SUPPLEMENTARY MATERIAL

Rank invariant method and Lowess extrapolation

When the number of genes is small in comparative experiments such as the 125 gene project we select the rank invariant set

$$S = \{g: |\text{rank}(\text{Cy5}_g) - \text{rank}(\text{Cy3}_g)| < d \;\&\; l < \text{rank}[(\text{Cy5}_g + \text{Cy3}_g)/2] < G - l\}$$

to perform normalization.

If the number of genes is large, such as in the 4129 gene project, we can afford to apply the rank invariant method in an iterative manner to select a more conserved set of genes. Firstly we select

$$S_0 = \{g\,|\,\text{rank}(\text{Cy5}_g) - \text{rank}(\text{Cy3}_g)| < p \times G \;\&\; l < \text{rank}[(\text{Cy5}_g + \text{Cy3}_g)/2] < G - l\}$$

Then at each iteration we select

$$S_i = \{g: g \in S_{i-1} \;\&\; |\text{rank}_{g \in Si-1}(\text{Cy5}_g) - \text{rank}_{g \in Si-1}(\text{Cy3g})| < p \times |S_{i-1}|\}$$

where $|S_i|$ is the number of genes in set S_i. The iteration stops at the kth step when $|S_k| = |S_{k-1}|$ and the set of genes S_k is the chosen rank invariant set.

After selection of rank invariant set S we can fit a normalization curve $\hat{M} = \hat{f}(A)$ between $\min_{g \in S} A_g$ and $\max_{g \in S} A_g$ by the Lowess method. In order to normalize genes with an average log intensity $> \max_{g \in S} A_g$ we perform linear fitting $M = \alpha + \beta A$ in the subset $T = \{g: g \in S \;\&\; \text{rank}_{g \in S}(A_g) > |S| - 50\}$. The estimated linear fit is used to normalize genes with an average log intensity $> \max_{g \in S} A_g$. The same extrapolation procedure is also used to normalize genes with an averaged log intensity $< \min_{g \in S} A_g$.

Testing the homogeneity of slide variation

Perform hypothesis testing $H_0: \tau^2 = \tau_1^2 = \ldots = \tau_G^2$ versus $H_A: \tau_g^2$ *not all equal*. Assume y_{gse} to be the normalized log ratio of gene g, slide s and comparative experiment e and $y_{gse} \sim N(\mu_{ge}, \tau_g^2)$. Note that under the null hypothesis $\hat{\tau}_g^2 = \sum(y_{gse} - y_{g\cdot e})^2/S - 1 \sim \chi_{S-1}^2 \tau_g^2/S - 1$. Therefore, the statistic $t = \text{var}(\hat{\tau}_g^2)/\text{mean}(\hat{\tau}_g^2)^2$ converges in probability to $2/S - 1$. We compute the statistic t from R1S1, R1S2 and R2S1, R2S2 in the 4129 gene project and t = 9.2, 7.7, which obviously rejects the hypothesis.

