Supporting Information

McMullen et al. 10.1073/pnas.1000938107

SI Text

Quality Control. The measured intensity B_i is, in principle, uncorrelated with the intensity of (*i*) feature regions, (*ii*) neighboring nonfeature regions, and (*iii*) nonfeature regions in the corresponding spot for the alternate dye system. In the absence of defects, intensities of the nonfeature regions represent an ensemble of independent realizations of the same process. Uncharacteristic values of nonfeature intensity necessarily indicate idiosyncratic characteristics of the spot that may mask the intended measurement. These include, but are not limited to, scratches and particulates (1).

Specifically, the fluorescence at the nonfeature region of a spot is expected to be near a characteristic value for the chip. We expect the largest contribution to the fluctuations of B_i to be a result of noise from the detector system, as discussed above. Let B_i^0 be the fluorescence at each nonfeature region from optical background. This quantity is identical for spots on the chip without problems; that is, $B_i^0 = B^0$. Photomultiplication of this signal is characterized by a dye-specific gain G^{nf} (which may be distinct from G) and a characteristic error. Following the assumptions of Eq. **9**, the expression for the observed fluorescence at nonfeature region is $B_i = B^0 G^{\mathrm{nf}} e^{\varepsilon^{\mathrm{nf}}}$, where $\varepsilon^{\mathrm{nf}}$ is a normally distributed variate with zero mean. Thus, B_i will be a log-normal variate.

We suspected, based on visual inspection of microarray images, that surface flaws are pervasive in publicly available data, and use this model of nonfeature intensity to screen for flawed spots. Because the source of these defects is often not conclusively known, and because their effect on spot feature intensities is unknown, it is prudent to exclude these spots from analysis. To this end, we identify spots that do not follow the above model and are, therefore, subject to exogenous factors such as scratches and particulates.

One expects a physical defect on a chip to affect the nonfeature intensities in both channels of a two-color microarray, and the chance that both channels have extreme nonfeature intensities is remote—assuming the spot has no defect. We take advantage of this to build a statistic for assessing the *irregularity* of a spot, $I_i = z_i^1 z_i^2$, where z_i^1 and z_i^2 are the z scores of $\log B_i^1$ and $\log B_i^2$, the nonfeature intensities recorded in the two channels, respectively (Fig. 1). Extreme values of I_i are rare in the absence of surface flaws. Spots for which $|I_i| > 3$ have less than a 2% probability of occurring by chance and are likely to be defective. More sophisticated filters could incorporate spatial correlation of extreme values into a metric of spot quality, resulting in a more specific metric.

PCR-Based Amplification. An alternative amplification procedure to the T7 expression system is PCR, a branching process in which stages are separated discretely. The fundamental difference between sample amplification by T7 expression and PCR is that the latter is a branching process in which products become reactants in successive rounds, while T7 transcripts do not. Each step in PCR amplification process can be modeled as a multiplicative growth process, yielding

$$n_i^{l+1} = n_i^l (2 - \delta + \epsilon_{il}^p). \qquad \text{[S1]}$$

For simplicity, we assume that ϵ_{il}^p is a Gaussian variate with mean zero and standard deviation σ^p, whereas $\delta \ll 1$. The parameters δ and σ^p are related to the probability of a duplication event for a species in a single round of the PCR branching process. If the PCR amplification involves k steps, then the number of copies in the sample will be

$$n_i^{\mathrm{PCR}} = E_i \prod_{l=1}^{k} (2 - \delta + \epsilon_{il}^p), \qquad \text{[S2]}$$

$$= E_i \prod_{l=1}^{k} (2-\delta)\left(1 + \frac{\epsilon_{il}^p}{2-\delta}\right), \qquad \text{[S3]}$$

$$= E_i (2-\delta)^k \exp\left(\sum_{j=1}^{k} \frac{\epsilon_{il}^p}{(2-\delta)}\right), \qquad \text{[S4]}$$

$$= E_i A_p e^{\nu_i^p}, \qquad \text{[S5]}$$

where, by the conditions of the central limit theorem, ν_i^p is a Gaussian variate with mean zero.

Distribution of Fluorescence Intensities. The fluctuations arising in a microarray experiment are nontrivial, but as we have shown, can be modeled explicitly. A successful model of microarray data should, among other things, describe the distribution of F_i measured in a single experiment. Assessment of quality of fit (Fig. 2) between the two discussed models and the observed distribution of F_i requires an expression for the probability density predicted by both models. Here, we derive an expression for the distribution of observed fluorescence intensities from a single microarray experiment for both the physically grounded model and the standard statistical model. Generally, we derive $p(F_i|\mathscr{R}(\boldsymbol{\theta}^R),\mathscr{M}^E(\boldsymbol{\theta}^E))$, an expression for the probability density of a value F_i, which depends on a model for the response function $\mathscr{R}(\boldsymbol{\theta}^R)$ and on a model for the expression levels in the sample $\mathscr{M}^E(\boldsymbol{\theta}^E)$. Here, we define $\mathscr{M}^E(\boldsymbol{\theta}^E)$ by Eq. **10**.

