Supporting Appendix: A re-analysis of the Choe et al. Affymetrix GeneChip control dataset

Alan R Dabney, John D Storey

Address: Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Email: adabney@u.washington.edu, jstorey@u.washington.edu;

*Corresponding author

Simulation details. Let \( X_{i0} = (X_{i0A}, X_{i0B})^T \) be the RNA amounts for gene \( i \), \( i = 1, 2, \ldots, m \), on a randomly-selected individual under conditions A and B. Suppose that \( X_{i0} \) is distributed according to \( N_2(\mu_i, \Sigma_i) \), the bivariate normal distribution with mean \( \mu_i = (\mu_iA, \mu_iB)^T \) and covariance matrix

\[
\Sigma_i = \sigma_i^2 \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}.
\] (1)

Equivalently, we can write

\[
X_{i0} = \mu_i + \epsilon_i,
\] (2)

where \( \epsilon_i \sim N_2(0, \Sigma_i) \), \( i = 1, 2, \ldots, m \). Let \( X_i \) be the observed expression from a microarray. We write

\[
X_i = \mu_i + \epsilon_i + \phi_i,
\] (3)

where \( \phi_i = (\phi_{iA}, \phi_{iB})^T \sim N_2(0, \tau_i^2 I_2) \) represents hybridization variability, \( i = 1, 2, \ldots, m \); by \( I_2 \), we mean the \( 2 \times 2 \) identity matrix.

Suppose we form three technical replicates on a single individual. Then, the three observations can be written as

\[
X_{ijT} = \mu_i + \epsilon_i + \phi_{ij},
\] (4)

\( i = 1, 2, \ldots, m \), \( j = 1, 2, 3 \). If, on the other hand, we sample three independent individuals, we obtain

\[
X_{ijI} = \mu_i + \epsilon_{ij} + \phi_{ij},
\] (5)

\( i = 1, 2, \ldots, m \), \( j = 1, 2, 3 \). Note that, in terms of Figure 1 in Choe et al, \( \epsilon \) represent variability introduced in columns 1-4, while \( \phi \) represents the variability introduced between column 4 and 5.

The simulations summarized in Figures 2-4 were generated as follows. For \( m = 15000 \) genes, baseline means \( \mu_{i0} \) were generated from the \( N(0,1) \) distribution. Then, for 6000 alternative genes, differential expression was created by sampling \( \delta_i \) from the \( N(0,1) \) distribution. The means for each alternative gene were formed as \( \mu_{iA} = \mu_{i0} + \delta_i \) and \( \mu_{iB} = \mu_{i0} \). The means for the remaining 9000 null genes were formed as \( \mu_{iA} = \mu_{iB} = \mu_{i0} \). The standard deviations \( \sigma_i \) and \( \tau_i \) were taken to be 0.3 and 0.2, respectively. The correlation parameter \( \rho \) was chosen as 0.85. In simulation one, three technical replicates were formed. In simulation two, three biological replicates were formed. The above scenario was simulated 30 times.