelled with bulky dye adducts) during reverse transcription into cDNA. One might be concerned that this characterization of the dye by gene interaction is incorrect, and that in fact the interaction effect varies among samples. If one is concerned about a GDP (or GDF) interaction, one can partially test for the presence of this effect by performing a Tukey single degree of freedom test for additivity (Scheffé, 1999) on those sample pairs that have been run both forward and backward. This test rejects if there is evidence against the assumption of no interaction, in our case the assumption $GDP \equiv 0$ for this gene. The power of the test is not known and it only tests for a particular form of interaction, but may still be informative. The statistical significance of the findings can then be checked with a permutation test. We ran the test on paired cancer data (unpublished data) with 5759 genes, where 18 arrays had been run both forward and

reverse, then performed a permutation test. The sample data tended to have a larger proportion of genes with significant GDP interactions than the data with labels randomly permuted. A third of the genes had more significant GDP interaction F-test statistics than all 200 permutations. But these effects tended to be very small. Out of more than 100,000 estimated GDP interactions, only 10 reflected a fold change of 2 or more. Also, 9 of the 10 with large fold change were associated with the same sample, suggesting that some particular technical issues with the sample could explain the larger effects. The median fold change was 1.01. Since small fold changes, even if statistically significant, are often considered to be artifacts of the microarray technology, these small interaction effects should not be a major concern.

For paired samples, including GVP in the model produces a loss of efficiency in the contrast of interest, $\widehat{GV}_{g1} - \widehat{GV}_{g0}$. This can be seen by drawing an analogy with two-stage sampling (Cochran, 1977; Sadooghi-Alvandi, 1986), or by direct calculation. Under the assumption that the GVP effects are normally distributed, the efficiency of the contrast for the balanced model with no repeated samples (right column of Table G) relative to the GVP-model with k repeated samples is $1 + \frac{2k\tau_g^2}{m(\tau_g^2+2\sigma_g^2)}$. Removing GVP terms from the paired model effectively lumps these terms in with error. In our calculations, we assume that the distribution of GVP, the effect of the cancer on log-expression level of a particular gene in the population, is approximately normally distributed with variance τ_g^2. Validity of F-tests for the no-GVP-models are based on the same assumption. The advantage of the GVP-model is that by estimating this term explicitly, no assumption about its distribution need be made. But this is not as great an advantage as it may sound because in place of the normality assumption, one is forced to introduce constraints on the GVP terms to make the parameters estimable, and the constraints do not appear more reasonable than the normality assumption. Moreover, the resulting GV estimate in the GVP model is calculated by averaging the individual array contrasts, and so will be sensitive to outliers and skewness in the GVP effects.

Appendix

A Paired Design Calculation With GVP

ANOVA models are highly over-parameterized and require constraints to ensure parameters are well defined. The constraint equations are

$$\begin{aligned} \sum_d GD_{gd} &= 0 \\ GA_{ga} + GA_{g(n+a)} &= 0, a = 1, 2, ..., k \end{aligned}$$

$$\begin{aligned}
\sum_{a=k+1}^{n} GA_{ga} &= 0 \\
\sum_{v} GV_{gv} &= 0 \\
\sum_{p=1}^{k} 2GP_{gp} + \sum_{p=k+1}^{n} GP_{gp} &= 0 \\
\sum_{p=1}^{k} 2GVP_{gvp} + \sum_{p=k+1}^{n} GVP_{gvp} &= 0, v = 1, 2 \\
\sum_{v} GVP_{gvp} &= 0, p = 1, 2, ..., n
\end{aligned}$$

