BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1470–1476

Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays

*William J. Lemon, Jeffrey J.T. Palatini, Ralf Krahe and Fred A. Wright**

Division of Human Cancer Genetics, The Ohio State University, Columbus, Ohio, USA

Received on September 18, 2001; revised on April 22, 2002; accepted on April 30, 2002

ABSTRACT

Motivation: Oligonucleotide expression arrays exhibit systematic and reproducible variation produced by the multiple distinct probes used to represent a gene. Recently, a gene expression index has been proposed that explicitly models probe effects, and provides improved fits of hybridization intensity for arrays containing perfect match (PM) and mismatch (MM) probe pairs.

Results: Here we use a combination of analytical arguments and empirical data to show directly that the estimates provided by model-based expression indexes are superior to those provided by commercial software. The improvement is greatest for genes in which probe effects vary substantially, and modeling the PM and MM intensities separately is superior to using the PM–MM differences. To empirically compare expression indexes, we designed a mixing experiment involving three groups of human fibroblast cells (serum starved, serum stimulated, and a 50:50 mixture of starved/stimulated), with six replicate HuGeneFL arrays in each group. Careful spiking of control genes provides evidence that 88–98% of the genes on the array are detectably transcribed, and that the model-based estimates can accurately detect the presence versus absence of a gene. The use of extensive replication from single RNA sources enables exploration of the technical variability of the array.

Availability: Scripts for computing the Li–Wong reduced and full models are available in C, Splus and Perl in the supplementary information.

Contact: fwright@bios.unc.edu

Supplementary information: http://thinker.med.ohio-state.edu

INTRODUCTION

Oligonucleotide DNA arrays are a powerful means to monitor expression of thousands of genes simultaneously Lipshutz *et al.* (1999). However, important challenges remain in estimating expression level from raw hybridization intensities on the array. Li and Wong (Li and Wong, 2001a,b) recently introduced a statistical model which better fits observed patterns of hybridization than the models implicitly employed by standard commercial software (Affymetrix, 1999). The Li–Wong model was derived for Affymetrix GeneChip arrays, but the model-based approach is likely to be useful in other photolithography or ink-jet arrays (Hughes *et al.*, 2001) in which genes are represented by multiple oligonucleotide probes. Individual probe effects are large and systematic, and by explicitly fitting such effects the Li–Wong model presumably provides an improved index of gene expression (Li and Wong, 2001b). However, the extent of improvement has not been shown directly or explored in the context of extensive replication.

The term 'expression index' here describes a statistic intended to reflect expression level for a particular gene, whether or not it is based on an explicit model. We have evaluated the theoretical relative efficiency of competing gene expression indexes, and consider a framework for empirical index comparison. To provide this comparison, we conducted a carefully designed mixing experiment involving the response of human fibroblasts to serum. The experiment provides insight into technical variability and the number of expressed genes. Supplemental plots, primary data, perl and S-plus scripts and C programs for decoding GeneChip files are available at our web site http://thinker.med.ohio-state.edu. A longer version of the manuscript is also online and includes additional figures and detailed analyses of differentially expressed genes.

The current generation of photolithographic arrays (McGall and Fidanza, 2001) have 250 000–500 000 probes arranged in pairs—a perfect match (PM) probe that is complementary to a 25-base pair segment of mRNA and a mismatch (MM) probe that is complementary to the same mRNA segment except for the 13^{th} nucleotide. A collection of 16–20 probe pairs, called a *probe set*, is used to

*To whom correspondence should be addressed at Present address: Department of Biostatistics UNC-CH, 3107B McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420, USA

represent a gene.

Following Li and Wong (2001a), I denotes the number of samples in an experiment, and J denotes the number of probe pairs in a probe set (e.g. $J = 20$ for the Affymetrix HuGeneFL array described in this study). The PM and MM intensities for the ith sample and jth probe pair of a given gene are modeled as

$$\begin{aligned} PM_{ij} &= \nu_j + \alpha_j\theta_i + \phi_j\theta_i + e \\ MM_{ij} &= \nu_j + \alpha_j\theta_i + e, \end{aligned} \tag{1}$$

where θ_i is the expression index and ν_j is a non-specific cross-hybridization term for the jth probe pair. The term α_j is the rate of increase of MM intensity with expression, and ϕ_j is the additional rate of increase in the PM intensity. The errors e are assumed independent with variance ξ^2. We consider the θ_is as the parameters of interest, and estimates may be obtained via least-squares, the maximum likelihood solution if the errors are assumed to be normally distributed.

