Quantitative noise analysis for gene expression microarray experiments

Y. Tu*†, G. Stolovitzky*, and U. Klein‡

*IBM T. J. Watson Research Center, Yorktown Heights, NY 10598; and ‡Institute for Cancer Genetics, Columbia University, New York, NY 10032

Edited by Robert H. Austin, Princeton University, Princeton, NJ, and approved August 29, 2002 (received for review March 20, 2002)

A major challenge in DNA microarray analysis is to effectively dissociate actual gene expression values from experimental noise. We report here a detailed noise analysis for oligonuleotide-based microarray experiments involving reverse transcription, generation of labeled cRNA (target) through *in vitro* transcription, and hybridization of the target to the probe immobilized on the substrate. By designing sets of replicate experiments that bifurcate at different steps of the assay, we are able to separate the noise caused by sample preparation and the hybridization processes. We quantitatively characterize the strength of these different sources of noise and their respective dependence on the gene expression level. We find that the sample preparation noise is small, implying that the amplification process during the sample preparation is relatively accurate. The hybridization noise is found to have very strong dependence on the expression level, with different characteristics for the low and high expression values. The hybridization noise characteristics at the high expression regime are mostly Poisson-like, whereas its characteristics for the small expression levels are more complex, probably due to cross-hybridization. A method to evaluate the significance of gene expression fold changes based on noise characteristics is proposed.

DNA microarray technology has a profound impact on biological research as it allows the monitoring of the transcription levels of tens of thousands of genes simultaneously. In the near future, it will be possible to profile the whole transcriptome of higher organisms, including *Homo sapiens*, with only a few DNA gene chips. This will allow us to obtain a global view of the genotypes corresponding to different cell phenotypes. Such capability will greatly accelerate and perhaps fundamentally change biomedical research and development in many areas, ranging from developing advanced diagnostics to unraveling complex biological pathways and networks, to eventually facilitating individual-based medicine (1, 2).

DNA microarray technology, however, is not without caveats. One of the major difficulties in deciphering high throughput gene expression experiments comes from the noisy nature of the data. In general, the changes in the measured transcript values between different experiments are caused by both biological variations (corresponding to real differences between different cell types and tissues) and experimental noise. To correctly interpret the gene expression microarray data, it is crucial to understand the sources of the experimental noise.

Previous works (3, 4) studied some aspects of the noise in DNA microarray experiments. In this article we report on detailed studies of the experimental noise occurring at subsequent steps in high-density oligonucleotide-based microarray (Affymetrix, Santa Clara, CA) assays. Elucidating the sources of noise may be of help for identifying the steps of the techniques that need to be modified to improve the signal-to-noise ratio. Our results show that it is the hybridization (including the subsequent readout) step, as opposed to the sample preparation step where most of the noise originates. Based on these results, we propose a data analysis method that takes into consideration the quantitative characterization of the noise, and thus provides a tool for evaluating the statistical significance of gene expression changes from different microarray experiments.

Materials and Methods

We study the measurement noise by replicate experiments in which gene expression levels of a cell line are measured multiple times. Two sources of experimental noise can be identified from the extracted mRNA to the final readout of the gene expression levels: the prehybridization target sample preparation steps and the hybridization and the subsequent readout processes (including staining and scanning). For simplicity, we refer to these two sources of noise as sample preparation noise and hybridization noise, respectively, throughout this article. To separate the noise sources caused by these two factors, we have carried out multiple replicate experiments, where at different stages of the experiment, the sample is divided equally into multiple aliquots, and the subsequent steps of the experiment are carried out independently. In this article, mRNA from cells of a human Burkitt's lymphoma cell line (Ramos) is used for the replicate experiments. Total RNA is extracted from the Ramos cells. The purified RNA sample subsequently is separated equally into several subgroups. Each subgroup independently goes through the target preparation steps, composed of the reverse transcription step and *in vitro* transcription (IVT) step. At the end of the target sample preparation, each of the subgroups is again split into several samples, each of which is independently hybridized to different Affymetrix U95A GeneChip arrays. The experimental design is shown schematically in Fig. 1. To have sound statistics and ensure the experimental statistics are independent of the starting mRNA, we have repeated the above replicate experiments with total RNA taken from two different cultures of the Ramos cells, as represented in Fig. 1, where experiments 1–4 and experiments 5–10 start from the different RNAs.

Sample preparation starting from 5 μg total RNA, hybridization, staining, and scanning were performed according to the Affymetrix protocol. Unless indicated otherwise, our analysis uses the (average difference-based) expression values obtained by Affymetrix MICROARRAY SUITE (MAS) version 5.0 with all of the default parameters and target intensity set to 250. The expression values from earlier versions of MAS (versions 4.0 and 3.1) were used only for comparison purposes.

Results and Discussion

From the experiments described above, we obtain a gene expression value matrix $\{E_{i,j}\}$, where $i = 1,2,\ldots,10$ represents all of the experiments shown in Fig. 1 and $j = 1,2,\ldots,J$ labels all of the individual genes being probed. For the U95A chip we used, $J \approx 12{,}600$. Due to the large variation in measured gene expression values, the analysis in this section is performed by using the logarithm of the expression level: $\theta_{i,j} = ln(E_{i,j})$.