Note that the first constraint allows us to correct for the gene-specific dye effects explicitly for each gene. Spot effects GA will appear in the expected value of the estimator of GV unless the GV estimate is based on within array contrasts between the red and green channels. Therefore, unbiased estimates will have the form

$$\begin{aligned}
\widehat{GV}_{g1} - \widehat{GV}_{g0} &= W_f \sum_{a=1}^{k} (r_{ga11a} - r_{ga00a}) + \\
&\quad W_b \sum_{a=n+1}^{n+k} (r_{ga01(a-n)} - r_{ga10(a-n)}) + \\
&\quad W_u \sum_{a=k+1}^{k+(n-k)/2} (r_{ga11a} - r_{ga00a}) + \\
&\quad W_u \sum_{a=k+1+(n-k)/2}^{n} (r_{ga01a} - r_{ga10a}) \\
&= W_f \sum_{a=1}^{k} (GD_{g1} - GD_{g0} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...) + \\
&\quad W_b \sum_{a=n+1}^{n+k} (GD_{g0} - GD_{g1} + GV_{g1} - GV_{g0} + GVP_{g1(a-n)} - GVP_{g0(a-n)} + ...) + \\
&\quad W_u \sum_{a=k+1}^{k+(n-k)/2} (GD_{g1} - GD_{g0} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...) +
\end{aligned}$$

$$W_u \sum_{a=k+1+(n-k)/2}^{n} (GD_{g0} - GD_{g1} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...)$$

The ... indicate the error terms, which are omitted. The $W's$ must satisfy the constraints

- In order to ensure we are estimating $GV_{g1} - GV_{g0}$, and not some constant multiple of the contrast, we must have $kW_f + kW_b + (n-k)W_u = 1$.
- $W_f = W_b$ ensures that the no dye effects appear in the expected value.
- $W_f = W_b = W_u$ combined with the constraints ensures that the GVP effects will cancel out of the expectation.

The unique solution is $W_u = W_b = W_f = \frac{1}{n+k}$, resulting in the variance

$$var\left(\widehat{GV}_{g1} - \widehat{GV}_{g0}\right) = \frac{2\sigma_g^2}{n+k}.$$

For fixed $m = n + k$, the variance is constant in k.

B Paired Design Calculation Without GVP

Spot effects GA will appear in the expected value of the estimator of GV unless the GV estimate is based on within array contrasts between the red and green channels. Therefore, unbiased estimates will have the form

$$\begin{aligned}
\widehat{GV}_{g1} - \widehat{GV}_{g0} = {} & W_f \sum_{a=1}^{k} (r_{ga11a} - r_{ga00a}) + \\
& W_b \sum_{a=n+1}^{n+k} (r_{ga01(a-n)} - r_{ga10(a-n)}) + \\
& W_u \sum_{a=k+1}^{k+(n-k)/2} (r_{ga11a} - r_{ga00a}) + \\
& W_u \sum_{a=k+1+(n-k)/2}^{n} (r_{ga01a} - r_{ga10a})
\end{aligned}$$

$$
\begin{aligned}
= \quad & W_f \sum_{a=1}^{k}(GD_{g1} - GD_{g0} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...) + \\
& W_b \sum_{a=n+1}^{n+k}(GD_{g0} - GD_{g1} + GV_{g1} - GV_{g0} + GVP_{g1(n-a)} - GVP_{g0(n-a)} + ...) + \\
& W_u \sum_{a=k+1}^{k+(n-k)/2}(GD_{g1} - GD_{g0} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...) + \\
& W_u \sum_{a=k+1+(n-k)/2}^{n}(GD_{g0} - GD_{g1} + GV_{g1} - GV_{g0} + GVP_{g1a} - GVP_{g0a} + ...)
\end{aligned}
$$