Physically Grounded Model. The probability density functions for the Gaussian stochastic variables ν_i^{sp} and ν_i^{nsp} can be written as

$$\phi_{\mathrm{sp}}(\nu_i^{\mathrm{sp}}) = \phi(\nu_i^{\mathrm{sp}}|\sigma^{\mathrm{sp}}) = \frac{1}{\sigma^{\mathrm{sp}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\nu_i^{\mathrm{sp}}}{\sigma^{\mathrm{sp}}}\right)^2},$$
$$\phi_{\mathrm{nsp}}(\nu_i^{\mathrm{nsp}}) = \phi(\nu_i^{\mathrm{nsp}}|\sigma^{\mathrm{nsp}}) = \frac{1}{\sigma^{\mathrm{nsp}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\nu_i^{\mathrm{nsp}}}{\sigma^{\mathrm{nsp}}}\right)^2}, \qquad \text{[S6]}$$

where $\phi(x|\sigma)$ denotes the probability density function of a Gaussian variate with zero mean and standard deviation σ.

Let us now consider the distribution of expression levels in a sample. Generally, a model $\mathscr{M}^E(\boldsymbol{\theta}^E)$ for expression levels specifies a probability density $p(E_i|\mathscr{M}^E(\boldsymbol{\theta}^E))$ of a value E_i that will depend on the set of parameters $\boldsymbol{\theta}^E$. Following Eq. **9**, the probability of a spot having fluorescence level F_i is

$$p(F_i|\mathscr{R}^{\mathrm{phys}}(\boldsymbol{\theta}^R),\mathscr{M}^E(\boldsymbol{\theta}^E)) = \int_0^\infty dE_i \int_{-\infty}^\infty d\nu_i^{\mathrm{sp}} \int_{-\infty}^\infty d\nu_i^{\mathrm{nsp}} \times p(F_i,E_i,\nu_i^{\mathrm{sp}},\nu_i^{\mathrm{nsp}}|\mathscr{R}^{\mathrm{phys}}(\boldsymbol{\theta}^R),\mathscr{M}^E(\boldsymbol{\theta}^E)), \qquad \text{[S7]}$$

where $p(F_i|\mathscr{R}^{\mathrm{phys}}(\boldsymbol{\theta}^R),\mathscr{M}^E(\boldsymbol{\theta}^E))$ denotes the conditional probability of observing F_i, given our physically grounded model $\mathscr{R}^{\mathrm{phys}}(\boldsymbol{\theta}^R)$ with parameters $\boldsymbol{\theta}^R = (\sigma^{\mathrm{sp}},\sigma^{\mathrm{nsp}},A,U)$. Because in our model E_i, ν_i^{sp}, and ν_i^{nsp} are independent, we can marginalize over their distributions, obtaining

$$p(F_i|\mathcal{R}^{\text{phys}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_{-\infty}^{\infty} d\nu_i^{\text{nsp}} \int_0^{\infty} dE_i \phi_{\text{sp}}(\nu_i^{\text{sp}})\phi_{\text{nsp}}(\nu_i^{\text{nsp}}) \times p(E_i|\mathcal{M}^E(\boldsymbol{\theta}^E))p(F_i|E_i,\nu_i^{\text{sp}},\nu_i^{\text{nsp}},\mathcal{R}(\boldsymbol{\theta}^R)). \quad \textbf{[S8]}$$

Following Eq. **9**, we can write,

$$p(F_i|\mathcal{R}^{\text{phys}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_{-\infty}^{\infty} d\nu_i^{\text{nsp}} \int_0^{\infty} dE_i \times \phi_{\text{sp}}(\nu_i^{\text{sp}})\phi_{\text{nsp}}(\nu_i^{\text{nsp}})p(E_i|\mathcal{M}^E(\boldsymbol{\theta}^E)) \times \delta[E_i A e^{\nu_i^{\text{sp}}} - (F_i - U e^{\nu_i^{\text{nsp}}})], \quad \textbf{[S9]}$$

where $\delta[\bullet]$ is the Dirac delta function. Here we make the substitution $\gamma = E_i A e^{\nu_i^{\text{sp}}}$, which allows us to write the delta function as

$$p(F_i|\mathcal{R}^{\text{phys}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_{-\infty}^{\infty} d\nu_i^{\text{nsp}} \int_0^{\infty} \frac{d\gamma}{A e^{\nu_i^{\text{sp}}}} \phi_{\text{sp}}(\nu_i^{\text{sp}})\phi_{\text{nsp}}(\nu_i^{\text{nsp}}) \times p\left(\frac{\gamma}{A e^{\nu_i^{\text{sp}}}}|\mathcal{M}^E(\boldsymbol{\theta}^E)\right)\delta[\gamma - (F - U e^{\nu_i^{\text{nsp}}})]. \quad \textbf{[S10]}$$