We refer to Model (1) as the Li–Wong 'full' model (LWF) because the PM and MM values are treated separately. However, most of the analyses in Li and Wong (2001a) are based on the Li–Wong 'reduced' model (LWR) using only the differences $y_{ij} = PM_{ij} - MM_{ij} = \phi_j\theta_i + \varepsilon$. The reduced model follows directly from Model (1), with $var(\varepsilon) = \sigma^2 = 2\xi^2$. Fitting either LWF or LWR requires an identifiability constraint (Li and Wong, 2001a), and we use $\sum_j \phi_j^2 = J$.

The reduced model has the advantage of fewer parameters than the full model, but potentially ignores information contained in the bivariate PM, MM data. This is especially true for probe pairs in which PM and MM are similarly sensitive to expression changes (i.e. α_j is large compared to ϕ_j). We have verified the large and systematic probe effects in several datasets and found numerous genes in which the reduced model is likely to perform poorly compared to the full model. Nonetheless, the reduced model offers considerable improvement over the most popular current indexes.

Average difference versus Li–Wong reduced

The MAS4 software (*Affymetrix Microarray Analysis Suite* version 4.0, Affymetrix, Santa Clara, Calif) computes the 'average difference' (AD), the simple average of the PM–MM differences across a probe set. Assuming Model (1) holds, we can contrast AD with LWR (Li and Wong, 2001a), implemented in our software and in the program DCHIP (www.dchip.org). For the limiting situation with large sample sizes, the ϕs will be estimated with great precision. If we assume the ϕs and σ^2 to be known, then for a single sample (the subscript i is suppressed) the maximum likelihood estimate is unbiased and can be shown to be $\hat{\theta}_{\text{reduced}} = \sum_j y_j\phi_j/J$ (Li and Wong, 2001a). In contrast, AD (denoted η) is computed as $\hat{\eta} = \sum_j y_j/J$. AD is not constructed as an explicit estimate of θ, and thus has considerable bias unless all $\phi_j = 1$. Thus, AD can be made comparable to $\hat{\theta}_{\text{reduced}}$ by applying a correction factor to create a new index $\hat{\hat{\eta}} = \left(J/\sum_j \phi_j\right)\hat{\eta}$. It is easy to show that $E(\hat{\theta}) = E(\hat{\hat{\eta}}) = \theta$. The variances of the estimates can be computed directly as $\text{var}(\hat{\theta}_{\text{reduced}}) = \sigma^2/J$ and

$$\text{var}(\hat{\hat{\eta}}) = \frac{J^2}{\left(\sum\phi_j\right)^2}\cdot\text{var}(\hat{\eta}) = \frac{\sigma^2 J}{\left(\sum\phi_j\right)^2}.$$

It can be shown that the variance of the ϕs across the J probe pairs is $var(\phi) = E(\phi^2) - E^2(\phi) = 1 - \left(\sum\phi_j\right)^2/J^2$, so $\left(\sum\phi_j\right)^2 = J^2\,(1 - var(\phi))$. Finally we obtain the relative efficiency

$$\begin{aligned} RE(\text{reduced}, AD) &= \frac{\text{var}(\hat{\hat{\eta}})}{\text{var}(\hat{\theta}_{\text{reduced}})} = \frac{J^2}{\left(\sum\phi_j\right)^2} \\ &= \frac{J^2}{J^2\,(1 - \text{var}(\phi))} = \frac{1}{1 - \text{var}(\phi)}, \end{aligned}$$

showing that LWR is superior to AD. Although the ϕs are considered fixed, $\text{var}(\phi)$ is the variance of these sensitivities across the J probe pairs. The efficiency result is quite sensible, because the Li–Wong reduced model accounts for the variability in probe pair effects, while the average difference does not.