The ... indicate the error terms, which are omitted. In order to ensure we are estimating $GV_{g1}-GV_{g0}$, and not some constant multiple of the contrast, we must have $kW_f+kW_b+(n-k)W_u = 1$. In order for an estimate of the variety by gene interaction to be unbiased, we must have $W_f = W_b$ so that the dye effects will cancel out. The GVP are random effects with mean zero, and so will cancel out when we take the expectation. Then we need to minimize the variance function:

$$
Var\left(\widehat{GV}_{g1} - \widehat{GV}_{g0}\right) = k(W_f + W_b)^2\tau_g^2 + 2kW_f^2\sigma_g^2 + 2kW_b^2\sigma_g^2 + (n-k)W_u^2(\tau_g^2 + 2\sigma_g^2)
$$

subject to the constraints. A calculation gives

$$
\begin{aligned}
W_f &= \frac{1}{2}\frac{\tau_g^2 + 2\sigma_g^2}{(n-k)(\tau_g^2 + \sigma_g^2) + k(\tau_g^2 + 2\sigma_g^2)} \\
W_u &= \frac{\tau_g^2 + \sigma_g^2}{(n-k)(\tau_g^2 + \sigma_g^2) + k(\tau_g^2 + 2\sigma_g^2)}
\end{aligned}
$$

Now let $a = \tau_g^2+2\sigma_g^2$, $b = \tau_g^2+\sigma_g^2$, and $c = (n-k)(\tau_g^2+\sigma_g^2)+k(\tau_g^2+2\sigma_g^2)$. Then $W_f^2 = W_b^2 = \frac{a^2}{4c^2}$, $W_u^2 = \frac{b^2}{c^2}$ and $(W_f + W_b)^2 = \frac{a^2}{c^2}$. Then, it can be shown that

$$
\begin{aligned}
var\left(\widehat{GV}_{1g} - \widehat{GV}_{0g}\right) &= \frac{ab}{ka + (n-k)b} \\
&= \frac{\left(\tau_g^2 + 2\sigma_g^2\right)\left(\tau_g^2 + \sigma_g^2\right)}{(m-k)\tau_g^2 + m\sigma_g^2}
\end{aligned}
$$

We want to pick k to maximize this denominator, so pick $k = 0$.

C Two Varieties, Unpaired Samples Calculations

Spot effects GA will appear in the expected value of the estimator of GV unless the GV estimate is based on within array contrasts between the red and green channels. Therefore, unbiased estimates of $GV_{g2} - GV_{g1}$ will be based on intra-array comparisons, and will therefore take the form

$$\begin{aligned}\widehat{GV}_{g2} - \widehat{GV}_{g1} &= W_f \sum_{a=1}^{k} \left(r_{ga11a} - r_{ga02(n+a)} \right) + W_u \sum_{a=k+1}^{n} \left(r_{gad_a1a} - r_{gad_a2(n+a)} \right) + \\ & \quad W_b \sum_{a=n+1}^{k+n} \left(r_{ga01(a-n)} - r_{ga12(n+a-n)} \right) \end{aligned} \tag{1}$$

Each term in the sums in Equation 1 is independent and has variance $\tau_{g1}^2 + \tau_{g2}^2 + 2\sigma_g^2$. This results in the variance function

$$var\left[\widehat{GV}_{g2} - \widehat{GV}_{g1}\right] = \left\{ k\left(W_f + W_b\right)^2 + (n-k)W_u^2 \right\} \left(\tau_{1g}^2 + \tau_{2g}^2 + 2\sigma_g^2 \right)$$

The constraints in this case are

Constraint 1: $kW_f + kW_b + (n-k)W_u = 1$ ensures that we are estimating the variety contrast and not some multiple of the contrast.

Constraint 2: $2W_u = W_f + W_b$ ensures that the sample effects cancel out of the variety estimate under the usual constraint $\sum_{f=1}^{n} n_f GF_{gf} = \sum_{f=n+1}^{2n} n_f GF_{gf} = 0$.

Constraint 3: $W_f = W_b$ ensures that the dye effects cancel out of the variety contrast.