Markov chain Monte Carlo (MCMC) procedures

Denote by x_{gse} the normalized log ratios of gene g, calibration slide s and calibration experiment e and by y_{gse} the normalized log ratios of gene g, slide s and comparative experiment e. We assume $y_{gse} \sim N(\mu_{ge}, \tau_g^2)$ and $\mu_{ge} \sim N(\theta_g, \sigma_g^2)$, where θ_g measures the true log-fold change in gene g. We pool information across the calibration slides to obtain a prior distribution for the slide effect variance $\tau_g^2: \tau_g^2 \sim k\tilde{\tau}_g^2/\chi_k^2$, where $\tilde{\tau}_g^2 = [(S-1) \times E \times \hat{\tau}_g^2 + \hat{\tau}_A^2)]/(S-1) \times E + 1$ is the weighted value of gene-specific and overall sample variances obtained from calibration slides. Here $\hat{\tau}_g^2 = \sum_{s,e}(x_{gse} - x_{g\cdot e})^2/(S-1)E$, $\hat{\tau}_A^2 = \sum_{g,s,e}(x_{gse} - x_{g\cdot e})^2/G(S-1)E$ [$x_{g\cdot e} = \text{mean}_s(x_{gse})$], G is the total number of genes, S is the number of slides, E is the total number of calibration experiments, χ_k^2 is the χ^2 distribution with degrees of freedom k and k is an adjustable degree of freedom. We observed that, on average, the between-slide variation in comparative experiments is 50% larger than that in calibration experiments for the 125 gene project. To account for this, we multiply $\tilde{\tau}_g$ by 1.5 in the 125 gene project to account for the increased variation in the comparative experiment. Similarly, the prior distribution for σ_g^2 is given by $\sigma_g^2 \sim h\tilde{\sigma}_g^2/\chi_h^2$, where $\tilde{\sigma}_g^2 = (E \times \hat{\sigma}_g^2 + \hat{\sigma}_A^2)/E + 1$. Here $\hat{\sigma}_g^2 = \sum_e x_{g\cdot e}^2/E$, $\hat{\sigma}_A^2 = \sum_{g,e} x_{g\cdot e}^2/GE$ and again χ_h^2 is the χ^2 distribution with degrees of freedom h and h is an adjustable degree of freedom. We note that $\tilde{\sigma}_g^2$ is biased upward as an estimate of σ_g^2. As a result, our procedure will tend to be conservative. In this paper we use the prior degree of freedom $k = h = 3$. The posterior distributions of the parameters do not have a closed form solution. Thus we apply the MCMC method (16) to simulate the distributions of these parameters.

Assume $y_{gse} \sim N(\mu_{ge}, \tau_g^2)$ with prior $\tau_g^2 \sim k\tilde{\tau}_g^2/\chi_k^2$ and $\mu_{ge} \sim N(\theta_g, \sigma_g^2)$ with prior $\sigma_g^2 \sim h\tilde{\sigma}_g^2/\chi_h^2$ where gene g = 1, 2, … G, experiment e = 1, 2, … E and slide s = 1, 2, … s_e. Data y_{gse} and $\tilde{\tau}_g^2$, $\tilde{\sigma}_g^2$, h, k are known. In order to obtain distributions of the parameters, an MCMC procedure has been developed.

(i) Compute $(\mu_{ge})^{(0)} = y_{g\cdot e}$.

(ii) Generate $(\sigma_g^2)^{(i)}$ from distribution $\sigma_g^2 | (\mu_{ge})^{(i-1)}$ where

$$\sigma_g^2 | \mu_{ge} \sim [\sum_e (\mu_{ge} - \mu_{\cdot g})^2 + h\tilde{\sigma}_g^2]/\chi_{E+h-1}^2$$

(iii) Generate $(\theta_g)^{(i)}$ from distribution $\theta_g | (\mu_{ge})^{(i-1)}$, $(\sigma_g^2)^{(i)}$ where

$$\theta_g | \mu_{ge}, \sigma_g^2 \sim N\{\mu_{g\cdot}, \sigma_g^2/E\}$$

(iv) Generate $(\tau_g^2)^{(i)}$ from distribution $\tau_g^2 | (\mu_{ge})^{(i-1)}, y_{gse}$ where

$$\tau_g^2 | \mu_{ge}, y_{gse} \sim [\sum_{j=1}^{E} \sum_{s=1}^{se} (y_{gse} - \mu_{ge})^2 + k\tilde{\tau}_g^2]/(\chi^2_{s1+\ldots+sE+k})$$

(v) Generate $(\mu_{ge})^{(i)}$ from distribution $\mu_{ge} | y_{gse}, (\tau_g^2)^{(i)}, (\theta_g)^{(i)}, (\sigma_g^2)^{(i)}$ where

$$\mu_{ge} | y_{gse}, \tau_g^2, \theta_g, \sigma_g^2 \sim N[(s_e y_{g\cdot e}\sigma_g^2 + \tau_g^2\theta_g)/(s_e\sigma_g^2 + \tau_g^2), (\tau_g^2\sigma_g^2)/(s_e\sigma_g^2 + \tau_g^2)]$$

(vi) Repeat procedures 2–5 N times. We found that N = 4000 is sufficient for mixing of the Markov chain whose steady-state distribution is the desired posterior distribution.