Because the delta function in Eq. **S10** is single-valued, the integral over this product is only nonzero when $\gamma = E_i A e^{\nu_i^{\text{sp}}} = F_i - U e^{\nu_i^{\text{nsp}}}$, allowing the substitution,

$$p(F_i|\mathcal{R}^{\text{phys}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_{-\infty}^{\infty} d\nu_i^{\text{nsp}} \frac{1}{A e^{\nu_i^{\text{sp}}}} \phi_{\text{sp}}(\nu_i^{\text{sp}})\phi_{\text{nsp}}(\nu_i^{\text{nsp}}) \times p\left(\frac{F_i - U e^{\nu_i^{\text{nsp}}}}{A e^{\nu_i^{\text{sp}}}}|\mathcal{M}^E(\boldsymbol{\theta}^E)\right). \quad \textbf{[S11]}$$

Following Eq. **10**, we complete the derivation by incorporating $\mathcal{M}^E(\boldsymbol{\theta}^E)$,

$$p(F_i|\mathcal{R}^{\text{phys}}(\boldsymbol{\theta}^R),\mathcal{M}^{PL}(\alpha)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_{-\infty}^{\infty} d\nu_i^{\text{nsp}} \times \frac{1}{A e^{\nu_i^{\text{sp}}} 2\pi\sigma^{\text{sp}}\sigma^{\text{nsp}}} e^{-\frac{1}{2}\left[\left(\frac{\nu_i^{\text{sp}}}{\sigma^{\text{sp}}}\right)^2 + \left(\frac{\nu_i^{\text{nsp}}}{\sigma^{\text{nsp}}}\right)^2\right]} \times \frac{(\alpha-1)}{\left(\frac{F_i - U e^{\nu_i^{\text{nsp}}}}{A e^{\nu_i^{\text{sp}}}} + 1\right)^{\alpha}}. \quad \textbf{[S12]}$$

Standard statistical model. The standard statistical model (Eq. **2**) involves two measured quantities at each spot, F_i and B_i. Here we derive an expression for the difference between these variates,

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_0^{\infty} dE_i \times p(F_i', E_i, \nu_i^{\text{sp}}|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)), \quad \textbf{[S13]}$$

where $F_i' = F_i - B_i$. Marginalizing over E_i and ν_i^{sp}, as above, we obtain,

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_0^{\infty} dE_i \times \phi_{\text{sp}}(\nu_i^{\text{sp}})p(F_i'|E_i,\nu_i^{\text{sp}},\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)). \quad \textbf{[S14]}$$

Incorporating the expression from Eq. **2**, we can write,

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_0^{\infty} dE_i \phi_{\text{sp}}(\nu_i^{\text{sp}}) \times p(E_i|\mathcal{M}^E(\boldsymbol{\theta}^E))\delta[F_i' - E_i A e^{\nu_i^{\text{sp}}}]. \quad \textbf{[S15]}$$

Again, we make the substitution $\gamma = E_i A e^{\nu_i^{\text{sp}}}$, allowing us to rewrite the integrals,

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \int_0^{\infty} d\gamma \frac{1}{A e^{\nu_i^{\text{sp}}}} \phi_{\text{sp}}(\nu_i^{\text{sp}}) \times p\left(\frac{\gamma}{A e^{\nu_i^{\text{sp}}}}|\mathcal{M}^E(\boldsymbol{\theta}^E)\right)\delta[F_i' - \gamma]. \quad \textbf{[S16]}$$

Integrating over this delta function, as above, yields

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^E(\boldsymbol{\theta}^E)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \frac{1}{A e^{\nu_i^{\text{sp}}}} \phi_{\text{sp}}(\nu_i^{\text{sp}}) p\left(\frac{F_i'}{A e^{\nu_i^{\text{sp}}}}|\mathcal{M}^E(\boldsymbol{\theta}^E)\right), \quad \textbf{[S17]}$$

and adding the expression from Eq. **10**, we can write,

$$p(F_i'|\mathcal{R}^{\text{stat}}(\boldsymbol{\theta}^R),\mathcal{M}^{PL}(\alpha)) = \int_{-\infty}^{\infty} d\nu_i^{\text{sp}} \frac{1}{A e^{\nu_i^{\text{sp}}} \sqrt{2\pi}\sigma^{\text{sp}}} e^{-\frac{1}{2}\left(\frac{\nu_i^{\text{sp}}}{\sigma^{\text{sp}}}\right)^2} \frac{(\alpha-1)}{\left(\frac{F_i'}{A e^{\nu_i^{\text{sp}}}} + 1\right)^{\alpha}}. \quad \textbf{[S18]}$$

Heuristics for estimating model parameters. Our physically grounded model is simple enough that one can obtain maximum-likelihood estimates of A, U, σ_{sp}, σ_{nsp}, and α by the method of steepest descent. However, because of the large number of data in microarray experiments, this is computationally expensive. Thus, in general, we turn to heuristics to more quickly estimate model parameters.

