SUPPORTING INFORMATION

Additional Terms for Study-Specific Models

The models considered in the main text included probe-specific adjustment variables and intensity-dependent array effects and dye effects. There exist other technical factors that have been shown to influence intensities in microarray data. One of the most important of these are effects are related to the nucleotide composition of the probe sequences (Naef and Magnasco, 2002). Early work in the microarray analysis field demonstrated that models based on the physical chemistry of DNA-RNA binding identified robust and consist effects related to probe composition (Zhang *et al.*, 2003). Here we demonstrate how to extend the framework described in the main text to include such effects.

Let $\mathbf{s}_i$ denote the nucleotide sequence of probe i, $\mathbf{S}$ consist of the set of sequences for all probes, and g_{uj} be the function that maps s_i to the uth probe sequence effect on array j. For example, g_{uj} could describe the effect of having a given nucleotide at each position in a 25-mer probe sequence for array j, or it could describe the effect of probe sequence i being b bases away from the 3' end of its target RNA molecule. The u-th probe sequence effect can be parameterized as either a smooth function of some variable (e.g., position in probe sequence) or as a scalar shift, such that for a given j, the $\mathrm{E}[g_{uj}(\mathbf{s}_i)] = 0$ for any u across all of $\mathbf{S}$. The model for each probe can be written as:

$$y_{ij} = \sum_{k=1}^{d} b_{ik}x_{kj} + \sum_{\ell=1}^{r_c} a_{i\ell}z_{\ell j} + \sum_{t=1}^{r_f} f_{tj}\left(m_{ij}\right) + \sum_{u=1}^{r_g} g_{uj}\left(\mathbf{s}_i\right) + e_{ij},$$

with all terms described above or in the main text. Similarly, we can write the model probe i data, $\mathbf{y}_i$, as :

$$\mathbf{y}_i = \mathbf{b}_i\mathbf{X} + \mathbf{a}_i\mathbf{Z} + \sum_{t=1}^{r_f} \boldsymbol{f}_t(\mathbf{b}_i\mathbf{X} + \mathbf{a}_i\mathbf{Z}) + \sum_{u=1}^{r_g} \boldsymbol{g}_u(\mathbf{s}_i) + \mathbf{e}_i,$$

where $\mathbf{g}_u$ = $(g_{u1}(\mathbf{s}_i), g_{u2}(\mathbf{s}_i), ..., g_{un}(\mathbf{s}_i))$. The entire data set $\mathbf{Y}$ can be written as:

$$\mathbf{Y} = \mathbf{BX} + \mathbf{AZ} + \sum_{t=1}^{r_f} \boldsymbol{f}_t(\mathbf{BX} + \mathbf{AZ}) + \sum_{u=1}^{r_g} \boldsymbol{g}_u(\mathbf{S}) + \mathbf{E}.$$

where all terms have been described. See Naef and Magnasco (2002) for an example of a model to account for probe sequence effects that easily translate to relevant $g_{uj}(\mathbf{s})$.

Example Construction of Biological and Adjustment Variables

Consider the *Array Effects Plus Batch Effects* simulation study in the main text. For this simulation, $\mathbf{X}$ parameterizes the membership in Group 2, while $\mathbf{Z}$ parameterizes both the probe intercept and batch terms. The matrices of probe-specific biological and adjustment variables are:

$$\mathbf{X} = \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right]$$

$$\mathbf{Z} = \left[\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}\right]$$

A similar construction of $\mathbf{X}$ and $\mathbf{Z}$ is possible for the main study designs encountered with microarrays, even a two sample time course study, for example.

Additional Details on the SNM Algorithm

Here we expand on several details and operating characteristics of the SNM algorithm.

