Optimal gene expression analysis by microarrays

Lance D. Miller,[1,2,8] Philip M. Long,[1,3] Limsoon Wong,[4] Sayan Mukherjee,[5,6] Lisa M. McShane,[7] and Edison T. Liu[1,8]

[1]Genome Institute of Singapore, Singapore 117528
[2]Microarray and Expression Genomics Laboratory
[3]Information and Mathematical Sciences
[4]Institute for Infocomm Research, Singapore 119613
[5]MIT Whitehead Institute, Cambridge, Massachusetts 02138
[6]Center for Biological and Computational Learning
[7]Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892
[8]Correspondence: gislm@nus.edu.sg (L.D.M.), gisliue@nus.edu.sg (E.T.L.)

DNA microarrays make possible the rapid and comprehensive assessment of the transcriptional activity of a cell, and as such have proven valuable in assessing the molecular contributors to biological processes and in the classification of human cancers. The major challenge in using this technology is the analysis of its massive data output, which requires computational means for interpretation and a heightened need for quality data. The optimal analysis requires an accounting and control of the many sources of variance within the system, an understanding of the limitations of the statistical approaches, and the ability to make sense of the results through intelligent database interrogation.

Expression array technology

Expression genomics is an approach that examines gene expression in a comprehensive and massively parallel fashion. The core technology in expression genomics is *microarrays*, whereby thousands of DNA *probes* are immobilized on a solid surface and hybridized against fluorophore-labeled cDNA or cRNA *targets* from template RNA sources. The two major platforms for microarrays are *spotted* arrays, where the probes are mechanically deposited onto modified glass slides by contact or ink jet printing, and *in situ* arrays, where oligo probes are synthesized *in silico* (e.g., via photolithographic synthesis as in Affymetrix GeneChip arrays [Affymetrix, Santa Clara, CA]). For a comparative review of the two platforms, we refer the reader to Harrington et al. (2000).

The power of expression genomics is that the simultaneous assessment of gene expression in such a massively parallel manner and across numerous cellular conditions uncovers higher-order organization in gene transcriptional behavior. Often, the more genes and biological conditions studied, the more obvious this underlying organization becomes. Thus, complexity is essential, and the analysis of this complex array data becomes the most critical issue in expression genomics. The power of microarray data is not in viewing the technology as a collection of individual "Northern" blots, but in generating a composite image of the expression profile of a cell.

In cancer research, the most intriguing use of expression arrays has been in the molecular classification of tumors. Many studies have now shown the ability of this approach to identify tumor subclasses that standard clinical indicators or histopathology could not (Alizadeh et al., 2000; Sorlie et al., 2001; Bhattacharjee et al., 2001). Taken together, these studies have highlighted several key points.

First, cell lineage has a primary role in determining the expression profile. Individual lineages may originate from a specific cell type such as germinal center B cells, "activated" B cells (Alizadeh et al., 2000), or basal or luminal breast epithelial cells (Sorlie et al., 2001). The importance of cellular lineage is also seen in the greater similarity of array profiles between a primary and metastatic tumor from the same patient rather than between profiles of primary cancers from different patients (Perou et al., 2000).

Second, array profiles may define distinct prognostic subgroups, and frequently these subgroups are associated with cell lineages. For example, diffuse large B cell lymphomas with an activated B cell-like profile have a worse prognosis than those with a germinal center B cell-like profile (Alizadeh et al., 2000). van't Veer et al. (2002) identified 70 genes from over 20,000 genetic elements that can collectively separate node negative breast cancer patients into distinctly good prognostic and poor prognostic groups. Singh et al. (2002) found five genes whose expression characteristics alone could discern prostate cancer patients with excellent versus poor survival. Based on the known function of these genes, no obvious biological explanation linked most of these specific transcripts to tumor behavior and prognosis. Thus, though the composite expression of many genes define profiles associated with clinical outcomes, it is not clear whether any are causal or merely surrogate markers.

Third, there is primacy of pathways over the effects of individual genes. Transgenic animal models suggest that genes involved in the same pathway generate tumors with similar expression profiles and are distinct from profiles of tumors arising from other transgene pathways (Desai et al., 2002). This, again, means that the information from the composite expression of a group of genes is likely to be more important than the behavior of any one particular genetic element.

Fourth, though pathways ultimately define the profiles, key oncogenetic events inducing downstream pathway changes are associated with distinguishable expression profiles. The approaches used to analyze arrays have also placed specific pathways in a hierarchy of importance for determining a tumor profile. For example, estrogen receptor (ER) status appears to define breast cancers into two major classes (Sorlie et al., 2001; van't Veer, 2002). Subclasses can be discerned only within the major ER groups by other parameters such as p53 status and the presence of HER-2 overexpression (Sorlie et al., 2001). Moreover, in the analysis of pediatric acute lymphoblastic

leukemias, specific chromosomal abnormalities such as *E2A-PBX*1, *BCR-ABL*, *TEL-AML*1, *MLL* rearrangement, and hyperdiploidy are associated with signature profiles (Yeoh et al., 2002).

The ability to identify tumor classes with clinical importance through an association with the expression behavior of a population of genes has made expression profiling an attractive approach in the study of malignant transformation and for the discovery of tumor markers. Like any new technology, the confusing and sometimes contradictory data that often emerges is due to an incomplete understanding of the limits of the technology and the optimal ways to analyze its unique data sets. Herein, we focus on the capabilities and limitations of the expression array technologies and suggest "best practices" for optimal study design and analysis. Formulation of the best technical and operational approaches is necessary to render this platform useful in clinical practice since, in its current form, expression arrays cannot be considered sufficiently reliable upon which to make clinical decisions.

Study design and scientific objectives

The focus on the array technology hardware has often overshadowed the importance of quality experimental design in array experiments. Declarations that functional genomics is not hypothesis testing or that arrays merely cast a wide net for "interesting" genes are misguided. Occasionally, arrays are used to define pathways to uncover the function of a gene of unknown activity and of unknown significance. This approach, while always generating data, usually leads to inconclusive and often irrelevant results. The complexity and diversity of the data output from array experiments are such that any narrow conclusion can be divined from random noise. Unless the conclusions are analytically or biochemically validated, or the experimental conditions well defined, the veracity of the results remain suspect. For experiments in expression genomics to be effective, the design must be anchored to defined phenotypic end points (for example, growth versus no growth or long-term survival versus short-term survival) or to a defined hypothesis such as "large cell lymphomas are likely to be subclassified into prognostically important molecular classes."