If one knew which genes were unexpressed, contributions from nonspecific binding could be estimated directly. Unfortunately, we have little knowledge about the set of unexpressed genes a priori, but we can infer them from $p(F)$. Because they are conditionally expressed, tissue specific, or cell-cycle dependent, a significant fraction of the genes in a cell are not being expressed at any particular time—that is, $E_i = 0$. We exploit this fact to estimate U. For these genes, Eq. **9** reduces to

$$F_i^{E_i=0} = U e^{\nu_i^{\text{nsp}}}. \quad \textbf{[S19]}$$

The distribution of observed feature intensities must be the sum of discrete valued mRNA counts, $\{E_i\}$, subject to the noises discussed previously (Eq. **9**). We show a cartoon of this in Fig. S1.

Spots with very low fluorescence have a high likelihood of belonging to the set $F_i^{E_i=0}$. Note that, from Eq. **S19**, the values $F_i^{E_i=0}$ are log-normally distributed. We have developed a heuristic for identifying the largest set of F_i that is consistent with this log-normal distribution as a means of estimating U and σ_{nsp}.

We define a critical feature intensity F_c such that there is a high probability that features below this threshold belong to the set of unexpressed genes, $\{F_c|F_i < F_c \in F_i^{E=0}\}$. From this set, we can estimate U and σ_{nsp} by fitting it to a right-truncated Gaussian distribution,

$$p(\log F_i|F_i < F_c) = \frac{\phi\left(\frac{\log F_i - \log\hat{U}}{\hat{\sigma}_{\text{nsp}}}\right)}{\Phi\left(\frac{\log F_c - \log\hat{U}}{\hat{\sigma}_{\text{nsp}}}\right)}, \quad \textbf{[S20]}$$

where ϕ and Φ are the probability density function and cumulative distribution functions for a Gaussian distribution, respectively.

Extraneous Quality Issues. Our model is inherently flexible to many chip-based technologies. Nevertheless, its scope here is limited to describing data for which consequences of amplification, hybridization, and detection determine the feature fluorescence in the manner we have described. A number of other effects, due to technical limitations or flawed equipment or protocol, have been

previously shown to affect results. Care must be taken to avoid or (if possible) correct these sources of bias, lest they influence the predictions of our physically grounded model. Here we discuss how some of these common biases can be identified.

Sequence-specific dye bias. Factors that affect data in a spot-dependent manner are particularly dangerous to microarray data (2–4). Because changes in expression level are sequence/gene- and dye/condition-dependent effects, these are difficult to distinguish from sequence-specific dye bias, in which a differential fluorescence is observed at a spot in the absence of expression change. Furthermore, in many cases, the extent of this bias has been observed to vary between experiments performed using the same system, making generalization difficult (4). It has recently been suggested that sequence-specific dye bias results from differences in the extent of label incorporation (4), indicating that many of these effects can be ameliorated through optimization of labeling protocols.

Detector saturation. Saturation during detection is a natural challenge when measuring fluorescences that span many orders of magnitude and has been shown to be widespread in microarray data (5–9). In many cases, affected spots are easily identified by truncated intensity values in one or both channels. However, saturation near detector limits may result in nonlinear gamma saturation (5) that is more difficult to identify. A hallmark of this saturation is a dependence of $\hat{R}$ on $\hat{E}$ for spots with large $\hat{E}$. Avoiding detector saturation requires careful calibration of the detector during data acquisition. Employing multiple scans of the chip, at varying sensitivities, allows saturation to be addressed during analysis (5).

Sensitivity of nonspecific binding parameters. Small variations in the estimates of U, the characteristic nonspecific noise contribution, may result in changes in the predictions of up- and down-regulated genes, particularly for weakly expressed genes. If the model disagrees with the data—that is, if there are processes at work other than amplification, hybridization, and detection as we have described in the text—the extent of nonspecific binding will be difficult to estimate. We have found the distribution of feature fluorescence intensities predicted by the model to provide a prudent assessment of the fit of the model to the data.

Spot quality. Scratches, particulates, or manufacturing flaws may influence spot feature intensities in an idiosyncratic way. We have described a procedure to identify spots influenced by these and other exogenous factors. Parameter estimates for our model are, in principle, sensitive to these *irregular* spots, and they should be removed.