For a pair of experiments i_1 and i_2, the overall differences in gene expression can be visualized by plotting $\theta_{i_1,j}$ versus $\theta_{i_2,j}$ for all genes on the microarray. In Fig. 2, two pairs of experiments (1 and 3 and 1 and 10) are shown. The deviation of the scattered points from the diagonal line represents the difference between

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: IVT, *in vitro* transcription; MAS, MICROARRAY SUITE.

†To whom correspondence should be addressed. E-mail: yuhai@us.ibm.com.

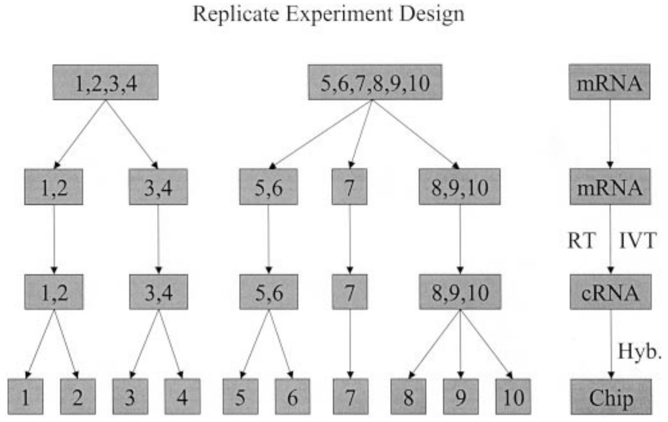

Fig. 1. Illustration of the replicate experiments setup. Two different mRNA samples are used, each being probed multiple times (replicates) with varying degrees of differences in measurement steps to separate the preparation error that occurred during the reverse transcription (RT) and IVT processes and the final hybridization (Hyb.) error.

the two measured transcriptomes. Although Fig. 2 *a* and *b* appear similar, the reasons for the deviation of the expression values from the diagonal line are different. Experiments 1 and 3 measure mRNA levels of exactly the same sample, so the observed expression differences between these experiments are caused by measurement error alone. On the other hand, samples 1 and 10 are from different cultures of the cell line, so the measured expression value differences as shown in Fig. 2*b* contain the combined effect of the genuine gene expression differences between the two cultures together with differences caused by measurement error. Therefore, to correctly assess the statistical relevance of the measured gene expression differences between two experiments, such as 1 and 10, it is crucial to characterize the fluctuation caused purely by experimental measurement, such as the noise shown in Fig. 2*a*.

Although experimental noise is known to be a feature of microarray experiments, only recently has it been studied systematically by replicate experiments (3, 4). In particular, for the oligonucleotide microarrays, Novak *et al.* (3) characterized the dispersion between two experiments by the SD of their corresponding gene expression levels. Using this measure of dispersion, they studied the different effects of experimental, physiological, and sampling variability, which provide important guidance for microarray experiment design. In this article, we focus on understanding how different experimental steps contribute to the total noise and what the possible mechanism for the noise could be. We also study the distribution of the noise in detail, which is used in devising a statistical method to determine differentially expressed genes.

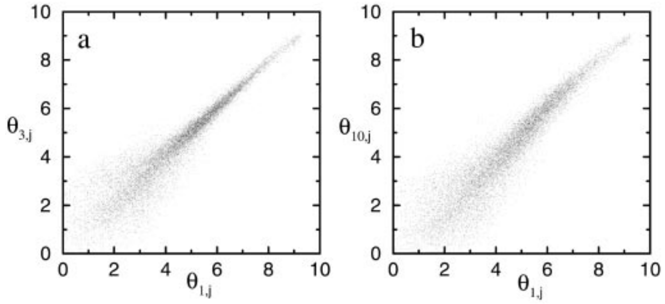

Fig. 2. The scatter plots of gene expression value pairs ($\theta_{i_1,j}, \theta_{i_2,j}$) for all genes $j \in [1,J]$ and for: (*a*) experiments pair (1 and 3), where the deviation from the diagonal axis is caused purely by experimental error; (*b*) experiment pair (1 and 10), where true differences exist between the two transcriptomes.

To separate the different noise sources, we group all of the replicate experiment pairs into two groups. Group G_1 consists of all of the pairs that differ only in the hybridization step:

$$G_1 = \{(1, 2), (3, 4), (5, 6), (8, 9), (9, 10), (8, 10)\}.$$

Group G_2 consists of all of the replicate experiment pairs that are carried out separately right after the extraction of the mRNA:

$$G_2 = \{(1, 3), (1, 4), (2, 3), (2, 4), (5, 7), (5, 8), (5, 9), (5, 10), (6, 7), (6, 8), (6, 9), (6, 10), (7, 8), (7, 9), (7, 10)\}.$$

Although gene expression differences between pairs of experiments in G_2 represent the full experimental noise, G_1 has been constructed to extract the noise caused by hybridization alone. For reference, we also group all of the nonreplicate experiment pairs into group $G_3 = \{(i, j), 1 \le i \le 4, 5 \le j \le 10\}$.