Li–Wong full model versus reduced model

For the Li–Wong full model, solving the likelihood equation and computing the variance gives (see Supplementary Appendix)

$$\text{var}(\hat{\theta}_{\text{full}}) = \frac{\sigma^2}{2\left[\sum_j \alpha_j^2 + \sum_j (\alpha_j + \phi_j)^2\right]}$$

and the relative efficiency compared to LWR:

$$\begin{aligned} RE(\text{full,reduced}) &= \frac{\text{var}(\hat{\theta}_{\text{reduced}})}{\text{var}(\hat{\theta}_{\text{full}})} \\ &= 2\left[\frac{\sum_j \alpha_j^2 + \sum_j (\alpha_j + \phi_j)^2}{J}\right]. \end{aligned}$$

It can be shown that $RE(\text{full, reduced}) \geqslant 2$, with equality holding when all of the αs are zero. This corresponds to the situation in which the mismatch probe pairs are not at all sensitive to gene expression, and so they simply add noise in computing the difference PM–MM (thus doubling the variability). In the case where each $\alpha_j = \phi_j$, the relative efficiency can be computed to be 10.0 for any

J. In practice, the result may be even more extreme, as the αs are often greater than the ϕs (see estimates in supplemental data). We believe the possible gains from using the full model are not generally appreciated.

The Log-average

Another expression index is the Log-Average (LA), computed by MAS4 as $10\left[\sum_j \log(PM_j/MM_j)\right]/J$. This index may perform poorly, as the Li–Wong model indicates that hybridization intensity increases proportionally to expression for both PM and MM. Thus, for a gene in which the nonspecific hybridization terms ν are small and the error variance is small (an otherwise favorable situation), we have $(PM_j/MM_j) = (\theta\alpha_j + \theta\phi_j)/(\theta\alpha_j)$, which does not depend on θ. This result and the empirical results below suggest that LA is of limited value.

In addition to providing expression indexes, MAS4 attempts to make an absolute determination of the presence/absence of each gene (Affymetrix, 1999). In our experience, typically fewer than 50% of the genes in an array are called 'present' by MAS4. We present evidence below that in fact the vast majority of genes on the array are expressed at detectable levels. MAS4 identifies numerous probe pairs in advance as potentially unreliable, and trims additional probes having outlying intensities before calculating AD and LA. Outlier detection is also a major feature in Li and Wong (2001a). Under the carefully controlled conditions of the present study, consideration of outliers has a minor effect and such trimming was not performed in our calculations.

MATERIALS AND METHODS

As shown in Figure 1, human fibroblast cells (GM 08330; Coriell Cell Repositories) were grown according to the distributor's recommendations in media supplemented with 20% FBS for 5 passages (27 flasks). Cultures were placed in serum-reduced media (0.1% FBS) for 48 h. After 48 h, 9 flasks were returned to 20% serum condition (Stimulated) and cells from the other flasks (Starved) were placed in RNA-Stat60 (Qiagen, Valencia, CA) according to manufacturer's instructions. More RNA is produced per cell in the stimulated condition, so fewer flasks of stimulated cells were needed to produce the same amount of RNA as those of starved cells. Twenty-four hours later, stimulated cells were harvested and placed in RNA-Stat60. Total cellular RNA was extracted using phenol:chloroform and was purified using RNeasy (Qiagen) according to manufacturer's specification. Extraction produced one Stimulated sample and one Starved sample. A third RNA sample (50:50) was produced from an equimolar mixture of these two.

On each of three days, two aliquots of RNA were taken from each group and processed separately as previously described (Virtaneva *et al.*, 2001; detailed protocols at