Similarly, we have observed examples of severe spatial correlation in nonfeature intensities on some chips. This can occur such that the distribution of nonfeature intensities is well behaved—that is, that nonfeature intensities are log-normally distributed. The extent to which this "drift" affects feature intensities is difficult to determine, and special care must be taken when fitting model parameters for these chips. We recommend fitting parameters for different regions of the chip separately.

Microarray Quality Control Project Data. We next investigate the implications of our model in the context of published data from the Microarray Quality Control Project (10). This study used commercially available Stratagene Universal Human Reference and Ambion Human Brain Reference RNA samples to assess technical concerns regarding the reproducibility of microarray results. These two samples were used by three different labs in four different array experiments (Fig. 4*A* and Fig. S4), each performed five times. We denote the vector of expression levels over all genes on a chip for these samples $\mathbf{R}^{elr}$, where $e \in \{A,B,C,D\}$ is the experiment type, $l \in \{1,2,3\}$ is the lab index, and $r \in \{1,\ldots,5\}$ is the replicate index.

In two of these experiments (A and B), there is no difference in the underlying expression levels for the two samples on the chip; that is $R_i^{A\cdot\cdot} = R_i^{B\cdot\cdot} = 0$ for all spots on the chip. Deviations of the best estimate $\hat{\mathbf{R}}$ predicted by a model from the expected value $\hat{\mathbf{R}} = (0,\ldots,0)$ are the result of our inability to precisely estimate stochastic fluctuations in the experiment. For the other two experiments (C and D), the two samples hybridized to the chip are different. We thus expect that $\mathbf{R}^{C\cdot\cdot} = \mathbf{R}^{D\cdot\cdot}$. Furthermore, we expect many components of $\mathbf{R}$ to be significantly different from zero.

For experiments A and B, there is no difference in expression between the two samples ($R_i^{A\cdot\cdot} = R_i^{B\cdot\cdot} = 1$); therefore $\hat{\mathbf{R}}^A$ and $\hat{\mathbf{R}}^B$ are the exact residuals of an estimation procedure. For a well-posed estimation procedure, the residuals are independent and uncorrelated across spots on a chip and across replicated experiments. Thus, the linear correlation of $\hat{\mathbf{R}}^{elr}$ and $\hat{\mathbf{R}}^{e'l'r'}$ is approximately zero when $e,e' \in \{A,B\}$ and $(e,l,r) \neq (e',l',r')$—the correlation is 1 when $(e,l,r) = (e',l',r')$ because these are the same array. An idealization of this expected behavior is depicted graphically in the correlation matrix in Fig. 4*A*.

Experiments C and D are designed to measure the differences in expression levels in the two samples. We expect $\hat{\mathbf{R}}^{e\cdot\cdot} \approx \mathbf{R}^{e\cdot\cdot}$, where $e \in \{C,D\}$. Deviations of $\hat{\mathbf{R}}^{e\cdot\cdot}$ from $\mathbf{R}^{e\cdot\cdot}$ are of the same order and have the same properties discussed above for the residuals of experiments A and B. It follows then that the correlation of $\hat{\mathbf{R}}^{elr}$ and $\hat{\mathbf{R}}^{e'l'r'}$ is large when $e,e' \in \{C,D\}$ and $(e,l,r) \neq (e',l',r')$.

Biases in MAQC Data. ***Spatial biases.*** Following recent reports of sequence-specific dye bias (2, 3, 11), we investigated whether estimates of R_i depend on i. The MAQC project provided an ideal opportunity to evaluate this effect and, if necessary, to identify probe sequences for which the effect is prominent. We calculated $\mathbf{R}$ for protocols A and B for lab 1 and used these to screen for spots that may be affected by this bias. Because these protocols are not measuring any change in expression, components of $\mathbf{R}$ should be zero-centered Gaussian variates. Following from the central limit theorem, the distribution of the sum of a sample of 10 instances (chips GSM128989–GSM128998) of R_i^0 is also a zero-centered Gaussian variable. Indeed, we find this to be true in general, but that this distribution has heavier tails than one would expect. Because our model does not address this problem directly, we restricted the set of spots used to compute these correlations using a Monte Carlo approach. From the 10 sampled chips, we permuted the spot assignments of the R_i values, thereby removing any sequence dependence of the estimates. Then we calculated the mean value $\langle R_i^{\text{shuf}} \rangle$ across the 10 chips. The resulting distribution of the means of $\langle R_i^{\text{shuf}} \rangle$ is equivalent to the distribution of the *actual* $\langle R_i \rangle$ if there exists no sequence-specific dye bias. As such, we restricted the set of spots we used for the correlation matrices to those with $\langle R_i \rangle$ falling within the 95% confidence interval of the distribution of $\langle R_i^{\text{shuf}} \rangle$.

By plotting the dependence of spot nonfeature fluorescent intensity on chip position, we found substantial spatial trending in the chips in this study (Fig. S2). Because the specific mechanism of this trend is difficult to determine, and because it is difficult to estimate its effect on F_i, we determined parameters separately for each quadrant of the chips.