The Noise Distribution. It is evident from Fig. 2 that the noise depends strongly on the expression level. Therefore, an expression-dependent distribution function is needed to characterize the variability between replicates. Given two measured gene expression values, θ_1 and θ_2, for the same gene from two replicate experiments, the estimated value of the true expression level, $\bar{\theta}$, and the size of the measurement error, $\delta\theta$, can be defined as: $\bar{\theta} = (\theta_1 + \theta_2)/2$, $\delta\theta = (\theta_1 - \theta_2)/2$. $\bar{\theta}$ is discretized with a relatively small bin size of 0.25 throughout this article to maintain a good resolution while having sufficient data points per bin. The results are insensitive to the exact choice of the bin size. For a given $\bar{\theta}$, the average of $\delta\theta$ between two experiments should be zero: $\langle\delta\theta|\bar{\theta}\rangle = 0$. Any significantly nonzero value of $\langle\delta\theta|\bar{\theta}\rangle$ is caused by systematic experimental errors whose source is beyond the scope of our current study. This error typically appears as a departure from the diagonal of the scatter plots of Fig. 2. A hint of it can be seen at the higher values of Fig. 2*b*. Even though this was not a big problem for our data sets, we compensated for such error whenever it occurred by subtracting any nonzero $\langle\delta\theta|\bar{\theta}\rangle$ from $\delta\theta$ for each replicate experiment pairs for all of the subsequent analysis.

Within each group G_k ($k = 1, 2$), the distribution of $\delta\theta$ for a given $\bar{\theta}$ can be obtained from each pair of replicate experiments, these distributions are found to be highly consistent with each other (data not shown). To gain better statistics, we use the gene expression values from all of the pairs of replicate experiments in G_k to construct the noise distribution: $P_k(\delta\theta|\theta_0) = Prob_k(\delta\theta|\bar{\theta} = \theta_0)$. In Fig. 3*a*, the noise distribution functions for different values of θ_0 are shown. We use the second-order moment to quantify the strength of the noise and its dependence on the value of the expected expression level θ_0:

$$\sigma_k^2(\theta_0) = \int_{-\infty}^{\infty} \delta\theta^2 P_k(\delta\theta|\theta_0) d\delta\theta. \quad [1]$$

In Fig. 3*c*, we show the dependence of σ_2 *on* θ_0. For reference, we have calculated σ_3, the difference in gene expression between pairs of experiments in group G_3 in the same way as we calculated $\sigma_{1,2}$ and plotted it in Fig. 3*c* as well. It is interesting that σ_3 is consistently larger than σ_2 for $\theta_0 \ge 2$, indicating the existence of signal beyond noise even for the small differences between the same cell line from different cultures.

For a given θ_0, we can define the rescaled noise $\delta\theta' = \delta\theta/\sigma_k(\theta_0)$ and obtain the distribution function for $\delta\theta'$: $Q_k(\delta\theta'|\theta_0)$. We find that except for very small values of θ_0, the $Q_k(\delta\theta'|\theta_0)$ collapse onto a single curve $\Phi(\delta\theta')$ independent of θ_0

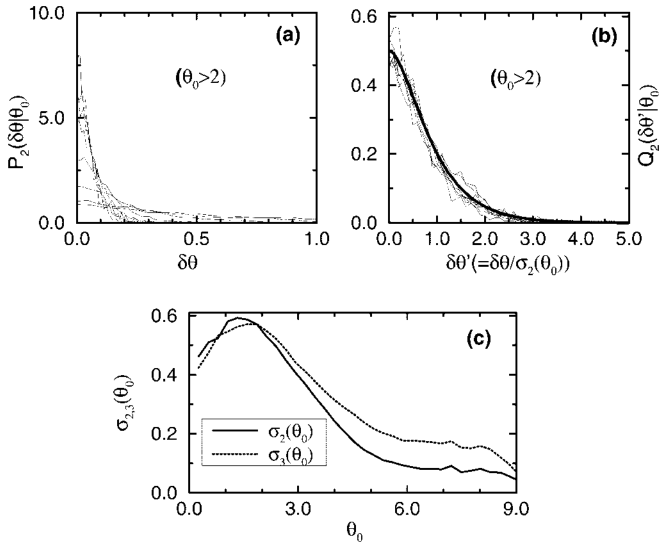

Fig. 3. The noise distribution functions at different values of mean expression values: θ_0 = 2,3,4,5,6,7,8,9 (*a*) before and (*b*) after rescaling by the SD $\sigma_2(\theta_0)$, which is shown in *c*. Only the positive region of $\delta\theta > 0$ is shown in *a* and *b* for symmetry reasons. The rescaled distribution functions collapse onto a single curve well fitted by $\Phi(\theta') = 1/2\exp(-x^2/0.5 + 0.6|x|)$, plotted as the thick line shown in *b*.

and k, as shown in Fig. 3*b* (for k = 2 only). Equivalently, this means the distribution for $\delta\theta$ can be well approximated by:

$$P_k(\delta\theta|\theta_0) \approx \frac{1}{\sigma_k(\theta_0)}\,\Phi(\delta\theta/\sigma_k(\theta_0)), \qquad \textbf{[2]}$$

for $\theta_0 \geq 2$, which includes more than 90% of the data. The rescaled distribution function is found to have an exponentially decaying tail in contrast with a Gaussian distribution. In fact, $\Phi(x)$ can be approximated very well by an empirical function $\Phi(x) \approx 1/2\exp(-x^2/0.5 + 0.6|x|)$ shown in Fig. 3*b* (thick solid line).