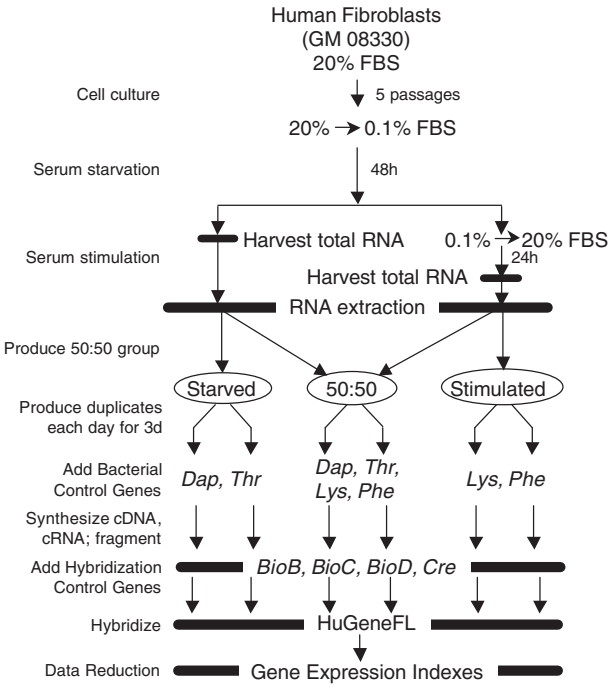

Fig. 1. Design of the replication experiment. RNA was derived from serum-starved and serum-stimulated fibroblasts in a single extraction, and a 50:50 mixture created. Six replicate arrays were hybridized in each group.

http://www.cancergenetics.med.ohio-state.edu/microarray/uArrayProtocols.html). Modifications are as follows. Stimulated RNA samples received bacterial control genes *Lys* and *Phe* RNAs at 0.08 ng/8 μ g total RNA, the Starved samples received the same amount of *Phe* and *Thr* and the 50:50 samples all four at 0.04 ng/8 μ g. Hybridization controls *BioB*, *BioC*, *BioD*, *Cre* were added to final concentrations of 1.5, 5, 25 and 100 pM, respectively. Each HuGeneFL array was loaded with 11 μg/200 μL labelled cRNA. This produced six replicates for each sample (18 arrays) in a day-balanced procedure. To minimize external variability, all arrays were from the same lot. To test for a day effect, analysis of variance against median summaries for entire arrays were performed, as well as individual analyses for each gene. P-value plots revealed no substantial day effect. Perl scripts were used to decode the probe information contained in Affymetrix GeneChip CEL files. Model-fitting and statistical analyses were performed using Splus v. 5.0, 6.0 and 2000 (Insightful, Seattle).

Empirical comparison of expression indexes

The relative efficiency results help clarify the advantages of model-based approaches, although the results are dependent on the applicability of the Li–Wong model.

In practice, such dramatic efficiency improvements are not achieved, in part because the difficulty in scaling arrays across the range of intensity values adds additional variation to all expression indexes. We now describe an approach for comparing expression indexes based on empirical data.

$\hat{\theta}_{\text{full}}$ and $\hat{\theta}_{\text{reduced}}$ are derived from the same model, so it may be appropriate to compare their variances directly, e.g. across replicate arrays under a fixed experimental condition. Direct variance comparisons for other expression indexes may not be meaningful. Instead, we propose to judge disparate indexes on the basis of their correlation with the underlying *true* expression. Although the true expression is generally not known, we propose a mixing experiment allowing the explicit estimation of this correlation. The use of correlation coefficients has intuitive appeal, but can also be placed in the framework of relative efficiencies as outlined below.

Suppose the true underlying gene expression for a given gene is τ. Consider two indexes of gene expression, $\hat{\theta}$ and $\hat{\eta}$, where

$$\begin{aligned} \hat{\theta} &= \beta_0 + \beta_1\tau + e_\theta, \quad e_\theta \sim N(0, \sigma_\theta^2) \\ \hat{\eta} &= \delta_0 + \delta_1\tau + e_\eta, \quad e_\eta \sim N(0, \sigma_\eta^2) \end{aligned} \tag{2}$$

In other words, we assume that each expression index has a linear relationship with the true gene expression and an uncorrelated error term. In practice the coefficients will not be known, but note that rewriting the estimate as $\hat{\hat{\theta}} = (\hat{\theta} - \beta_0)/\beta_1$ gives an unbiased estimate of τ, and $var(\hat{\hat{\theta}}) = \sigma_\theta^2/\beta_1^2$. Similarly, the variance of the unbiased estimate $\hat{\hat{\eta}} = (\hat{\eta} - \delta_0)/\delta_1$ is σ_η^2/δ_1^2, for an overall relative efficiency of $\hat{\theta}$ compared to $\hat{\eta}$:

$$RE(\hat{\theta}, \hat{\eta}) = \frac{var(\hat{\hat{\eta}})}{var(\hat{\hat{\theta}})} = \frac{\sigma_\eta^2/\delta_1^2}{\sigma_\theta^2/\beta_1^2}.$$