The design of an array experiment often begins with the selection of the most relevant reference RNA. This is especially important in the spotted array format. Often, in time course experiments, all time points are compared to an untreated reference at time zero or one that parallels each time point. This is much more preferable to time series studies where each time point is compared to the subsequent time point. For clinical samples such as tumors, some have suggested using the matched patient's normal tissue as reference. This approach can be problematic because, often, sufficient quantities of a tumor's normal counterpart are not available or identifiable, such as the normal tissue for a head and neck cancer. Moreover, the normal tissues may actually be commonly abnormal, as in the case of hepatocellular carcinoma where the adjacent liver tissues are usually cirrhotic. At the Genome Institute of Singapore, we consistently use a specific universal reference RNA comprising a calibrated mix of a number of defined cell lines for experiments in human cells (Universal Human Reference RNA, Stratagene, La Jolla, CA). In this manner, most spots give a signal in the reference channel, thus avoiding having a small, near zero denominator in calculating ratios. More importantly, all experiments over time and across operators can be compared since the reference RNA is the same. This concept of a "perpetual array platform" allows for the collective experimental history of a laboratory to be used to uncover new hypotheses. Other microarray experimental designs involving multifactorial comparisons have also been proposed and may be useful in unique experimental circumstances; however, for the majority of array experiments, the simplest study design provides the most clearly interpretable results (Yang and Speed, 2002; Dobbin and Simon, 2002).

The main objectives of most microarray studies can be broadly classified into one of the following categories: class comparison, class discovery, or class prediction. For the class comparison aim, the interest is in establishing whether expression profiles differ between classes, and if they do, what genes are differentially expressed between the classes. Examples include establishing that expression profiles differ between two histologic subtypes of cancer and identifying genes whose expression levels are altered by exposure of cells in vitro to an experimental drug. For class discovery, the goal is to elucidate subclusters or structure among specimens or among genes. Examples include discovery of previously unrecognized subtypes of lymphoma and identification of coregulated genes. The goal of class prediction is to predict a phenotype using information from a gene expression profile. Examples include predicting which patients are likely to experience severe drug toxicity versus who will have none and predicting which breast cancer patients will relapse within two years of diagnosis versus who will remain disease free. For excellent discussions of sample size and other design considerations relevant to the various types of study aims, the reader is referred to Simon et al. (2002) and Yang and Speed (2002).

Higher-order analysis: Analytic approaches appropriate for the scientific objectives

When class discovery is the goal, *unsupervised* analysis strategies such as clustering methods can be used. For class comparison or class prediction, *supervised* analysis methods that use known class information (such as tumor versus normal designations) are most effective.

Prior to conducting any of these analyses, data from low signal intensity or inconsistent/nonreciprocating spots and from genes exhibiting little variation across the collection of arrays should be excluded. The rationale is that low-intensity spots are unstable, and genes that exhibit little variation across arrays do not contribute useful information for distinguishing among specimens. Eliminating genes showing little to no variation is an important "filtering" step in the data analysis that is sometimes overlooked. Typically, genes exhibiting high calculated variance across all samples (above a specified threshold) are included in the final analysis. By removing such "noise" from the system, the results are cleaner but without biasing the outcome.

Unsupervised clustering methods seek structure inherent in the data and assume no a priori classifications of the genes and samples (see Figures 1A and 1B). Their goal is to separate specimens or genes into subgroups of related expression patterns in an unbiased manner. One of the most widely used clustering approaches for microarray data is hierarchical agglomerative clustering (Eisen et al., 1998). In this procedure, each individual (specimen or gene) starts as its own cluster, and then pairs of clusters that are most similar in some sense are iteratively merged to form new clusters. There are nonhierarchical clustering alternatives as well. Many of these require that the number of clusters be predefined by the user. The clas-

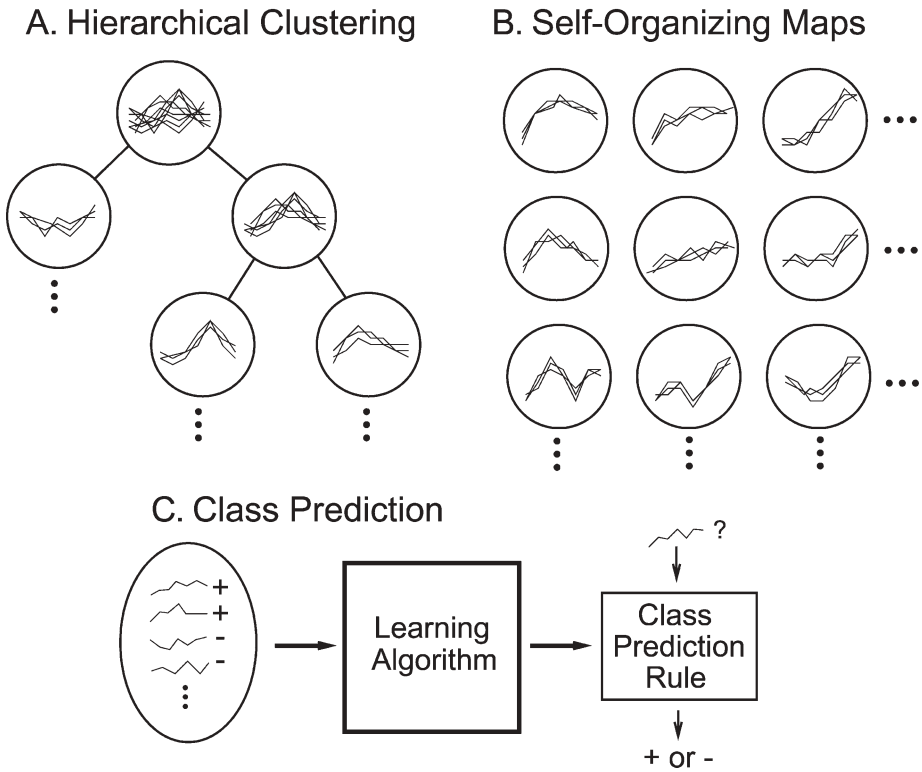

Figure 1. Three approaches to the analysis of high-dimensional expression data

A: Hierarchical clustering identifies cases in which groups of genes (or samples) with similar expression profiles contain smaller groups whose members are even more similar to one another. **B:** Self-organizing maps promote the choice of clusters that bear a special relationship to one another; shown is how the clusters can be laid out on a two-dimensional grid, with similar clusters near to each other. **C:** In class prediction, an algorithm takes as input a collection of expression profiles paired with preassigned class designations and outputs a rule for predicting class designations. The leave-one-out approach generates a class prediction rule using all but one sample, then subjects that "held-out" sample to the rule. By holding out each of the samples in turn, a measure of accuracy is obtained.