The unique solution to this maximization problem is $W_f = W_b = W_u = \frac{1}{n+k}$, resulting in the variance function

$$\begin{aligned} var\left[\widehat{GV}_{g2} - \widehat{GV}_{g1}\right] &= \left\{ \frac{4k}{(n+k)^2} + \frac{n-k}{(n+k)^2} \right\} \left(\tau_{1g}^2 + \tau_{2g}^2 + 2\sigma_g^2 \right) \\ &= \left\{ \frac{n+3k}{(n+k)^2} \right\} \left(\tau_{1g}^2 + \tau_{2g}^2 + 2\sigma_g^2 \right) \\ &= \frac{m+2k}{m^2} \left(\tau_{1g}^2 + \tau_{2g}^2 + 2\sigma_g^2 \right). \end{aligned}$$

For fixed $m = nn + k$, the variance is minimized when $k = 0$.

D Reference Design Calculations 1

Spot effects GA will appear in the expected value of the estimator of GV unless the GV estimate is based on within array contrasts between the red and green channels. Therefore, unbiased estimates of $GV_{1g} - GV_{0g}$ will be based on intra-array comparisons, and will therefore take the form

$$\begin{aligned}\widehat{GV}_{1g} - \widehat{GV}_{0g} &= W_f \sum_{a=1}^{k} (r_{ga11a} - r_{ga000}) + W_u \sum_{a=k+1}^{n} (r_{ga11a} - r_{ga000}) + \\ & \quad W_b \sum_{a=n+1}^{k+n} \left(r_{ga01(a-n)} - r_{ga100}\right) \end{aligned} \tag{2}$$

The constants W_f, W_u, and W_b are to be selected to minimize the variance of the contrast subject to three constraints.

Constraint 1: $2W_u = W_b + W_f$ assures that no sample effects appear in the expected value of the variety contrast estimate; so that the estimate is unbiased for all values of the sample effects. This is true because of the usual constraint on the sample effects $0 = \sum_{f \in V_1} n_f GF_{gf}$ where n_f is the number of times sample f appears on an array.

Constraint 2: $(n-k)W_u = k\,(W_b - W_f)$ assures that no dye effects appear in the expected value of the variety contrast estimate.

Constraint 3: $kW_f + (n-k)W_u + kW_b = 1$ assures that we are estimating the contrast, and not some multiple of the contrast.

This gives three constraint equations for three unknowns. The unique solution is:

$$W_f = \frac{3k-n}{2k(k+n)}; W_b = \frac{1}{2k}; W_u = \frac{1}{k+n}.$$

Note that $var\left[r_{ga11a} - r_{ga000}\right] = \tau_g^2 + 2\sigma_g^2$, and that $var\left[W_f\left(r_{ga11a} - r_{ga000}\right) + W_b\left(r_{ga01(a-n)} - r_{ga100}\right)\right] = \left(W_f + W_b\right)^2 4\tau_g^2 + \left(W_f^2 + W_b^2\right) 4\sigma_g^2$. Therefore, the contrast variance is:

$$\begin{aligned} var\left[\widehat{GV}_{g1} - \widehat{GV}_{g0}\right] &= \left[k\left(W_f + W_b\right)^2 + (n-k)W_u^2\right]\tau_g^2 + \\ & \quad \left[2kW_f^2 + 2(n-k)W_u^2 + 2kW_b^2\right]\sigma_g^2 \\ &= \frac{n+3k}{(n+k)^2}\tau_g^2 + \frac{n^2+3k^2}{k(n+k)^2}\sigma_g^2 \end{aligned} \tag{3}$$

If we fix the number of arrays $m = n + k$, then the k which minimizes the variance is

$$k = max\left[1, \frac{m\sigma_g}{\sqrt{2\tau_g^2 + 4\sigma_g^2}}\right].$$

E Reference Design Calculations 2

In this case, the "treatments" are balanced with respect to the dyes, so that each appears half the time with each dye tag. As a result, the dye effects will automatically cancel out of the variety contrast estimate in Equation 2 as long as $W_f = W_b$, but constraint 2 of the last Appendix no longer need be satisfied. The other two constraints still need to be satisfied. So we add this new constraint, $W_f = W_b$, to the constraint equations and solve.