We also note that the overall variance of $\hat{\theta}$ is $var(\hat{\theta}) = \beta_1^2 var(\tau) + \sigma_\theta^2$, so that the ratio of explained to residual variance in the model is $ER_\theta = \beta_1^2 var(\tau)/\sigma_\theta^2$. This ratio is $r^2/(1 - r^2)$, where r is the Pearson correlation coefficient in the regression of $\hat{\theta}$ on τ. Similarly, $ER_\eta = \delta_1^2 var(\tau)/\sigma_\eta^2$, and we have $ER_\theta/ER_\eta = RE(\hat{\theta}, \hat{\eta})$, with cancellation of the $var(\tau)$ term. Thus ER_θ and ER_η reflect the respective efficiencies of the indexes, quantities that can be directly estimated from the regression of the expression indexes on τ. The quantities can be equivalently (and importantly) estimated using any predictor variable that is a linear transformation of τ. We show below how ER_θ can be estimated from empirical data using the following experimental design.

RESULTS

A mixing experiment with replication

Consider a mixing experiment involving two different RNA sources (samples), A and C. After a single extraction of mRNA from each sample, a third sample (B) is formed as a 50:50 mixture of A and C, so that the true expression for each gene in sample B is the average of that in samples A and C. Replicate arrays are then prepared for each of the three samples. As described above, we can perform a regression of any expression index on a suitable predictor variable (e.g. $x = 1, 2, 3$ for groups A, B and C, respectively), with the assurance that x is a linear transformation of the unknown true expression τ. Within this framework, the magnitude of the correlation coefficient for each gene is equal to that of Model (2), i.e. $r^2_{\hat{\theta},\tau} = r^2_{\hat{\theta},x}$.

Bioscaling

To make arrays comparable, most studies have used forms of linear or nonlinear global scaling (Li and Wong, 2001b). However, such global scaling is inappropriate if a large portion of the genes are substantially up- or down-regulated in a set of arrays, as such scaling will largely eliminate this important feature of the data. We observed that median hybridization intensities were 30% lower in Starved than Stimulated (supplemental data, $p < 0.002$, Kruskal–Wallis test). As mRNA is thought to represent only 2–8% of total RNA, this is clear evidence of reduced transcription in the Starved group. To scale arrays while preserving evidence for large scale changes in transcription, we used the *Bio* hybridization control genes added to all samples at constant concentrations (**Materials and Methods**). Probe intensities were divided by the average intensity of these probe sets, a procedure we term *bioscaling*. This moderately reduced the median within-group coefficients of variation for all genes—$\sim$ 6% for LWF and 11.8% for LWR. Except where otherwise indicated, our analyses were performed using the LWF and LWR indexes after bioscaling, and AD and LA after the (essentially linear) scaling using MAS4 (Affymetrix, 1999).

Efficiency and clustering results

Pairwise correlations among the arrays are shown as supplemental figures. Median within-group coefficients of variation (std dev/|mean|, LWF 16.2%, LWR 14.4%, AD 26.3%, LA 27.9%) are consistent with the apparent lower heterogeneity of model-based estimates. Figure 2 shows that model-based estimates also improve the precision in estimating expression. With or without scaling, the efficiency $r^2/(1 - r^2)$ follows LWF > LWR > AD > LA. The reduction in efficiency after bioscaling reflects overcorrection in the Stimulated