sical K-means algorithm is a partitional clustering that is often used (Tibshirani et al., 1999), and self-organizing maps (SOMs) have also been profitably applied (Tamayo et al., 1999). Loosely speaking, self-organizing maps promote the choice of clusters whose centers can be rearranged on a two-dimensional grid without distorting the distances between them too much. This allows for a graphical presentation of the results that can yield useful insights. It is important to keep in mind that clustering algorithms will always find clusters, even in data that is random noise. Here, again, a clear understanding of the study objective and careful study design are important for optimal analysis. McShane and colleagues (McShane et al., 2002) discuss this issue and suggest some methods for assessing the reproducibility of clustering patterns found in analyses of microarray data.

For class comparison (a method for supervised analysis), one usually identifies genes that are differentially expressed between known classes of specimens using univariate analyses on each gene, such as Wilcoxon tests. In doing so, it is important to take into account the problem of multiple testing in order to avoid generating many false leads (sometimes referred to as "false discoveries"). For example, if 10,000 genes were tested, we would expect 500 genes to be falsely declared as significantly different between the classes at $p < 0.05$, even if there were no real differences. Various multiple comparisons adjustment procedures have been applied to microarray data for purposes of controlling the number or proportion of false discoveries (see Tusher et al., 2001; Efron et al., 2001). One simple method to control for false discoveries is to conduct each univariate test using a small significance level. If each univariate test is conducted at a significance level α and there are N tests, the expected number of false discoveries will be $\alpha \times N$ or less. Thus, if 10,000 tests are conducted, each at a specified probability level of 0.001, on average the number of false discoveries will be 10 or less. If the analysis of these 10,000 genes reveals 100 that achieve this statistical level of significance, then one can be confident that true classification markers are within the 100 selected genes.

Sometimes it is desired to develop a multivariate predictor of tumor classification (Figure 1C). For example, there may be a number of gene markers whose collective behavior may predict with substantial accuracy whether a tumor will respond to a particular chemotherapeutic agent. Here, the tissues are already divided into classes based on a putative "gold standard" assessment of response, and the question is how to best mathematically combine the gene expression measurements into a single function that can delineate those classes. Several analytical strategies have been successfully used in array studies, such as Fischer linear discriminant analysis (Dudoit et al., 2002), nearest centroid (Tibshirani et al., 2002), and the support vector machine (SVM) (Furey et al., 2000; Mukherjee et al., 1999). These approaches are especially important in the potential clinical application of microarrays where the power of the technology is in its ability to use somewhat imprecise composite patterns of expression rather than exact thresholds of individual markers.

Leave-one-out crossvalidation is a popular method for estimating the accuracy of the output class prediction rule: each of the samples is individually removed from the data set, the remaining data is used to train a class prediction rule, and the resulting rule is applied to predict the class of the held-out sample. The accuracy of the class prediction rule is assessed by the number of true or false assignments in the samples that have been held out. For a discussion of the pros and cons of this and other methods, including *k*-fold crossvalidation and bootstrapping, see Efron and Tibshirani (1997) and Dudoit et al. (2002).

A common experimental question is whether statistically significant differences between the expression profiles of samples from different classes are seen. This can be assessed by estimating whether the accuracy achieved by a class prediction rule is better than would be obtained by chance. In assessing single markers, a standard t test would suffice. For the analysis of a multitude of markers as are generated from microarray experiments, this is assessed by randomly permuting the class

Table 1. Sources of technical variance in microarray data and proposed solutions

Domain	Problem	Solution
Printing/Processing-dependent	Missing spots	Guard against particulates that might clog pins, i.e., keep dust levels low, sonicate pins; facilitate pin wicking with 0.5× SSC in wash buffer and sufficient probe volumes; carefully calibrate for consistent pin-slide contact and faithful sample loading
	Merging spots	Reduce acceleration at which pins exit probe samples; reduce velocity of pin-surface contact; use slide "blot pads" and multiple tapping to remove excess probe; clean pins in EtOH after use
	Probe carry-over contamination	Experiment with pin cleaning procedure to ensure adequate cleaning and drying—check on regular basis; avoid reduction of vacuum pressure by keeping dry station and vacuum pump free of particulates and salt build-up
	Variability in probe volume and concentration	Ensure adaquate pin cleaning and drying; tightly seal print plates when not in use to prevent evaporation of probe buffer; routinely dry down probes and resuspend in appropriate volume
	Variability in slide surface-DNA affinity	For both commercially acquired and in-house prepared slides, routinely check the capacity of processed arrays to retain probe via flourescent staining and quantitation of DNA spots
	Comet tailing (i.e., probe smearing) and background	During array postprocessing, i.e., where the unbound glass is blocked to prevent background, dunk slides vigorously in blocking solution for at least 20 s when first wetted to disperse "loose" DNA; perform this step in low humidity as spots should also be completely dry prior to entering blocking solution
Hybing/Scanning-dependent	Damage to array surface	Take care to prevent contact of coverslip, or other solid matter, with the array surface; minimize bubble formation when adding target to the array
	Dye bias	Minimize exposure of fluorophores to bright light; for target labeling, try amino-allyl-coupling protocol instead of direct incorporation; use dye swapping strategy to detect such events
	Signal gradients and high background	Be consistent with hybridization volumes, conditions, and technique; avoid evaporation of target by humidifying hyb chamber with an appropriate volume of water or SSC and by rapidly transferring slide from warm hyb chamber (after hybridization) to first wash buffer (otherwise instantaneous evaporation can occur)
	Signal saturation	This occurs when the scanner PMTs or laser power is set too high to detect signal in the linear range; this results in an underestimation of expression levels and ratio compression of differentially expressed genes. This can be avoided on some scanners by pixel coloring that indicates saturation has occurred
Sample-dependent	Tissue heterogeneity	When resecting tissues, prepare all samples in a consistent way such that minimal (but equal) amounts of associated nontarget tissues are cocollected
	RNA degradation	Take precautions to minimize RNAse activity during RNA purification, run RNA on gel to check (by size) for degradation; avoid multiple freeze-thaw cycles
	Limited RNA quantity	Purify total RNA instead of mRNA (for increased polyA RNA yield); use validated linear RNA (or cDNA) amplification
Probe-dependent	Probe-target cross hybridization	Avoid cDNA probe sequences that span conserved coding regions; select oligos that have high hamming distance or are otherwise scored for high specificity
	Poor hybridization characteristics	Select cDNA or oligo sequences with similar annealing temperatures and with minimal predicted secondary structure

