Supplement 1. Effects of cross-hybridization and their relevance to correlation analysis of microarray data

There is no theoretical reason to expect that the multiple targeting process induces only positive correlations. Consider the simplest case of two probe sets targeting the same transcript (RNA molecule). Suppose that the first probe set binds to this transcript with probability p, and, given the transcript remains unbound, the second probe set binds to it with probability $1-p$. Assuming in addition that no saturation effects are present and denoting the total amount of transcripts of a given type by U, the first probe-set signal is

$$Z_1 = \sum_{j=1}^{U} \xi_j,$$

where ξ_j are i.i.d. indicator variables taking on values 1 and 0 with the probabilities p and $1-p$, respectively. For the second probe set we have

$$Z_2 = \sum_{j=1}^{U} (1-\xi_j).$$

It can be easily verified that the covariance between Z_1 and Z_2 can be of any sign under this natural model. Indeed,

$$\text{Cov}\,(Z_1, Z_2) = \mathbb{E}\{\sum_{j=1}^{U} \xi_j \sum_{k=1}^{U} (1-\xi_k)\} - \mathbb{E}\{\sum_{j=1}^{U} \xi_j\}\,\mathbb{E}\{\sum_{k=1}^{U} (1-\xi_k)\} =$$

$$p(1-p)\mathbb{E}\{U(U-1)\} - p(1-p)\big(\mathbb{E}\{U\}\big)^2 = p(1-p)\big(\text{Var}\{U\} - \mathbb{E}\{U\}\big),$$

so that the sign of Cov (Z_1, Z_2) depends on the relationship between $\text{Var}\{U\}$ and $\mathbb{E}\{U\}$, both parameters being unobservable, of course. In particular, Cov $(Z_1, Z_2) = 0$ if U has a Poisson distribution.

In a recent paper, Okoniewski & Miller (2006) reported evidence in favor of the presence of spurious positive correlations induced by the process of multiple targeting (MT). Their work, however, is open to criticism and we plan to discuss it in detail in another paper. Briefly, there are several problems with their approach, the most serious of which is that the Novaritis Gene Atlas data set used in the paper is not amenable to correlation analysis. The reason is that the chosen data set does not represent a random sample drawn from a relatively homogeneous population but a mix of arrays derived from diverse biological specimens, each being of a different origin and each representing a single copy of the corresponding set of expression measurements. Our independent analysis of two other data sets is in conflict with the observations reported

by Okoniewski & Miller (2006). After removing multiple probe sets that target the same transcript, we observed either no change or even the tendency to a decrease in the average (across probe sets) value of correlation coefficients.

A simulation study presented by Okoniewski & Miller (2006) as proof of principle is also questionable. Their essentially deterministic model amounts to assuming that $Z_1 = pU$, $Z_2 = (1-p)U$, which is why its simulation counterpart displays a positive covariance. It is implicit in the latter model that each probe set "knows" exactly what proportion of the random amount U it must "catch" in the course of hybridization, a very implausible presumption.

The phenomenon of type A dependence is allowable under the above-considered stochastic model of cross-hybridization. However, its abundance cannot be explained by this mechanism. For HG_U133A arrays, Okoniewski & Miller (2006) identified 3859 non-overlapping families containing MT probe sets with an average number of 2.56 probe sets per family. Even if all gene pairs formed within each family display the type A dependence, the total number of such pairs will be of order 10^4. In contrast, the total number of type A pairs in a set of 22000 genes is expected to be over 10^8 on the average, quite a dramatic difference. In summary, we see no compelling reason to attribute the phenomenon of type A dependence to the effects of cross-hybridization.

References

Okoniewski, M.J. & Miller, C.J. (2006), 'Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations,' *BMC Bioinformatics* **7**, Article 276.