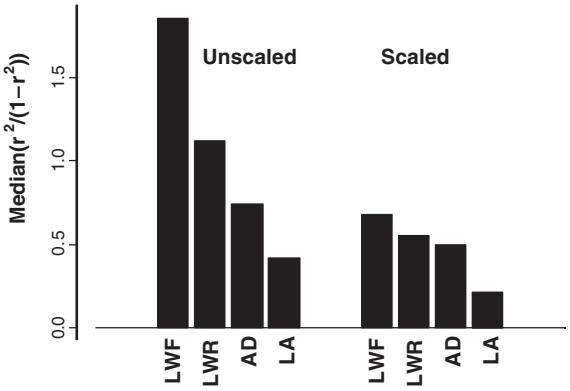

Fig. 2. Median efficiency of 7129 gene expression indexes, according to criterion derived in text. Unscaled indexes were computed from CEL files by the authors without probe trimming. Scaling of AD and LA was performed by MAS4 using default settings. Scaling of Li–Wong models was performed using bioscaling as described in text. Both types of scaling reduce between-group variation (and thus efficiency) because of the global difference in mRNA expression in Stimulated versus Starved.

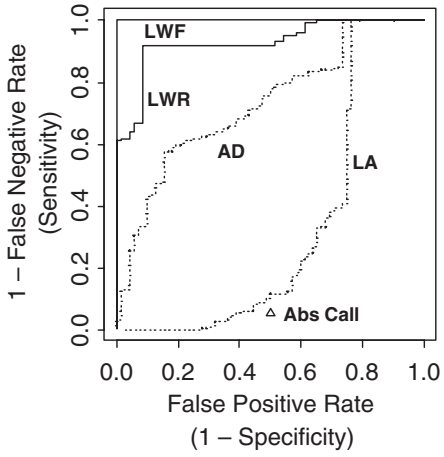

Fig. 3. ROC curves, using varying thresholds of expression indexes in declaring spiked bacterial control genes present/absent. Lines represent LWF, LWR, MAS4 AD and MAS4 LA. Triangle represents the overall assessment of the MAS4 absolute call.

group. The globally increased expression in Stimulated produces increased cross-hybridization, producing higher expression estimates for the *Bio* control genes. We note that inferences based on the unscaled data are valid, and view the efficiency comparisons based on the scaled data to be conservative.

Hierarchical clustering of samples was performed for each model (supplemental figures) using the 2104 genes called 'Present' by MAS4 in at least 15 of the samples. The full model classifies the samples into three distinct groups corresponding to the three mRNA groups, while the reduced model and MAS4 estimates each produce a classification error.

Duplicated genes

A total of 149 genes are represented twice (or more) on the HuGeneFL array, although not necessarily with the same set of probes. The median correlation of each pair of duplicated genes across the 18 samples was substantially higher for the model-based estimates (LWF median $r = 0.74$, LWR median $r = 0.43$, AD median $r = 0.12$, LA median $r = 0.09$).

Present versus absent calls

The simple z-statistic $\hat{\theta}/SE(\hat{\theta})$ can be used to test for whether a gene is 'present' (i.e. expressed). The z-statistic is related to the Wald statistic (Cox and Hinkley, 1974), except that in our implementation $SE(\hat{\theta})$ is the conditional standard error from the likelihood equation, i.e. assuming the nuisance parameters known (Li and Wong, 2001a). Such an approach tends to underestimate the standard error, so that fairly large z-statistic values may be necessary to call a gene present. This is especially true for LWF, which contains many more parameters than LWR. In our fibroblast experiment, the spiked bacterial control genes (*Lys, Phe, Dap, Thr*) serve as samples of genes known to be expressed and unexpressed (a total of 12 probe sets and 18 samples, further discussion below). Using the criterion $\hat{\theta}/SE(\hat{\theta}) > 5.0$, we obtain an error rate for LWF of 3.2% (144/144 probe sets properly called present, 65/72 properly called absent). All but one of the misclassified observations were due to a single probe set. Using the same criterion, the error rate for LWR was 9.3%. In contrast, the error rate of the MAS4 present/absent calls was a much larger 79.7% (8/144 properly called present, 36/72 properly called absent).

In a more comprehensive comparison, we present receiver–operator characteristic (ROC) curves in Figure 3 for detecting the control genes as present/absent. The z-statistics based on LWF and LWR are able to detect the genes with high sensitivity and specificity. Moreover, using the θ estimates directly gives very similar results, with the ROC curves almost coinciding with the respective curves for the z-statistics. For comparison we also plot the curves for AD and LA. The AD and LA values calculated by MAS4 are very poor—the LA curve is worse than chance variation (the 45° line) for much of the curve. The curves based on our own calculations of AD and LA perform somewhat better than those provided by MAS4. We have determined that this effect is not due to the scaling approaches used, and we speculate that MAS4 masking and probe trimming may in certain cases produce inferior estimates. The present/absent absolute calls from MAS4 appear as a single point.