From Eq. **2**, we see that all of the expression-dependent information in the noise is given by the variance $\sigma_k^2(\theta_0)$ for $\theta_0 \geq 2$. In the following two subsections, we focus on analyzing the dependence of the noise strength $\sigma_k^2(\theta_0)$ on the expression value.

Sample Preparation Noise. To dissect the origins of noise, we divide the total measurement noise into two parts: the first is sample preparation noise $\delta\theta_{prep}$ caused by the prehybridization steps such as reverse transcription and IVT; the second is hybridization noise $\delta\theta_{hyb}$. For replicate pairs in group G_1 and G_2, the noise can be expressed, respectively, as: $\delta\theta_1 = \delta\theta_{hyb}$, $\delta\theta_2 = \delta\theta_{prep} + \delta\theta_{hyb}$. Assuming the two sources of noise are independent of each other, their variances can be obtained by: $\sigma_{hyb}^2 = \langle\delta\theta_{hyb}^2\rangle = \sigma_1^2$, $\sigma_{prep}^2 = \langle\delta\theta_{prep}^2\rangle = \sigma_2^2 - \sigma_1^2$, where $\sigma_{1,2}^2$ can be computed from Eq. **1**.

In Fig. 4, we show $\sigma_1^2(\theta_0)$ (dotted line) and $\sigma_2^2(\theta_0)$ (solid line) versus the expected value of the expression level θ_0. Although the difference between σ_2 and σ_1 is small in comparison with σ_2, $\sigma_1(\theta_0)$ is consistently smaller than $\sigma_2(\theta_0)$ for all of the values of $\theta_0 \geq 2$. This should be so because the difference between σ_2 and σ_1 accounts for the sample preparation noise: this difference, albeit small, is real.

We have plotted the dependence of σ_{prep}^2 versus θ_0 in Fig. 4 *Inset*. We find that the dependence of σ_{prep}^2 on the expression level θ_0 can be well approximated by:

$$\sigma_{prep}^2 \approx 1.9 \times 10^{-3} + 0.12e^{-\theta_0}. \qquad \textbf{[3]}$$

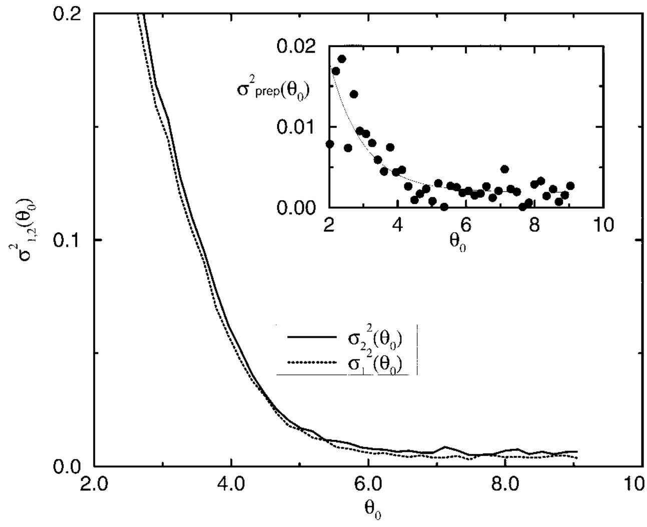

Fig. 4. The dependence of the noise strength $\sigma_{1,2}^2$, on the expected values of the gene expression for replicates in groups G_1 and G_2. (*Inset*) The variance of the sample preparation noise $\sigma_{prep}^2 = \sigma_2^2 - \sigma_1^2$ is shown. σ_{prep} has very weak dependence on the expression value for the large expression levels $\theta_0 \geq 4.0$ and can be fitted by $1.9 \times 10^{-3} + 0.12 \times e^{-\theta_0}$ for $\theta_0 \geq 2$.

The constant first term dominates the sample preparation noise for expression values $\theta_0 \geq 4$.

To understand the possible mechanisms for such noise behavior as shown in Eq. **3**, it is convenient to translate the above noise strength in θ (= $ln(E)$) to the noise strength in intensity E: $\sigma_E^2(E_0) \equiv \langle\delta E^2\rangle \approx E_0^2\langle\delta\theta^2\rangle$, where $E_0 = \exp(\theta_0)$ and $\delta E = E - E_0$. By using the numerical fit for σ_{prep}^2, the variance of the sample preparation noise δE_{prep}, $\sigma_{E,prep}^2$, can written as:

$$\sigma_{E,prep}^2(E_0) \equiv \langle\delta E_{prep}^2\rangle \approx 1.9 \times 10^{-3}E_0^2 + 0.12E_0. \qquad \textbf{[4]}$$

The two terms in the above expression represent two independent sources of noise, which we discuss in the following.