Importance of iteratively updating the estimated set of null probes. All probes are estimated to be null for the first iteration of the SNM algorithm. (A refined initial estimate can be provided if the investigator has prior information from another experiment.) It can be seen that this initial estimate quickly becomes more accurate over the first several iterations of the algorithm. An important point is that failing to adaptively update the set of null probes leads to poor normalization. This reinforces the fact that the relevant biological signal should be taken into account when performing a normalization. It should be noted that some unsupervised normalization procedures effectively treat all probes as null probes.

As an example, we considered the *Array Effects Plus Batch Effects* simulation from the main text. We fit model 3 from the main text in two extreme cases. The first is where all probes are estimated to be null probes and only a single iteration of the SNM algorithm is performed. The second case is where only the true null probes are utilized in a single iteration of the SNM algorithm. The π_0 estimates resulting from these two cases are equal to 33% and 72%, respectfully. Figure 0 presents the error associated with the normalized data obtained from these two cases. There is a clear relationship between the errors and the simulated biological variable in the first case (Figure 0A), which is reflected by the biased π_0 estimate of 33%. Figure 0B shows that by utilizing only the true null probes we have removed the bias in the estimated array effects. This example demonstrates the importance of SNM being an iterative algorithm where the estimated set of null probes is refined and made more accurate, with the goal to come as close to the set of true null probes as possible.

Model for intensity-dependent effects. Within each iteration the algorithm estimates intensity dependent effects using a linear mixed effects model. The intensity dependent variables are parameterized using B-spline basis functions available through the `ns` function in the `splines` library of the `R` statistical software package. The coefficients applied to the B-spline basis functions are modeled as random effects. The remaining probe-specific adjustment variables are included as fixed effects. Due to the large amount of data available in most microarray studies we fit this mixed effects model to a data matrix that summarizes the probe-specific intensities. This matrix is formed as follows. The range of the data is split into K equally spaced bins. Denote the kth such bin s_k. The matrix of summarized intensities is then defined as m^*_{kj} = $|s_k|^{-1}\sum_{i\in s_k} m_{ij}$, where $i \in s_k$ implies that $\sum_{j=1}^{n} m_{ij}/n$ is within the intensity range spanned by s_k. Note the dimension of this matrix is determined by the number of bins, not the number of probes, which reduces the computational burden associated with estimating the random effects terms.

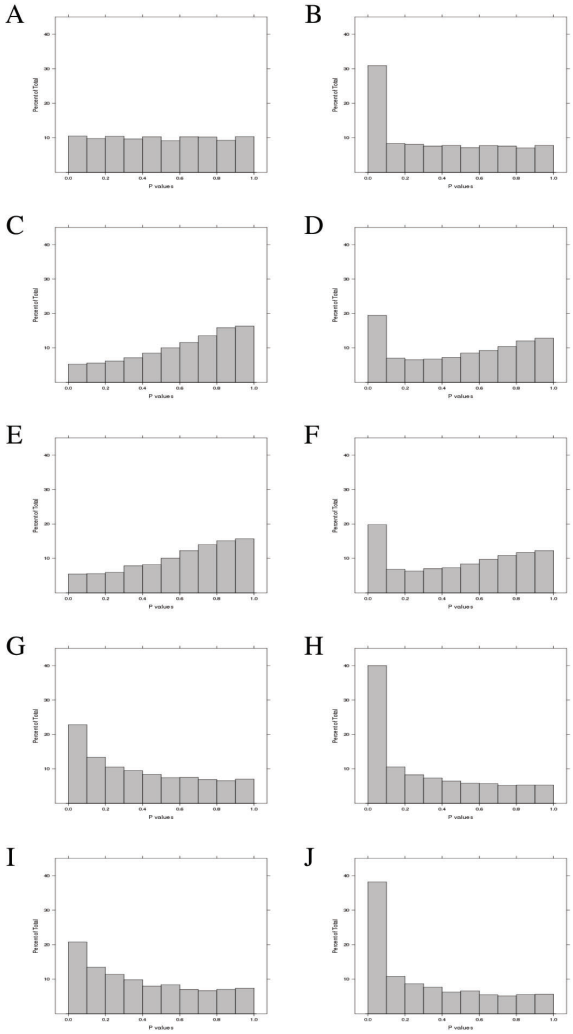
Adjustment variable coefficient estimation. Another noteworthy characteristic of our algorithm pertains to estimates of the coefficients of the adjustment variables, denoted by $\mathbf{A}$. Our algorithm is designed to produce unbiased estimates of the biological effects $\mathbf{BX}$ as well as residual variation independent from adjustment variables and technical factors. In doing so it is not necessary to produce unbiased estimates of the coefficients $\mathbf{A}$. It is only necessary to unbiasedly estimate the combined terms $\mathbf{AZ} + \sum_{t=1}^{r_f} \boldsymbol{f}_t(\mathbf{BX} + \mathbf{AZ})$ from the model $\mathbf{Y} = \mathbf{BX} + \mathbf{AZ} + \sum_{t=1}^{r_f} \boldsymbol{f}_t(\mathbf{BX}+\mathbf{AZ})+\mathbf{E}$ (equation 3 in the main text), for example.