designations and measuring whether this degrades the predictive accuracy. The confidence in the putative association between expression profiles and their class designations can be estimated by the frequency of false positives (see Radmacher et al., 2002).

Self-fulfilling oracles

An insidious problem that has been seen in array analysis is that of circular reasoning. The most obvious example is when a t test is used to identify genes that distinguish between two defined classes (such as tumor versus normal). This gene list is then used in hierarchical clustering of the tissues. As expected, the cluster dendrogram recapitulates the distinction between normal and tumor tissues. It is then claimed that the clustering result validates the t test gene list. This assertion of validation is not justified because one must validate on an independent test set.

Biases like this have been discussed previously (Furey et al., 2000; Ambroise and McLachlan, 2002) and are illustrated in the following simulation experiment. We constructed an artificial data set with 100 samples, each with 100,000 random expression values and randomly assigned class designations. We then selected the 20 genes with the smallest p values determined by the Wilcoxon rank sum test. Next, we evaluated the accuracy of using these 20 genes in class prediction by leave-one-out crossvalidation using only the 20 selected genes. The resultant estimated accuracy was 88%, despite the fact that the true accuracy must be only 50% (because the data are derived from random assignments). The proper approach would be to reselect the top 20 genes each time a sample is held out in the leave-one-out cycle. Otherwise, information about the held-out sample is inappropriately "leaked" to the process that generates the class prediction rule. Alternatively, the 20 most significant genes could have been derived from one data set and their association with class membership validated on a second set of

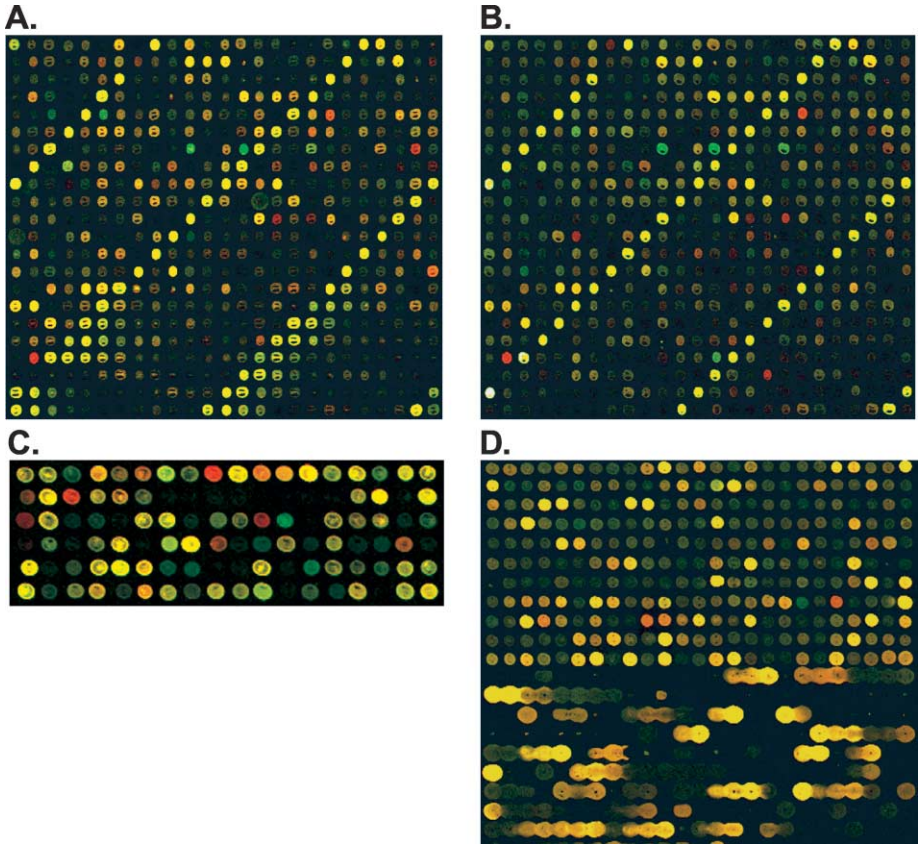

Figure 2. Array manufacturing artifacts that contribute to error

Arrays were printed from left to right, top to bottom. **A:** Example of probe carryover contamination. Note the three diagonal lines (top left to bottom right) of yellow spots (each is identical GAPDH probe). In the lower third of the image, probe carryover resulting from reduced vacuum pressure during tip cleaning is seen in spots adjacent to those comprising the diagonals by virtue of a diminishing yellow characteristic not observed in **B**. **B:** Same array location as in **A**, but from a print batch not affected by carryover contamination. **C and D:** Examples of missing spots (**C**, second row, eight consecutive missing spots) and merging spots (**D**, lower half) owing to print pin clogging.

samples. It is very important that the test data used to evaluate the system plays no role in training the classifier. The paper by Ambroise and McLachlan (2002) contains additional examples, some quite dramatic, using natural data.

Validating findings

Once microarray results have been analyzed, an orderly validation of the molecular and biological findings is necessary. Such a validation scheme should include some or all of the following elements. (1) When a particular gene is crucial to a hypothesis, the behavior of its transcript should be validated by an alternative RNA quantitation method. (e.g., Northern Blot, quantitative RT-PCR, etc.). (2) The pathway identified should be confirmed using biochemical means. (3) The importance of that pathway to the cellular phenotype should be confirmed by perturbing the pathway using chemical or molecular modifiers if available. (4) In the case of classification of tumors, validation may be performed on a new set of tumors or in assessing common features with other studies.

