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 $\boldsymbol{X}_{i0} = (X_{i0A}, X_{i0B})^T$ be the RNA amounts for gene i, $i = 1, 2, \ldots, m$, on a randomly-selected individual under conditions A and B. Suppose that $\boldsymbol{X}_{i0}$ is distributed according to $\boldsymbol{N}_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, the bivariate normal distribution with mean $\boldsymbol{\mu}_i = (\mu_{iA}, \mu_{iB})^T$ and covariance matrix

$$\boldsymbol{\Sigma}_i = \sigma_i^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{1}$$

Equivalently, we can write

$$\boldsymbol{X}_{i0} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\boldsymbol{\epsilon}_i \sim \boldsymbol{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, $i = 1, 2, \ldots, m$. Let $\boldsymbol{X}_i$ be the observed expression from a microarray. We write

$$\boldsymbol{X}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i + \boldsymbol{\phi}_i, \tag{3}$$

where $\boldsymbol{\phi}_i = (\phi_{iA}, \phi_{iB})^T \sim \boldsymbol{N}_2(\boldsymbol{0}, \tau_i^2 \boldsymbol{I}_2)$ represents hybridization variability, $i = 1, 2, \ldots, m$; by $\boldsymbol{I}_2$, we mean the 2×2 identity matrix.

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

$$\boldsymbol{X}_{ijT} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i + \boldsymbol{\phi}_{ij}, \tag{4}$$

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

$$\boldsymbol{X}_{ijI} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij} + \boldsymbol{\phi}_{ij}, \tag{5}$$

$i = 1, 2, \ldots, m$, $j = 1, 2, 3$. Note that, in terms of Figure 1 in Choe it et al, ϵ represent variability introduced in columns 1-4, while ϕ 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 μ_{i0} were generated from the $N(0, 1)$ distribution. Then, for 6000 alternative genes, differential expression was created by sampling δ_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 σ_i and τ_i were taken to be 0.3 and 0.2, respectively. The correlation parameter ρ 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.