To demonstrate this we present the intensity dependent relationship between two arrays from the same biological condition but different batches in Figure 0C. The solid grey line captures the relationship in the unnormalized data for these two arrays, the solid blue line the relationship after only the intensity-dependent array effect estimates have been removed, and the solid red line after the simultaneously estimated intensity-dependent array effect and batch adjustment variable have been removed. The solid black line is the line y = 0 and indicates that the technical effects have been removed. Notice that only removing the array effect produces data that still contain intensity dependent effects between the two arrays. However, the data have been rotated so that any two arrays from different batches have this difference. Simultaneously fitting the entire SNM model produces data with no remaining intensity dependent effects. Taken together, this figure demonstrates that $\hat{\mathbf{f}}(\hat{\mathbf{M}})$ and $\hat{\mathbf{A}}$ are biased, but their total quantity $\hat{\mathbf{A}}\mathbf{Z} + \hat{\mathbf{f}}(\hat{\mathbf{M}})$ is unbiased.

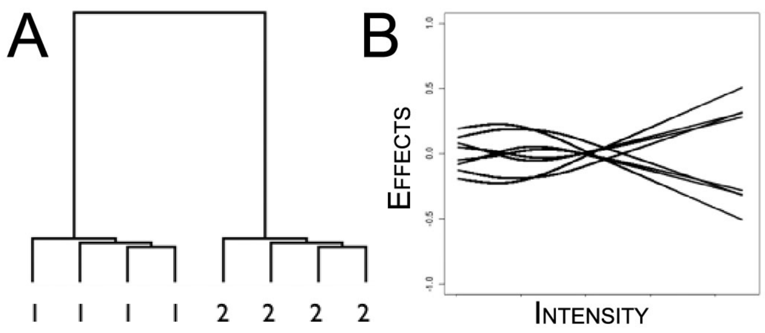
REFERENCES

Naef, F. and Magnasco, M. (2002). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Nucleic Acids Res Phys Rev E*, **68**, 011906.