One way to focus on the critical genes involved in a specific pathway is to use clean genetic models such as knockout animals or cell lines. In this approach, a pathway can be inferred if the elements are operative in the wild-type cells, absent or reduced in the knockout condition, and recovered in the knockin or reconstituted cells (Aprelikova et al., 2001). We have found that the analysis of several knockout lines or of a knockout with a reconstituted cell line or animal is very helpful in narrowing the affected genes to only those most likely to be important. Interestingly, gene dose effects are not linear in that overexpression of a cDNA, such as an oncogene, does not always activate the same genes reduced after gene disruption. This suggests that inferences about function from overexpressing a gene may be different from those after gene attenuation (Guo et al., 2000).

For tumor classification, a variant of the leave-one-out crossvalidation is commonly employed where for any tumor set, two-thirds are randomly selected and used to generate the class determination functions that are then tested on the remaining "held-out" one-third of the samples. This is done many times to arrive at an estimate of the performance of the classification algorithm. Alternatively, testing a gene list of classifiers from one study on the data of another can provide an assessment of the robustness of the classifiers (Shipp et al., 2002).

Database integration

The ability to assign biological importance to any array-based result is dependent on the quality and ease of access to information from a variety of databases. There are two types of databases that are relevant to the use of microarray technology as above. The first concerns the microarray measurements and experimental attributes and involves data management. The second concerns the linking of genes represented on the arrays with functional information that facilitates biological interpretation of the results. We note that, optimally, the two types of databases should be linked.

Managing expression data and experimental parameters

An effective array database tracks large-scale microarray data in a precise fashion and captures certain experimental attributes that may affect the expression ratio measurements. The essential components of such a database are as follows: first, the coordinates of the array probes and their corresponding signal measurements must be correctly mapped to a relational table that uniquely identifies each probe. Next, each set of array measurements must be linked to the biological or clinical information associated with the RNA samples, and this must be done in accordance to a strict controlled vocabulary. Finally, it must capture the experimental procedure and the protocols of sample collection and processing. This database infrastructure facilitates the extraction of information necessary for analysis of multiple data sets across a wide range of biological conditions. In addition to the database, uniform and effective experimental procedures are also important. A properly organized tissue

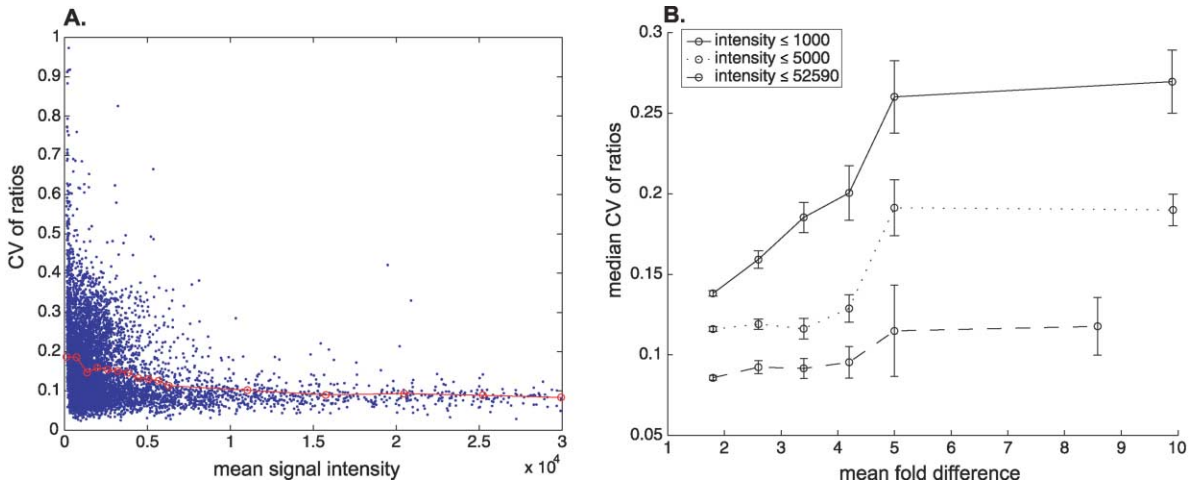

Figure 3. High variance in expression ratio measurements correlates with low signal intensities and higher ratios

A: Shown is the distribution of the CVs for replicate expression ratio measurements (vertical axis) for 9000 cDNA probes as a function of mean spot signal intensity (horizontal axis; using the average of Cy3 and Cy5 signals) where CV and mean intensity are derived from eight replicate array experiments. The red line shows the trend of the average CV at different intensity ranges. **B:** Using the same data set described in **A**, the median CV of ratio measurements (vertical axis) for probes belonging to six mean fold difference (ratio) bins (horizontal axis; bins are: 1.0 ≤ 1.8, >1.8 ≤ 2.6, >2.6 ≤ 3.4, >3.4 ≤ 4.2, >4.2 ≤ 5.0, >5.0 ≤ 10) is plotted at three mean signal intensity ranges (<1,000, >1,000 ≤ 5,000, >5,000 ≤ 52,590 signal units). The CV standard error is shown for each ratio bin at each signal intensity range.

repository, a strictly adhered to sample collection protocol, a uniform microarray platform, and consistent experimental procedures are needed to maximize the use and reuse of experiment results. At the Genome Institute of Singapore, we conform to this principle of a "perpetual array platform" whereby all microarray experiments performed use the same comprehensive microarrays, protocols, and pooled reference RNA, rendering the composite data crossreferable. Thus, over time, we will amass a transcriptional database ideally suited for hypothesis generation.

Gene features on microarrays: Annotation of meaning through database mining

Critical to the interpretation of array data is an understanding of the function and biological characteristics of the gene elements represented on a microarray. Database-integrated information regarding gene functions, domains, interactions, and pathways is useful in deciphering signaling components of biological systems. The detailed functional annotation of these genes requires interrogating a large number of disparate databases, many available in the public domain. For a comprehensive list of databases that are useful for microarray design and analysis, see Supplemental Table S1 at http://www.cancercell.org/cgi/content/full/2/5/353/DC1.