For the first term, δE_{prep} is proportional to the gene expression E_0 itself. To understand this term, it is important to realize that during sample preparation the mRNA is first reverse-transcribed into cDNA, and cRNA is subsequently generated from cDNA by IVT. The number of RNA molecules is amplified during the IVT, i.e., $N_{\mathrm{cRNA}} = A \times N_{\mathrm{mRNA}}$, where A is the amplification rate and N_{mRNA}, N_{cRNA} are the numbers of mRNA and cRNA molecules, respectively. A varies between one sample preparation process and another due to fluctuations in the reaction conditions, including fluctuation due to handling of the sample (human factors). The fluctuation of A between different sample preparation processes, denoted as δA, leads to a fluctuation in N_{cRNA} of the form $\delta A \times N_{\mathrm{mRNA}}$. Because N_{mRNA} is proportional to E_0, the first term in Eq. **4** can thus be explained by the fluctuation in A. Furthermore, σ_A, the SD of A, can be estimated: $\sigma_A \equiv \langle\delta A^2\rangle^{1/2} \approx (1.9 \times 10^{-3})^{1/2}\bar{A}$, where $\bar{A}$ is the mean amplification rate. Assuming a typical value of $\bar{A}$ around 100 (5), we have $\sigma_A \sim 4.4$.

For the second term in Eq. **4**, δE_{prep} is only proportional to the square root of E_0, which is thus indicative of a Poisson-like noise. Such Poisson-like noise in the sample preparation may arise naturally from the probabilistic nature of the amplification process (IVT).

The accuracy of the sample preparation process inevitably depends on human factors, whose influence is difficult to estimate. Our result here can be best viewed as an upper limit for the noise caused by the intrinsic chemical processes involved in the sample preparation.

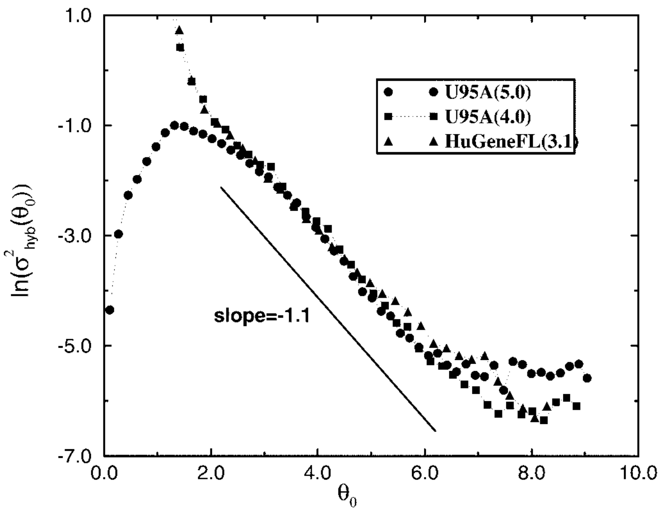

Fig. 5. Logarithm of the hybridization noise versus the expression level for our data obtained by different versions of Affymetrix MAS. ●, Results from MAS 5.0; ■, results from MAS 4.0. We have also calculated the noise strength from Lemon *et al.*'s replicate data (see ref. 6), which was performed with a different type of GeneChip array (HuGeneFL), with a different type of sample (human fibroblast cells) and by using MAS 3.1. The agreement between the three results indicates a certain universality of the noise characteristics in the region $3.0 \leq \theta_0 \leq 7.0$, where the variance of the noise decays exponentially (see text).

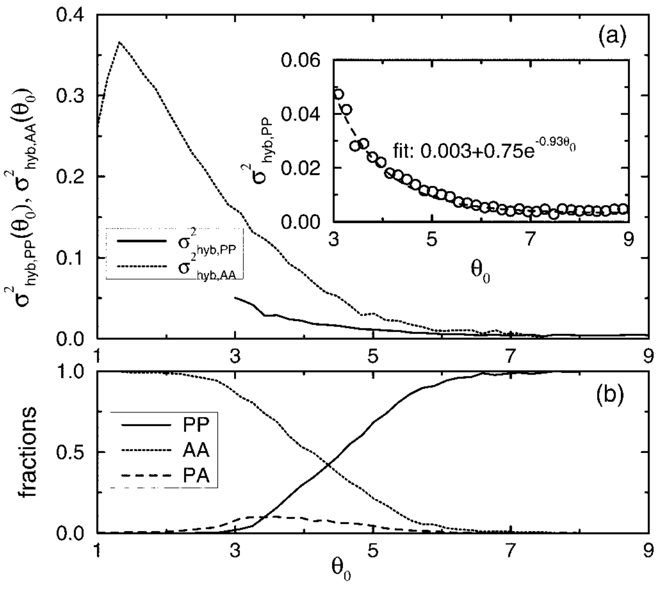

Fig. 6. (*a*) The overall hybridization noise (black line) is decomposed into two parts: the hybridization noise for genes that are labeled by MAS 5.0 as present ($\sigma_{hyb,PP}$, solid line) or absent ($\sigma_{hyb,AA}$, dotted line). (*Inset*) $\sigma^2_{hyb,PP}$ is fitted by $3.2 \times 10^{-3} + 0.75 \times \exp(-0.93\theta_0)$. For reference, the fractions of the PP, PA, and AA pairs in all of the replicate experiments at a given mean expression value θ_0 are plotted in *b*.