Intensity-dependent dye bias. Plots of $p(\log E^1 E^2, R)$, often called *MA plots*, are often considered during microarray experiments. When they exist, nonlinearities are typically approached in one of two ways: (*i*) thresholding of weakly expressed genes and (*ii*) nonlinear detrending. Next, we discuss both of these approaches.

Most microarray feature extraction software packages flag spots as "present" or "absent," depending on how close the spot feature intensity is to the spot nonfeature intensity. This designation, which is consistent with the classical concept of the spot nonfeature intensity as a measure of additive background, suggests whether a signal is present in sufficient quantity that it is recognizable over the local additive noise. Absent flags at a spot are used to demark low-quality estimates of expression change. Recognizing that in practice, these spots are often excluded from analysis, we computed correlation matrices as in Fig. S3, excluding spots marked with absent flags (Fig. S4). It is important to note that, typically, a substantial fraction of spots are marked as such—43% of the spots in the MAQC arrays. If this method is used to filter estimates, nearly half of the potential information obtained in the experiment is lost.

Locally Weighted Regression (LOWESS) is a scatterplot smoothing algorithm commonly used to detrend intensity-dependent dye bias in microarray data (12). Indeed, this procedure does improve R_i estimates, but at the cost of mechanistic interpretation and perhaps overfitting the data (Fig. S3). Furthermore, LOWESS detrending inappropriately adjusts saturated data, hiding these effects from the user and potentially misconstruing the results.

Analysis of Resveratrol Data. Because the experimental design was slightly unbalanced due to poor quality RNA for one of the HCR chips, we removed one SD chip and one HC chip from our analysis (13). Incidentally, we suspect that the database labeling for one SD chip (GSM140958) and one HC chip (GSM140962) may have been inadvertently switched (Fig. S6). As such, we removed these from further analysis.

We did not find spatial correlations for these chips, as we did with the MAQC chips, so this correction was unnecessary. However, we found that sequence-specific dye bias was prevalent with these chips, so we eliminated affected spots as described above using the SD chips as training.

1. O'Neill P, Magoulas GD, Liu XH (2003) Improved processing of microarray data using image reconstruction techniques. *IEEE T Nanobiosci* 2:176–183.
2. Rosenzweig B, et al. (2004) Dye-bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ Health Persp* 112:480–488.
3. Kelley R, Feizi H, Ideker T (2008) Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood. *Bioinformatics* 24:71–77.
4. Margaritis T, et al. (2009) Adaptable gene-specific dye bias correction for two-channel DNA microarrays. *Mol Syst Biol* 5:266.
5. de la Nava JG, van Hijum S, Trelles O (2004) Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. *Stat Appl Genet Mo B* 3:11.
6. Lyng H, et al. (2004) Profound influence of microarray scanner characteristics on gene expression ratios: Analysis and procedure for correction. *BMC Genomics* 5:10.
7. Bengtsson H, Jonsson G, Vallon-Christersson J (2004) Calibration and assessment of channel-specific biases in microarray data with extended dynamical range. *BMC Bioinformatics* 5:177.
8. Dodd LE, Korn EL, McShane LM, Chandramouli GVR, Chuang EY (2004) Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics* 20:2685–2693.
9. Shi L, et al. (2005) Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics* 6:S11.
10. Shi LM, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24:1151–1161.
11. Elkon R, Agami R (2008) Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS Comput Biol* 4:e1000189.
12. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74:829–836.
13. Baur JA, et al. (2006) Resveratrol improves health and survival of mice on a high-calorie diet. *Nature* 444:337–342.

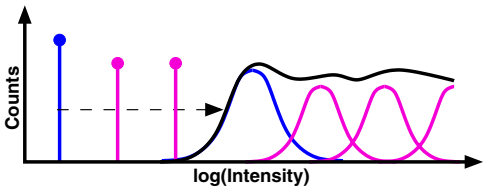

Fig. S1. The microarray experiment is a process that converts gene expression levels to observed feature intensities. Observed intensities F_i are a function of E_i and stochastic noises that convolute the readout. We observe the black distribution $\{F\}$ but wish to estimate $\{E_i\}$ (the points at left). To do this one must determine the parameters that define these stochastic fluctuations.