Importance of quality assurance: Garbage in, garbage out

The caveat "garbage in, garbage out" is an apt concern in microarray experiments. It is intuitive that if the quality of the arrays, probes, or samples is poor, the results and conclusions will be suspect. The challenge with microarrays lies in the difficulty in identifying these errors and their sources in such a massively parallel and multiprocess platform (see Table 1 for common sources of technical variance). Thus, a primary, but often neglected, consideration in array analysis is attention to quality control measures in the production and hybridization of arrays.

For spotted arrays, we separate the issues of quality assurance into two categories: manufacturing quality of the arrays and RNA quality. Though cDNAs have been the major source of probes, over time we and others discovered several troubling facets: high clone set error rates stemming from clone cross-contamination, mislabeling, missing inserts, and phage contamination were persistent problems, particularly in large, commercially available clone sets (Knight, 2001). Moreover, we encountered situations where crosshybridization limited the specificity of the probes, i.e., a cDNA probe would detect the summed expression of a number of gene paralogs (Aprelikova et al., 2001). As a result of such problems, many laboratories have begun migrating from cDNA to oligonucleotide arrays. The advantages of oligonucleotide probes are many and include greater specificity, uniformity of hybridization, minimization of contamination, and improved quality control measures; however, standard parameters for maximal sensitivity (e.g., oligo length, spotting concentration, attachment chemistry, etc.) are not yet well defined. We also find that oligo-based probes provide greater array design flexibility in that the availability of specific cDNA clones is not a limiting factor. Lastly, microarray printing artifacts that give rise to batch-to-batch and lab-to-lab variation are a common problem and should be considered when planning experiments (see Figure 2, Table 1). To minimize the disruption of experimental continuity, microarrays from the same print batch should be used for a given study when possible. Though this problem is more acute for "in-house" spotted arrays, similar caution should be used for all commercially available arrays. (For a comprehensive review of the microarray printing process and discussion of the technical challenges, see Eisen and Brown, 1999.) Similarly, it appears important to use consistent methods for RNA isolation and generation of labeled cDNA target (Wildsmith et al., 2001).

Perhaps one of the most important determinants of microarray data quality is the quality and quantity of the input RNA. Degraded, contaminated, or diluted RNA samples give rise to ineffective target with poor signal and reduced dynamic range. In these cases, attempts to salvage the data by statistical manipu-

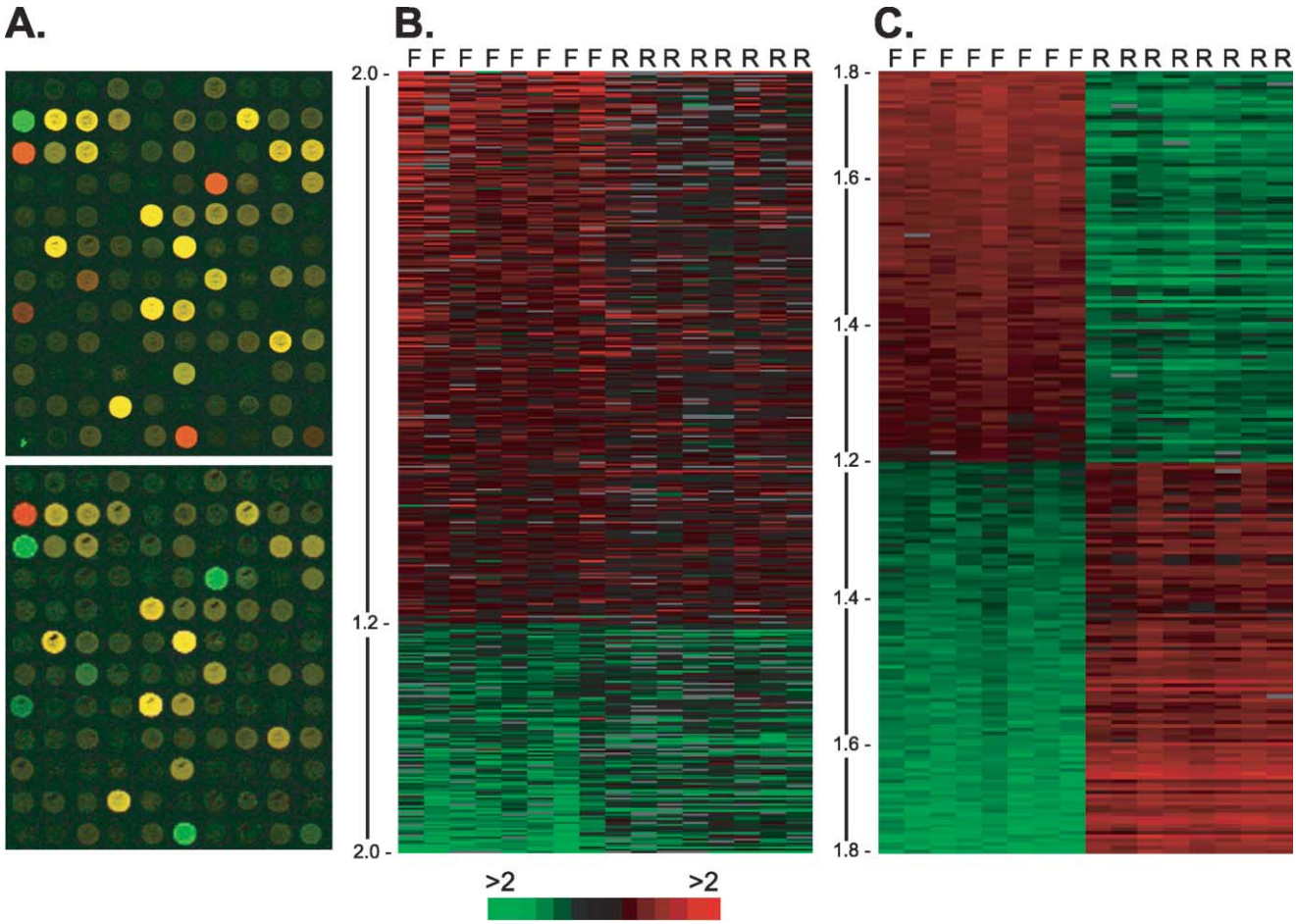

Figure 4. Dye swapping improves data reliability and increases confidence in the determination of outlier genes