Few unexpressed genes

Our implementation of the Li–Wong models does not constrain the parameter estimates to be positive. For LWF and LWR, only about 0.2% of the probe sets in the entire dataset have a negative expression index. This is lower than in other studies, but in our experience about 1–5% are negative. Intriguingly, this percentage is only slightly higher for the Starved group than for Stimulated, and is lowest for the 50 : 50 group (consistent with a smaller set of genes unexpressed in both Starved and Stimulated). The overall percentage of negative values for AD and LA was near 20%. Moreover, the number of genes we called 'present' is very high—approximately 7000 for each array. The results suggest that the model-based estimates may be extremely sensitive to genes expressed at low levels, and that very few genes on the array are unexpressed or undetectable by model-based estimates. We consider the spiked out genes as samples of 'unexpressed' genes under these conditions. Because each of these control genes is represented by three non-overlapping probe sets, these provide 6 independent unexpressed genes in each of the Starved and Stimulated conditions. From symmetry considerations, one would expect a gene to have $\hat{\theta} < 0$ with 50% probability in a sample in which the gene is truly unexpressed. The results are roughly in line with expectations—LWF produced negative indexes for 42% of the instances in which the control genes were spiked out, while LWR produced 35% negative.

As an additional, conservative approach to estimating the number of expressed genes, we treat the unexpressed genes as a contaminating population of unknown size. Let U denote the unknown number of unexpressed genes. If all genes are ranked from lowest to highest apparent expression using an expression index, then

$$U \leqslant 2 \times (\text{median rank of } U \text{ genes among all genes}),$$

with equality if the populations are disjoint. From a random sample of unexpressed genes, we can calculate an upper confidence bound for the median rank, which in turn generates a conservative upper bound for U. Other possible statistics, such as the maximum rank, are overly sensitive to sporadic outliers. For six observations from a continuous distribution, the fifth order statistic forms an approximate 90% distribution-free upper confidence bound for the population median (Hollander and Wolfe, 1999).

The sample of unexpressed genes can be examined among all genes in Figure 4, with the variance versus the mean for LWF among Stimulated samples shown on the rank scale. The variance clearly is lower for genes with lower expression, and note that this relationship holds even for the genes with the lowest expression. The relationship between mean and variance for AD is curved—genes with rank lower than ∼1650 have negative AD values when averaged over the replicates. These plots suggest that accounting for probe effects can improve detection limits and precision for lowly expressed genes.

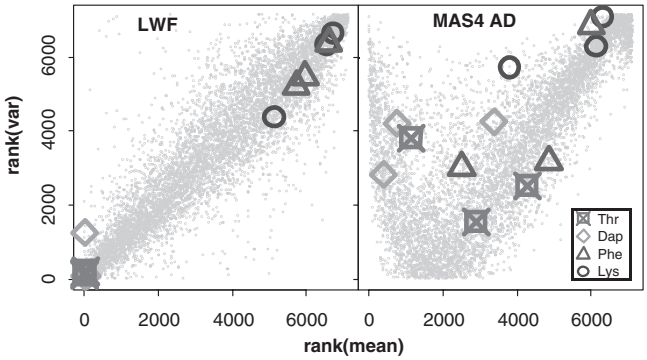

Fig. 4. Rank of expression index variance across the 6 Stimulated arrays versus rank of index mean. **Left:** Results from LWF with data points for spiked in/out control genes highlighted. The low ranks of *Dap* and *Thr* indicate that few genes are unexpressed. **Right:** MAS4 AD, with the same genes highlighted. *Dap* and *Thr* were truly absent, *Lys* and *Phe* truly present.