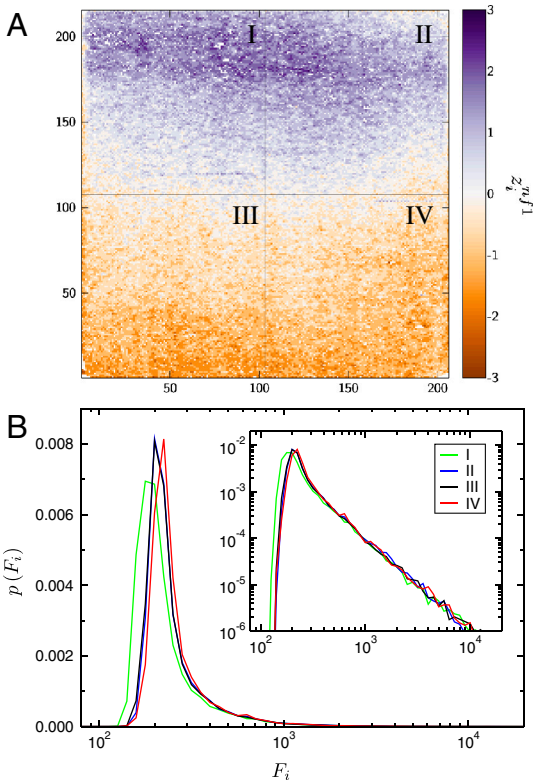

Fig. S2. Spatial bias in MAQC data. (*A*) The nonfeature intensity (standard score, z_i^{nf1}) has a strong dependence on the position of the spot in the chip. (*B*) The distributions $p(F_i)$ are drastically different for the four quadrants (I–IV) of the chip. (*B Inset*) Log-linear representation of the distributions. To address this problem, we determine model parameters for each quadrant separately.

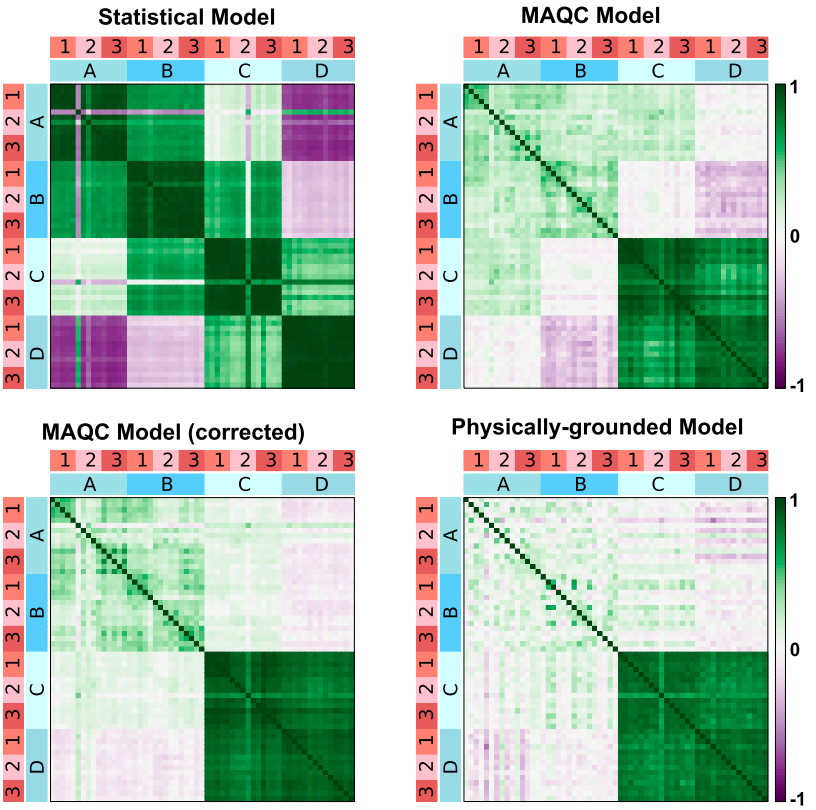

Fig. S3. Correlation of R_i in the MAQC project (Fig. 4). Pairwise correlations of $\hat{R}_i$ estimates derived from the statistical model, the MAQC model (10), the lowess-corrected MAQC model, and our physically grounded model. Our model imparts a much higher signal-to-noise ratio than the other models.

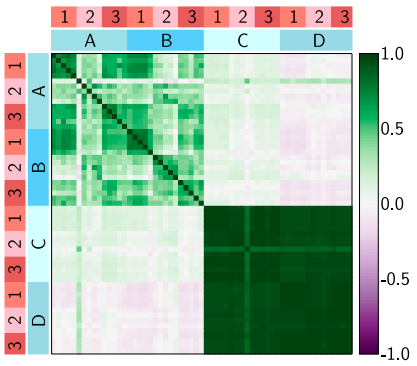

Fig. S4. A common strategy for improving expression estimate correlations for statistical models. Correlation matrix for MAQC arrays (see Fig. 4) using only spots flagged present. This approach reduces the information extracted from the array experiment by nearly a factor of 2.