A: Array images demonstrating the principle of dye swapping. The Cy dye labeling scheme used for one array (upper image) is reversed for the second array (lower image). **B:** Clustergram of 382 genes having mean ratios indicating fold change between 1.2 and 2.0 (eight replicates, labeled F) but do not reciprocate in dye swap experiments (eight replicates, labeled R) and therefore likely represent false positives owing to dye bias. **C:** Clustergram of 212 genes having mean fold change between 1.2 and 1.8 (arrays labeled F). Means were derived as described in **B** and were used to order the genes (see scale on left). Eight dye swap hybridizations (labeled R) show reciprocating ratios and thus provide added confidence that these results do not represent dye bias effects. Note, for clustergrams, columns represent array experiments and rows represent probes. The direction of expression ratios is indicated by red and green, and the magnitude of the ratios is reflected by the degree of color saturation (see color scale at bottom).

lations are rarely helpful and the experiment should be discarded. When RNA is limiting and RNA integrity cannot be physically assessed, RNA integrity can be estimated by array performance. We have found that the standard deviation of the log signal intensity in the target (e.g., tumor) channel is a good indicator of the normal diversity of gene expression seen in biological systems. When the standard deviation is less than 0.25–0.30, the RNA is generally unreliable (Assersohn et al., 2002).

In cases where it is known that the RNA quantity is insufficient for a single hybridization, an effective solution is linear RNA amplification. Most amplification approaches to date are derivatives of the T7-based method developed by the J. Eberwine laboratory (Van Gelder et al., 1990). Whereas 20 μg or more of total RNA is generally needed for a traditional microarray hybridization (or 5–10 μg for Affymetrix arrays), only 0.1–1 μg is required as starting material with amplification. Here, the linearity of the amplification reaction is of critical importance, for biases in the reaction that skew relative transcript ratios, even by a small amount, would render expression measurements unreliable. Independent evaluations have largely found the Eberwine amplification technique to be reliable for microarray use (Wang et al., 2000; Hu et al., 2002). In more in-depth comparisons of amplified versus unamplified RNA, we have detected only subtle biasing of the array readout from these amplification approaches. We found an overall compression of ratios such that the magnitude but not the direction of the differences is reduced. The net effect is that if standard cutoffs for outlier detection are used, the number of putative outliers may decrease by 8%–15% (Wang et al., 2000). Importantly, however, the variance of expression measurements does not appear to be increased following amplification. When taken together, RNA amplification does not appear to alter the structure of class distinctions and only marginally reduces the outlier determination for the purposes of individual gene discovery. It is worthwhile to recognize that the standard Affymetrix array platform routinely amplifies the RNA using the T7-based method and remains comparable to spotted arrays.

Understanding sources of variance in expression data

The ultimate goal of many microarray studies is to make biological inferences from gene expression patterns derived from population data. The robustness of the biological inference will be

dependent on the "noise" within the system that confounds the true differences between distinct biological populations. In expression genomics, the contributions to variance can be found in physical-chemical *noise* of molecular technologies and in the diversity of expression found in populations. The physical-chemical noise comes from RNA extraction, labeled target preparation, hybridization, or scanning and represents *within sample* variation. There are, however, gene expression variations from individual to individual within any population (*organismal* variation) that may have an impact on the outcome of the analysis. An example of such population variance is that despite identical experimental parameters, the same organs from isogenic mice will show distinct and detectable variability in expression (Pritchard et al., 2001). This form of "organismal" noise can only be assessed by multiple repeats at the organismal level (i.e., multiple mice for each biological group). Thus, in repeating array measurements, replicating at the level of individual samples is far better than repeating multiple arrays on the same sample. More specifically, RNA from ten different tumors, each subjected to a microarray analysis, provides more biological information than ten replicate arrays run on a single batch of RNA extracted from either a single tumor or pooled tumors.

Microarray experiments often seek to identify genes that vary significantly between biological states that are detectable above the background "noise." These are often referred to as "outlier" genes. One criterion for outlier detection uses an arbitrarily fixed cutoff, which is commonly in the range of 1.8- to 3-fold change. Lower ratios can sometimes be used to define outliers if there is sufficient replication and the variances of the system are taken into account. In replicate microarray hybridizations, it has been widely observed that the magnitude of the variance in signal is inversely correlated with the level of gene expression. Signals in the upper ranges are generally more reproducible and result in fairly stable expression ratios when measured across replicate hybridizations. In contrast, fluorescence intensities in the low range of detection (i.e., close to background) tend to be less reproducible and give rise to expression ratios that may fluctuate with considerable variance (Figure 3A). Interestingly, in our experience, we find that the highest variance at all signal intensities is associated with probes that report higher expression ratios, especially those greater than 4-fold difference (Figure 3B). Conversely, the most stable data are at ratios less than 2-fold at all signal intensities. This suggests that differences at lower ratios, even between 1- and 2-fold, have greater inherent accuracy. This finding has implications in the selection of the appropriate reference RNA. In microarray studies that seek to compare multiple tissues (e.g., tumor samples) via a common reference RNA, it is advisable to use a reference RNA that will target as many of the relevant arrayed probes as possible. If the reference channel has some baseline signal on the majority of the probe spots, then the magnitude of the expression ratios will be relatively limited (i.e., the denominator will infrequently approach zero). Thus, the variance in the ratio measurements will be kept low, resulting in more robust statistical analyses.

A corrective measure to identify and filter signal biases in spotted arrays is an experimental approach called *dye swapping*. Here, microarray hybridizations are performed in duplicate with the exception that the Cy3-Cy5 target labeling scheme is reversed or *swapped* between hybridizations, such that an outlier with a ratio in the "red" direction on one array should have a ratio in the "green" direction when the dyes are swapped, i.e., there is reciprocation (Figure 4A). This strategy is more rigorous than straight replication—which may otherwise identify dye bias artifacts as reproducible outliers. Through dye swapping, outliers on one array that fail to reciprocate on the other can be identified and flagged as unreliable (Figure 4B). The causes of this data unreliability may be related to signal intensity since nonreciprocating spots tend to have lower signal intensities. However, this approach also uncovers a small number of outliers that do not reciprocate despite having relatively high signal intensities. Without dye swapping, these spots may be counted as significant outliers. We routinely eliminate these nonreciprocating spots from further analysis. By combining dye swapping and filtering out spots with signal intensities near background (e.g., <2 SD above background), we can detect highly reproducible differential expression of genes with ratios as low as 1.2-fold (Figure 4C).

Can microarrays be used as a clinical test?