Hybridization Noise. Most of the total measurement error comes from the hybridization noise, which depends strongly on the expression level (see Fig. 4). For expression level $\theta_0 \geq 2$, the hybridization noise σ^2_{hyb} decreases rapidly with increasing expression level as shown in Fig. 5, where $ln(\sigma^2_{hyb})$ is plotted versus θ_0. Empirically, σ^2_{hyb} can be fitted by:

$$\sigma^2_{hyb}(\theta_0) \approx \beta e^{-\gamma\theta_0}, \qquad [5]$$

with $\beta = 4.6 \pm 0.2$ and $\gamma = 1.1 \pm 0.1$ for the region $3.2 \leq \theta_0 \leq 6.2$, before saturating to a constant (3.2×10^{-3}).

Also in Fig. 5, we have included the hybridization noise calculated by using expression values obtained from MAS version 4.0 [for 4.0 and earlier versions of MAS, θ_{ij} is defined as: $\theta_{ij} \equiv ln(\max(E_{ij}, E_c))$, where we choose a small $E_c = 0.1$ as a cutoff in avoiding negative expression values]. It is reassuring to see the results from the old and new versions of the software are consistent in the high-expression value region. The different behavior at low expression values reflects the major difference between versions 4.0 and 5.0 in dealing with negative differences between perfect match and mismatch probe pairs. This difference may be irrelevant because most of the genes with low expression values $\theta_0 \leq 3$ are considered to be absent from both versions of the software (see Fig. 6*b*).

To examine the robustness of the hybridization noise characteristics, we have also calculated the hybridization noise strength $(\sigma'_{hyb})^2$ for nine pairs of replicate experiments (6), which were performed with a different type of Affymetrix GeneChip array (HuGeneFL), with a different type of cell (human fibroblast cells) and in a different laboratory. The results are shown in Fig. 5 along with our data. It is remarkable that the exponentially decaying part of the hybridization noise seems universal regardless of the type of genechip and the sample being used. Notice also the agreement of the noise behavior in the full θ_0 range between our data generated with MAS 4.0 and the independently generated data of ref. 6 with MAS 3.1, which uses the same analysis algorithm as MAS 4.0. This observation indicates that the noise as characterized in the present analysis seems to show a degree of universality; more work is needed in confirming this behavior.

Noise in the hybridization signal can come from fluctuations in both the target molecule binding and cross-hybridization (nonspecific binding), which may have different behaviors. To roughly separate between specific and nonspecific hybridization, we use the Affymetrix "present" (PP) and "absent" (AA) calls. In particular, we calculate the noise strengths $\sigma^2_{hyb,PP}$ and $\sigma^2_{hyb,AA}$ for only those genes whose calls are both present and both absent for the replicate experiment pair from G_1. The results are shown in Fig. 6*a*. For reference, we also plot the fractions of the PP, AA, and PA pairs for a given mean expression value θ_0 in Fig. 6*b*. From Fig. 6*a*, it becomes clear that the noise characteristics are different for $\sigma^2_{hyb,PP}$ and $\sigma^2_{hyb,AA}$. This is most evident in the region $3 \leq \theta_0 \leq 6$, where PP pairs and AA pairs are both populated (see Fig. 6*b*). Their different behavior suggests that $\sigma_{hyb,PP}$ and $\sigma_{hyb,AA}$ have different origins.

For $\sigma^2_{hyb,PP}$, we can fit the PP hybridization noise strength with

$$\sigma^2_{hyb,PP}(\theta_0) \approx \alpha_{PP} + \beta_{PP} e^{-\gamma_{PP}\theta_0}, \qquad [6]$$

for $\theta_0 \geq 3.0$ and with $\alpha_{PP} = 3.2 \times 10^{-3} \pm 2.0 \times 10^{-4}$, $\beta_{PP} = 0.75 \pm 0.1$, and $\gamma_{PP} = 0.93 \pm 0.04$ as shown in Fig. 6*a Inset*. The origins of the two noise terms in Eq. **6** may be understood as follows. In general, for a gene with a present call, the final expression readout E should be proportional to N_{cRNA}, the number of cRNA molecules of the gene: $E = qN_{\text{cRNA}}$. However, the proportional factor q, which depends on the hybridization and the subsequent readout processes, can vary between different gene chips [for example, due to differences in purity of the probes on different gene chips (7)]. Such fluctuation in q between different experiments can give rise to the (constant) first term in Eq. **6**. The second term in Eq. **6**, with $\gamma_{PP} \sim 1$, indicates a Poisson-like noise (see earlier discussion of the sample preparation noise). Such Poisson-like noise may arise naturally from the probabilistic nature of the hybridization and the subsequent readout processes.