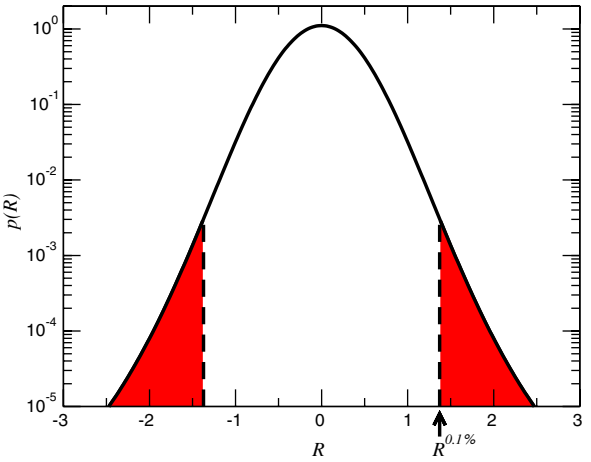

Fig. S5. For any level of expression, E_i, one can derive the corresponding distribution $p(R^0|E_i)$ to accompany the null model that there is no expression change. Imparting significance criteria to this distribution for each gene allows one to establish the significance of any observed expression change (Fig. 5). Traditionally, this exercise has required arbitrary thresholding in the absence of a model-based expression for how experimental fluctuations vary with E.

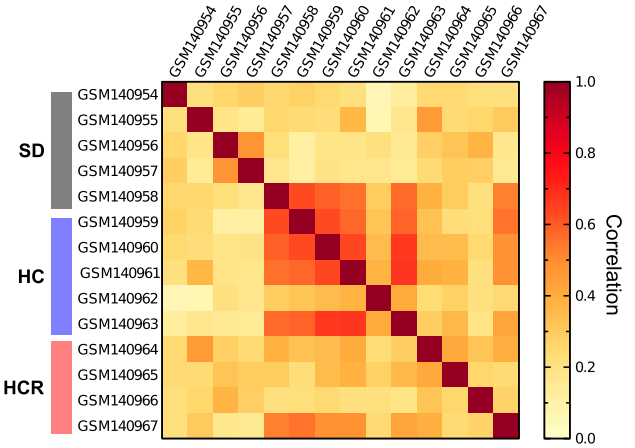

Fig. S6. There exists a possibility that the database labels on two of the chips (GSM140958 and GSM140962) for the Baur data (13) may have been inadvertently switched in the Gene Expression Omnibus database. Note that chip GSM140958 is much more similar the HC chips than to the SD chips. Since we could not unambiguously attribute these chips to either class, and because their removal balances the experimental design, we removed them from downstream analysis.

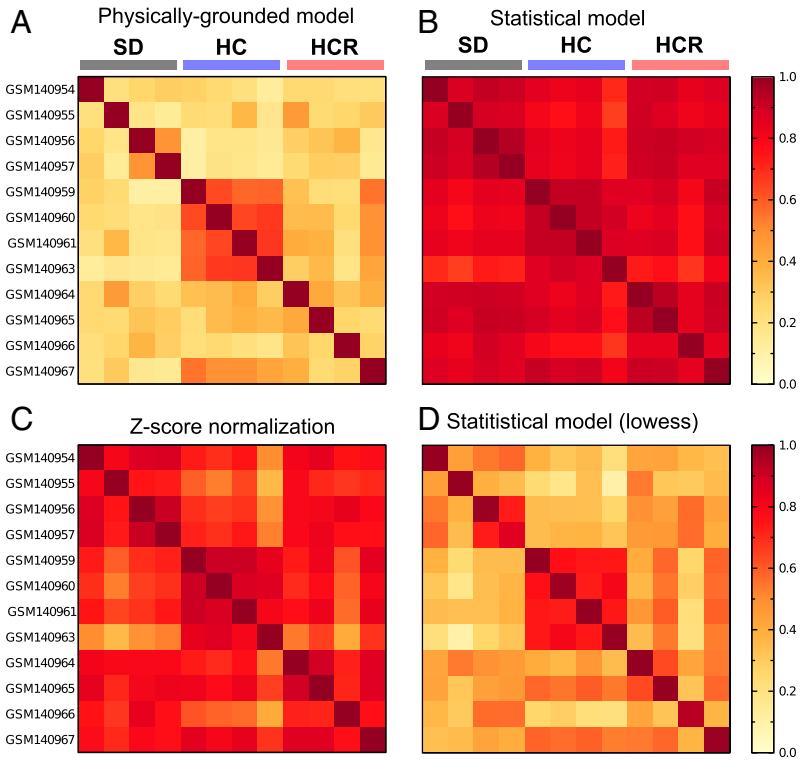

Fig. S7. Correlation matrices for Resveratrol chips (Fig. 6*B*). We computed the pairwise correlation of $\hat{R}$ for samples derived from animals fed standard diet (SD), high-calorie diet (HC), and high-calorie diet supplemented with resveratrol (HCR) for four different models. The physically grounded model yields much higher statistical power, as evidenced by the high correlation of HC chips, relative to the correlations of SD and HCR chips.