In the current format, expression profiling using microarrays is not sufficiently reliable to be used in making clinical decisions. This is because of the problems in standardization and performance as outlined above. Clearly, given the technical variance in these hybridization-based systems, precise and quantitative measurement of single genes using microarrays is not expected. However, the collective expression behavior of a set of gene classifiers, albeit imprecise, can provide valuable diagnostic information. This suggests that the more important approach to clinical diagnostics is to focus on algorithms that can integrate a large number of diagnostic markers into a clinically meaningful term, rather than on the exact technical platform that extracts this multiplex information. Even if the technical problems are solved, important conceptual questions remain: what is the contribution of stromal and infiltrating inflammatory cells? Will tumor heterogeneity affect the consistency of the results? How will the regulatory agencies evaluate a diagnostic where the individual determinations may have 20%–30% variance? We anticipate that these questions will be answered, but a concerted effort by the clinical trials community will be necessary.

Concluding remarks

The optimal analysis of microarray data should not only be focused on the statistical analysis of the array output, but also take into account quality control issues, study design factors, and database applications. As with any maturing technology, the applications and the approaches will migrate over time. Our prediction for the future is that dense arrays covering not only all possible genes but also noncoding RNAs will be in demand. In addition, the number of arrays used for each experiment will increase significantly so as to better address the scientific questions and refine the analyses. The linkage of these experiments into a relational database will require different data retrieval and analysis approaches, exchangeable formatting, and greater computational capabilities. All these predictions will require array costs to be dramatically reduced and the hybridization-scanning cycle to be automated—standard demands as a genomic tool advances. Given the power of this technology in speed and in permitting higher-order structuring of data, the possibilities are immense.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Ambroise, C., and McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA *14*, 6562–6566.

Aprelikova, O., Pace, A.J., Fang, B., Koller, B.H., and Liu, E.T. (2001). BRCA1 is a selective co-activator of 14–3-3 sigma gene transcription in mouse embryonic stem cells. J. Biol. Chem. *13*, 25647–25650.

Assersohn, L., Gangi, L., Zhao, Y., Dowsett, M., Simon, R., Powles, T.J., and Liu, E.T. (2002). The feasibility of using fine needle aspiration from primary breast cancers for cDNA microarray analyses. Clin. Cancer Res. *8*, 794–801.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA *98*, 13790–13795.

Desai, K.V., Xiao, N., Wang, W., Gangi, L., Greene, J., Powell, J.I., Dickson, R., Furth, P., Hunter, K., Kucherlapati, R., et al. (2002). Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc. Natl. Acad. Sci. USA *99*, 6967–6972.

Dobbin, K., and Simon, R. (2002). Comparison of microarray designs for class comparison and class discovery. Bioinformatics, in press.

Dudoit, S., Fridlyand, J., and Speed, T.P. (2002). Comparison of discrimination methods for the classication of tumors using gene expression data. J. Am. Stat. Assoc. *97*, 77–87.

Efron, B., and Tibshirani, R. (1997). Improvements on crossvalidation: the .632+ bootstrap method. J. Amer. Stat. Assoc. *92*, 548–560.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. J. Amer. Stat. Assoc. *96*, 1151–1160.

Eisen, M.B., and Brown, P.O. (1999). DNA arrays for analysis of gene expression. Methods Enzymol. *303*, 179–205.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *8*, 14863–14868.

Furey, T.S., Cristiannini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*, 906–914.

Guo, Q.M., Malek, R.L., Kim, S., Chiao, C., He, M., Ruffy, M., Sanka, K., Lee, N.H., Dang, C.V., and Liu, E.T. (2000). Identification of c-myc responsive genes using rat cDNA microarray. Cancer Res. *60*, 5922–5928.

Harrington, C.A., Rosenow, C., and Retief, J. (2000). Monitoring gene expression using DNA microarrays. Curr. Opin. Microbiol. *3*, 285–291.

Hu, L., Wang, J., Baggerly, K., Wang, H., Fuller, G.N., Hamilton, S.R., Coombes, K.R., and Zhang, W. (2002). Obtaining reliable information from minute amounts of RNA using cDNA microarrays. BMC Genomics *3*, 16.

Knight, J. (2001). When the chips are down. Nature *415*, 860–861.

McShane, L.M., Radmacher, M.D., Freidlin, B., Yu, R., Li, M.C., and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics, in press.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P., and Poggio, T. (1999). Support vector machine classification of microarray data. Technical Report *182*, AI Memo 1676, CBCL, 59–60.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Pritchard, C.C., Hsu, L., Delrow, J., and Nelson, P.S. (2001). Project normal: defining normal variance in mouse gene expression. Proc. Natl. Acad. Sci. USA *6*, 13266–13271.

Radmacher, M.D., McShane, L.M., and Simon, R. (2002). A paradigm for class prediction using gene expression profiles. J. Comput. Biol. *9*, 505–511.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. *8*, 68–74.

Simon, R., Radmacher, M.D., and Dobbin, K. (2002). Design of studies using DNA microarrays. Genet. Epidemiol. *23*, 21–36.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell *1*, 203–209.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps. Proc. Natl. Acad. Sci. USA *96*, 2907–2912.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., and Brown, P. (1999). Clustering Methods for the Analysis of DNA Microarray Data. (Stanford, CA: Stanford University Department of Statistics Technical Report).

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA *99*, 6567–6572.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. Proc. Natl. Acad. Sci. USA *98*, 5116–5121.

Van Gelder, R.N., von Zastrow, M.E., Yool, A., Dement, W.C., Barchas, J.D., and Eberwine, J.H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc. Natl. Acad. Sci. USA *87*, 1663–1667.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Wang, E., Miller, L.D., Ohnmacht, G.A., Liu, E.T., and Marincola, F.M. (2000). High-fidelity mRNA amplification for gene profiling. Nat. Biotechnol. *18*, 457–459.

Wildsmith, S.E., Archer, G.E., Winkley, A.J., Lane, P.W., and Bugelski, P.J. (2001). Maximization of signal derived from cDNA microarrays. Biotechniques *30*, 202–206, 208.

Yang, Y.H., and Speed, T. (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. *3*, 579–588.

Yeoh, E.J., Ross, M.E., Shurtleff, S.A., William, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Reilling, M.V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell *1*, 133–143.