The estimates based on our sample of unexpressed genes are particularly revealing. Five of the six *Dap* and *Thr* genes have very low rank (50 or less) among the 7129 probe sets, while *Lys* and *Phe* show high ranks. The confidence bound procedure leads to a 90% upper bound for U as 100. Out of 7129 probe sets (and 6800 genes), this provocatively implies that over 98% of the genes on the array are expressed in Stimulated (consistent with our z-statistic criterion for present/absent calls). The relatively high ranks of spiked-out genes and low ranks of spiked-in genes using AD suggests a high background noise level, and AD may fail to detect many expressed genes. Similar results hold for the Starved group, with the rank-based procedure indicating that over 88% of the genes on the array are expressed.

Technical variability

We describe variation among our replicates as technical variation, with essentially standard protocols followed after extraction into three pools of RNA. Within the Stimulated replicates, the median coefficient of variation (standard deviation/|mean|; CV) for individual probe intensities was 12%, comparable to that of carefully constructed cDNA arrays (Yue *et al.*, 2001). Unscaled Stimulated expression indexes yielded median CVs as follows: LWF 14%, LWR 14.9%, AD 26.3%, LA 27.9%. The Starved replicates showed a similar pattern, but with higher CVs due to the reduced mean in the denominator. We have used the approximately linear relationship

between log(variance) versus log(mean) of the replicates to describe sample sizes necessary to two-fold changes in expression values, with conservative bounds designed to account for biological heterogeneity (further details in supplementary data).

DISCUSSION AND CONCLUDING REMARKS

We have developed a theoretical and experimental framework for evaluating indexes of gene expression in high-density oligonucleotide arrays. We have demonstrated the improvements provided by model-based estimates over simple averaging methods and proposed an improved and simple approach to present/absent gene calls. The model-based estimates and results of our spiking experiment suggest that the vast majority of genes on the array are expressed—thus the present/absent calls may not be a meaningful distinction for many genes. It will be important to investigate this phenomenon in a wider variety of tissues to better understand the relationship between quantity and activity of a gene. To our knowledge, our experiment involves more extensive replication than other studies reported thus far, and we hope that our data will serve as a useful resource for further investigations by statisticians and geneticists.

Several challenges remain in the evaluation of gene expression. We have noted that probe error variation appears somewhat dependent on hybridization intensity, suggesting that an expression-dependent error structure may offer further improvements. The Li–Wong approach involves fitting the probe strengths as fixed effects, and is technically possible using only a few arrays (e.g. $2IJ$ data points for LWF, with only $3J + I + 1$ parameters). However, the artificial constraints on the ϕs do not enable estimates of absolute expression, or even relative comparisons of different genes on the same array. A mixed-model approach incorporating random probe effects would enable the comparison of different genes. Our bioscaling procedure is a step towards scaling using absolute concentrations of control genes, and should be developed further using genes across a wide variety of spiked concentrations. Additional study of the sequence-dependence of probe sensitivity is also warranted. To further all of these efforts, it will be important that genetic researchers distribute primary probe intensity data (CEL files) with their published studies, so that further improvements may be explored.

ACKNOWLEDGEMENTS

We thank Timothy Wise, Gustavo Leone, Daolong Wong, Karl Kornacker, Wing H. Wong, and Cheng Li. Supported in part by NIH GM58934 and the Solove Research Institute.

REFERENCES

Affymetrix (1999) *Gene Chip Analysis Suite User Guide*. Affymetrix, Santa Clara, CA.

Cox,D.R. and Hinkley,D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.

Hollander,M. and Wolfe,D.A. (1999) Nonparametric Statistical Methods. *Nonparametric Statistical Methods*. Wiley, New York, pp. 75.

Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.

Li,C. and Wong,W.H. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Li,C. and Wong,W.H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**, 1–11.

Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**, 20–24.

McGall,G.H. and Fidanza,J.A. (2001) Photolithographic synthesis of high-density oligonucleotide arrays. *Methods Mol. Biol.*, **170**, 71–101.

Virtaneva,K., Wright,F.A., Tanner,S.M., Yuan,B., Lemon,W.J., Caligiuri,M.A., Bloomfield,C.D., de La Chapelle,A. and Krahe,R. (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl Acad. Sci. USA*, **98**, 1124–1129.

Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M. and Johnston,R. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, E41–41.