We need to minimize Equation 3 subject to constraints 1 and 3 of the previous Appendix, and the new constraint $W_f = W_b$. The unique solution $W_f = W_b = W_u = \frac{1}{n+k}$.

Plugging these into the contrast variance Equation 3 gives

$$var\left[\widehat{GV}_{g1} - \widehat{GV}_{g0}\right] \quad = \quad \frac{\tau_g^2 + 2\sigma_g^2}{m} + \frac{2k\tau^2}{m^2}.$$

The variance will be minimized when $k = 0$, i.e., when each array contains an unique sample.

F Proof of the Loss of Efficiency When Samples Are Replicated in a Paired Design

Assume there are n individuals randomly selected from some population of interest, and m pairs of samples from each individual. Also assume that both members from each pair are to be placed together on each array, and we wish to estimate the average difference between the pairs for this population of individuals. We wish to prove that if there is no dye bias, then replicating individuals over the arrays decreases the efficiency of the population estimates.

Our model is

$$y_{ij} \quad = \quad \mu + s_i + \epsilon_{ij}$$

where μ is the population mean we wish to estimate, s_i is the effect of individual i, and ϵ_{ij} is the experimental error. Here, $i = 1, ..., n$ indexes the individuals and $j = 1, ..., m$ indexes the different

replicates for each individual. Assume s_i is normal with mean 0 and variance τ^2, and ϵ_{ij} is normal with mean zero and variance σ^2, and that all of these are independent. Then τ^2 represents biological variation and σ^2 represents experimental variation. The total number of observations is nm. The variance of the estimated population average log-ratio is

$$\begin{aligned} var\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right) &= var\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i\cdot}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} var\left(\bar{y}_{i\cdot}\right) \\ &= \frac{1}{n}\left(\tau^2 + \frac{\sigma^2}{m}\right) \\ &= \frac{\tau^2}{n} + \frac{\sigma^2}{nm} \end{aligned}$$

The number of arrays is $a = nm$. For fixed a, the variance is $\frac{m\tau^2}{a} + \frac{\sigma^2}{a}$. The variance is an increasing function of m, so pick $m = 1$ to minimize the variance.

Therefore, we have shown that the variance of the estimated population mean is minimized when no samples are replicated, and increases with the number of replicates. Therefore, the most efficient design is one that uses a single pair of samples from each individual.

G Tables

Source of Variation	*Degrees of Freedom* GP and GVP included	*Degrees of Freedom* GP only included	*Degrees of Freedom* GP and GVP omitted
G	1	1	1
$GA(GP)$	n	n	n+k-1
GD	1	1	1
GV	1	1	1
GP	n-1	n-1	0
GVP	n-1	0	0
Error	k-1 [a]	n+k-2 [b]	n+k-2
Total	2 (n+k)	2 (n+k)	2 (n+k)

[a] GP and GA effects for arrays $k+1$ to n are completely confounded; hence the predictors are linearly dependent, and the dimension of the predictor space is $3n + 1 - (n - k) = 2n + k + 1$ leaving $k - 1$ df for error.

[b] In this case, the dimension of the predictor space is $2n + 2 - (n - k) = n + k + 2$ leaving $n + k - 2$ degrees of freedom for error.

Table A: Analysis of Variance Tables for Paired Design. $GA(GP)$ notation indicates that the GA gene by array (spot) effects are nested within the GP gene by participant effects.

Source of Variation	*Degrees of Freedom* (with GF)	*Degrees of Freedom* (without GF)
G	1	1
GA	n+k-1	n+k-1
GD	1	1
GV	1	1
GF	n-1	0
Error	k-1	n+k-2
Total	2 (n+k)	2 (n+k)

Table B: Analysis of Variance Table for Reference Design