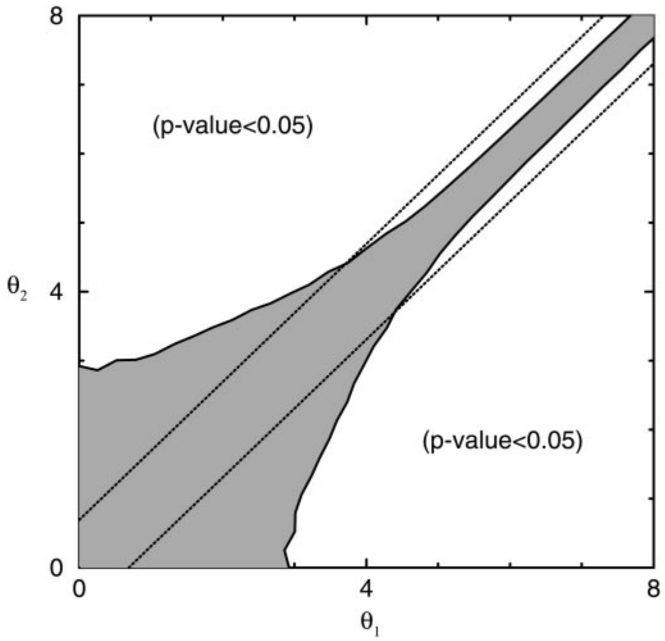

Fig. 7. The contour line of p value equal to 0.05. Any pair of expression values (θ_1,θ_2) outside the shaded area represents differently expressed genes beyond experimental noise with a p value of 0.05 or smaller. The two dotted lines represent 2-fold expression changes.

For $\sigma_{hyb,AA}$, it cannot be fitted with any simple form that would allow speculations about its origin. The best fit with an exponential function in the region $2 \le \theta_0 \le 5.0$ is (not shown in Fig. 6*a*): $\sigma^2_{hyb,AA} \sim \beta_{AA}e^{-\gamma_{AA}\theta_0}$ with $\beta_{AA} = 1.3 \pm 0.1$ and $\gamma_{AA} = 0.72 \pm 0.1$. Indeed, it is not clear what the expression intensity means when the gene is deemed absent by the Affymetrix call. Most likely, the intensity value and its fluctuation, if meaningful at all, are affected by cross-hybridization. The final intensity values and their fluctuations depend very much on the way one deals with negative differences between perfect match and mismatch probe pairs, which occur most frequently in the absent genes. This is consistent with our finding (data not shown) that $\sigma_{hyb,AA}$ changes significantly when we use the intensity values from MAS 4.0 instead of MAS 5.0, whereas the change in $\sigma_{hyb,PP}$ between the two versions is minimal.

USE-Fold: A Method for Uniform Significance of Expression Fold Change. The results presented in the previous sections can be used to design a method for determining the statistical relevance of gene expression changes. The idea is simply that the fold change experienced by a gene under different biological conditions has to be larger than the fold change expected from the noise. We shall use the full noise distribution function discussed previously to evaluate the significance of the difference between a pair of gene expressions (θ_1,θ_2) for the same gene but different experiments. By using the fluctuation between replicate experiment pairs in G_2 as the null hypothesis, a gene expression-dependent p value can be defined as:

$$p(\theta_1,\ \theta_2|\theta_0) = \int_{|\delta\theta| \ge \Delta\theta_0} P_2(\delta\theta|\theta_0)d\delta\theta, \tag{7}$$

where $\Delta\theta_0 = |\theta_1 - \theta_2|/2$, $\theta_0 = (\theta_1 + \theta_2)/2$.

For $\theta_0 \ge 2$, we can use Eq. **2**, and the p value can be expressed simply as a function of the signal-to-noise ratio $R \equiv \Delta\theta_0/\sigma_2(\theta_0)$: $p(\theta_1,\theta_2|\theta_0) = 2\int_R^\infty \Phi(x)dx$. In Fig. 7, the contour lines for $p(\theta_1,\theta_2|\theta_0) = 0.05$ are shown together with two lines corresponding to a uniform 2-fold expression value change [$|\theta_1 - \theta_2| = ln(2)$]. This clearly shows that given a fixed confidence level (p value = 0.05), a requirement of a uniform 2-fold expression change is too stringent for the high expression level, while being inadequate for the low expression level ($\theta_0 \le 4$). In fact, given the strong expression level dependence of the noise, no significance criterion based solely on the expression fold change is appropriate. Instead, to guarantee a fixed level of statistical relevance p_0, one can enforce a uniform (i.e., expression level independent) lower bound on the signal-to-noise ratio $R \ge R_0(p_0)$.