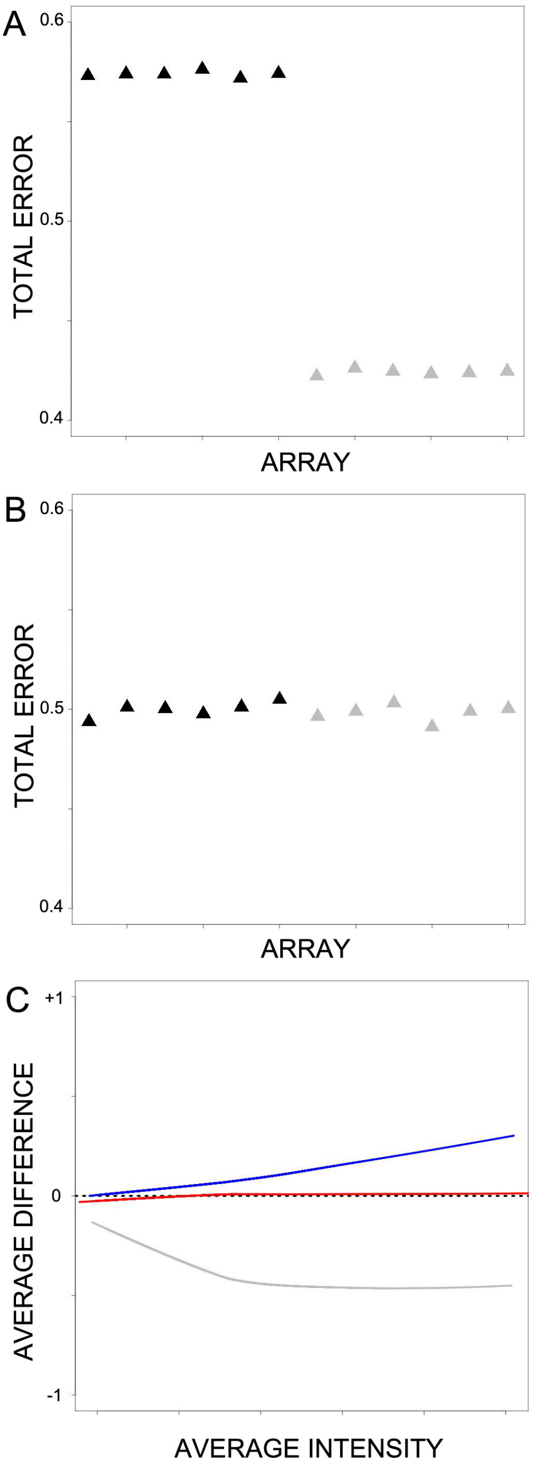
Zhang, L., Miles, M., and Aldape, K. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, **21**(7), 818–821.



Supp. Fig. 1. Results from simulated data with biologically relevant differential expression, batch, and array effects. The true proportion of null probes is π_0 = 0.70. (A) P-value histogram of null probes after SNM normalization. (B) P-value histogram of all probes after SNM normalization. (C) P-value histogram of null probes after QN normalization. (D) P-value histogram of all probes after QN normalization. (E) P-value histogram of null probes after ISN normalization. (F) P-value histogram of all probes after ISN normalization. (G) P-value histogram of null probes after QN normalization using a model that includes a term for batch. (H) P-value histogram of all probes after QN normalization using a model that includes a term for batch. (I) P-value histogram of null probes after ISN normalization using a model that includes a term for batch. (J) P-value histogram of all probes after ISN normalization using a model that includes a term for batch.



Supp. Fig. 2. Results from supervised approach used to identify study-specific variables for the vascular development study. (A) Hierarchical clustering of samples, where the distances between samples are defined on the residuals obtained by regressing the estimated biology from the observed, raw probe-level intensities (via $\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}$). The leaves are labeled according the batch for that array. (B) The relationship between the array-specific residuals and average intensity. These figures show that biology, batch, and intensity dependent array effects all influence the observed intensities.



Supp. Fig. 3. Operating characteristics of the SNM algorithm. (A, B) The total error between the normalized data and the optimal data. The black and grey colors describe the two conditions. (A) The total error after one iteration of the SNM algorithm when all probes are estimated to be null. (B) The total error after one iteration of the SNM algorithm using only the true null probes. (C) A demonstration that SNM accurately estimates the total quantity $\mathbf{AZ} + \sum_{t=1}^{rf} \boldsymbol{f}_t(\mathbf{BX} + \mathbf{AZ})$. The intensity dependent relationships between two arrays from the same condition but different batches using the raw data (grey line), the SNM normalized data after removing the array effects (blue line), and the SNM normalized data after removing the array and batch effects (red line). The dashed black line is the line y=0.