

Statistical Appendix

We propose a simple model for exploring the effects of amplification on gene expression:

$$X_{ij} = Z_i + e_{ij}, \quad (1)$$

where X_{ij} is the measured expression value for gene i in replication j of a specific microarray experiment, Z_i is the “true” expression value for that gene, and e_{ij} is the measurement and experimental error that will change across replications. Note that Z_i is a theoretical quantity, since we can never measure Z_i directly. If we assume that the e_{ij} average 0 across replications, then we can think of Z_i as the average expression value for gene i if we replicated the experiment infinitely many times. Likewise $\sigma_i^2 = \text{Var}(e_{ij})$ is the variance of e_{ij} for infinitely many replications.

Treating the measurements on different genes as “exchangeable”, we can talk of the variance of Z across genes:

$$\text{Var}_g(Z) = \frac{1}{p} \sum_{i=1}^p (Z_i - E_g(Z))^2, \quad (2)$$

where p is the total number of genes under study (here around 11,000), and $E_g(Z)$ is the mean of Z :

$$E_g Z = \frac{1}{p} \sum_{i=1}^p Z_i \quad (3)$$

Likewise,

$$\text{Var}_g(e) = \frac{1}{p} \sum_{i=1}^p \sigma_i^2, \quad (4)$$

the average measurement-error variance across genes.

Since the Z_i and e_{ij} are not directly observable, we cannot measure $\text{Var}_g(Z)$ and $\text{Var}_g(e)$ directly; however, we can estimate them both after making some mild assumptions. We assume that $E_g(e_j) = 0$ for each experiment, and that the e_{ij} are independent and identically distributed across replications. The first assumption is technical, and can be avoided with some extra detail. It can be achieved exactly by removing the mean expression value for each microarray, and assuming that $E_g(Z) = 0$. The second assumption is standard. Here a single subscript, such as X_j , refers to the entire collection $\{X_{ij}\}$ of observed values for the j th experiment.

It can then be easily seen that

$$E[\text{Var}_g X_j] = \text{Var}_g(Z) + \text{Var}_g(e), \quad (5)$$

where the unsubscripted E refers to replication expectation.

Likewise, let

$$\text{Cov}_g(X_j, X_k) = \frac{1}{p} \sum_{i=1}^p (X_{ij} - E_g(X_j))(X_{ik} - E_g(X_k)) \quad (6)$$

be the covariance across genes between replicate j and replicate k . Then

$$\mathbb{E}[\text{Cov}_g(X_j, X_k)] = \text{Var}_g(Z). \quad (7)$$

Hence by averaging the variances and covariances between replicates for a particular amplification method, we can get estimates of $\text{Var}_g(Z)$ and $\text{Var}_g(e)$ for that particular method. This is the approach used to produce table 5.

A similar approach was used to produce the decomposition in table 6, although there we used the covariances between virtual arrays.