The above discussion suggests the following method of selecting differently expressed genes with user-defined statistical significance:

- Evaluate the noise level from replicate experiments such as those in group G_2. Ideally, each laboratory should carry out its replicate experiments to determine the noise level. If this is not possible, the results of this article may be used with some degree of confidence, as we have shown consistency between two sets of replicate data produced in different laboratories (our data and that of ref. 6, see Fig. 5).
- After obtaining $\sigma_2(\theta_0)$ from the previous step, pick a significance level p_0, and compute the corresponding threshold for the signal-to-noise ratio R_0 such that $p_0 = 2\int_{R_0}^\infty \Phi(x)dx$, where $\Phi(x)$ is the noise distribution function. Using the empirical form of $\phi(x) = 1/2 \exp(-x^2/0.5 + 0.6|x|)$ found in this article, for significance level $p_0 = 0.05$, we find the corresponding $R_0 \approx 2.1$.
- Given two expression values E_1 and E_2, corresponding to the fluorescence intensity of the same gene from different gene chips, compute $\theta_1 = ln(E_1)$ and $\theta_2 = ln(E_2)$, and define $\theta_0 = (\theta_1 + \theta_2)/2$. The fold change $\phi = \max(E_1/E_2,\ E_2/E_1)$ is statistically significant with a p value less or equal than p_0 if the signal-to-noise ratio $ln(\phi)/(2\sigma_2(\theta_0)) \ge R_0$.

To demonstrate the utility of this method, we have applied it to discover differentially expressed genes between two developmentally distinct types of B lymphocytes, a centroblast (CB) and a naive (N) B cell (see Tables 1 and 2, Fig. 8, and additional *Text*, which are published as supporting information on the PNAS web site, www.pnas.org, for details). A total of 1,490 genes were found to change more than 2-fold in their expression values and have at least one present call in either of the two experiments. However, more than 10% of these genes do not pass the USE-Fold noise test with $p_0 = 0.05$. For example, one gene (GenBank accession no. AA143021) has present calls in both experiments with expression values $E_1 = 48.3$ and $E_2 = 21.7$ for CB and N, respectively. Even though the fold change $\phi = E_1/E_2 = 2.23$ is greater than 2, at their mean (logarithmic) expression level of $\theta_0 = (ln(E_1) + ln(E_2))/2 = 3.48$, the noise level is also large, $\sigma_2(\theta_0) = 0.32$ (see Fig. 4) and the signal-to-noise ratio $ln(\phi)/(2\sigma_2(\theta_0)) = 1.25$ is smaller than $R_0 = 2.1$. Therefore, this gene cannot be considered to be differentially expressed with high confidence by just these two experiments. To test whether or not such gene is differentially expressed between the two types of B cell, more experiments need to be done to average out the effect of the random experimental noise (8). This is necessary particularly for genes with low expression, because the relative noise is much larger at low expression levels.

All of the data used in this article and free software implementing the USE-Fold method can be found at our web site (www.research.ibm.com/FunGen/index.html).

Conclusions

In this article, we have systematically studied the experimental noise characteristics of Affymetrix GeneChip microarray experiments. By designing replicate experiments that differ from each other at different stages of the experiments, we are able to decompose the total experimental noise into two parts: the sample preparation (prehybridization) noise and the hybridization (including the subsequent readout processes) noise. We have characterized these two sources of noise quantitatively, and in particular, their dependence on the gene expression level itself. For the sample preparation noise, we find that it is dominated by an expression-independent constant and is in general much smaller than the hybridization noise. For the hybridization noise, except for a small constant component, the noise strength is found to depend strongly on the expression level. Specifically, for the genes labeled by the Affymetrix call as present, the dependence of the hybridization noise strength on the expression indicates a Poisson-like noise, in accordance with the probabilistic nature of the hybridization process; for the absent genes, however, the hybridization noise characteristics does not have a simple explanation, because the noise and even the gene expression readout itself are affected by cross-hybridization.

Overall, the importance of this work is 2-fold. First, our study provides a quantitative measure of the experimental noise, which served us as a base for designing a simple method for determining statistical meaningful biological information from gene expression microarray data. Second, our study provides insight into the sources of the noise by decomposing the noise according to the individual steps of the genechip experiment. The insights gained from this study may help to further reduce the errors arising in DNA microarray experiments.

We thank V. Miljkovic for technical assistance and R. Dalla-Favera for his support. We also thank G. Grinstein for discussion and reading of the manuscript. U.K. was a recipient of a fellowship granted by the Human Frontiers Science Program.

1. Lockhart, D. J. & Winzeler, E. A. (2000) *Nature* **405,** 827–836.
2. Brown, P. O. & Botstein, D. (1999) *Nat. Genet.* **21,** Suppl., 33–37.
3. Novak, J. P., Sladek, R. & Hudson, T. J. (2002) *Genomics* **79,** 104–113.
4. Lee, M.-L. T., Kuo, F. C., Whitemore, G. A. & Sklar, I. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 9834–9839.
5. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Hortan, H. & Brown, E. L. (1996) *Nat. Biotechnol.* **14,** 1675–1680.
6. Lemon, W. J., Palatini, J. J. T., Krahe, R. & Wright, F. A. (2001) preprint, http://thinker.med.ohio-state.edu/projects/fbss/index.html.
7. Forman, J. E., Walton, I. D., Stern, D., Rava, R. P. & Trulson, M. O. (1997) *Am. Chem. Soc. Symp. Ser.* **682,** 2208–2228.
8. Pan, W., Lin, J. & Le, C. T. (2002) *Genome Biol.* **3,** 1–10.