The above methodology can also be applied when calibration experiments are not available to provide prior information. In such cases we assume the same hierarchical model with prior distribution $\tau_g^2 \sim k\tilde{\tau}_g^2/\chi_k^2$ and $\sigma_g^2 \sim h\tilde{\sigma}_g^2/\chi_h^2$ where $\tilde{\tau}^2 = \sum_{g,s,e}(y_{gse} - y_{g\cdot e})^2/G(S-1)E$ and $\tilde{\sigma}^2 = \sum_{g,e}(y_{g\cdot e} - y_{g\cdot\cdot})^2/G(E-1)$ become non-gene-specific. When S and E are small relative to the prior degrees of freedom h and k the posterior distributions will tend to be non-gene-specific, while if S and E are large the posterior distributions are dominated by gene-specific observations.

Cancellation of non-linearity in reverse labeling

Assume that u_{1g}, v_{1g} are the Cy5 and Cy3 intensities of gene g on slide 1 and u_{2g}, v_{2g} the intensities on slide 2. In the reverse labeling design applying the ANOVA model (9) the logarithmic expression ratio of each gene is estimated as

$$1/2\{\log(u_{1g}/v_{1g}) - \log(u_{2g}/v_{2g}) - \text{mean}_g[\log(u_{1g}/v_{1g}) - \log(u_{2g}/v_{2g})]\}$$

which in calibration experiments should have a distribution centered at 0 and independent of the absolute intensity of the

gene. Note also that if ANOVA is adequate, $\log(u_{1g}/v_{1g}) - \text{mean}_g[\log(u_{1g}/v_{1g})$ should also behave in the same way.

In the first and second plots of Figure 8 log(Cy5) – log(Cy3) versus the average log intensity of Cy5 and Cy3 has a positive slope trend. This is probably due to the need for a normalization function between Cy3 and Cy5.

We decompose Cy5 intensity $u = \alpha v + \Delta(v)$, where $\Delta(v)$ is the non-linear part which cannot be explained by αv. By Taylor expansion

$$\log(u/v) = \log[\alpha + \Delta(v)/v] \approx \log(\alpha) + [\Delta(v)/v\alpha]$$

If the non-linear part $\Delta(v)$ on the two slides of reverse labeling are highly positively correlated the second term will be partially cancelled out in the reverse label average, as in the third plot of Figure 8.

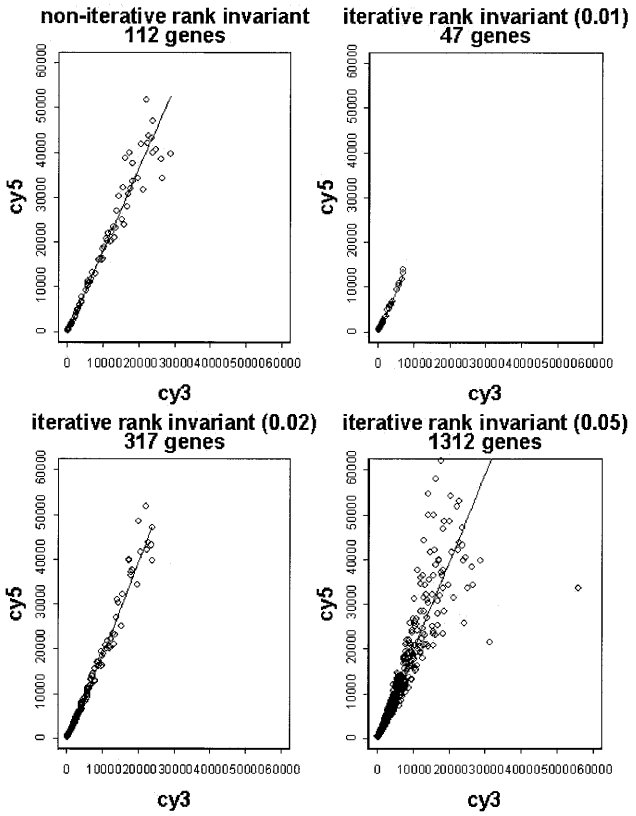

Figure S1. Intensity plots of genes chosen by non-iterative and iterative (with P = 0.01, 0.02, 0.05) rank invariant methods for R1S1 from the 4129 gene project, showing that iteration helps to select a more conserved set of genes.