



Microarray standard data set and figures of merit for comparing data processing methods and experiment designs

Yudong D. He, Hongyue Dai, Eric E. Schadt, Guy Cavet, Stephen W. Edwards, Sergey B. Stepaniants, Sven Duenwald, Robert Kleinhanz, Allan R. Jones, Daniel D. Shoemaker and Roland B. Stoughton*

Rosetta Inpharmatics Inc.[†], 12040 115th Avenue Northeast, Kirkland, WA 98034, USA

Received on July 24, 2002; revised on October 16, 2002; December 17, 2002; accepted on December 22, 2002

ABSTRACT

Motivation: There is a very large and growing level of effort toward improving the platforms, experiment designs, and data analysis methods for microarray expression profiling. Along with a growing richness in the approaches there is a growing confusion among most scientists as to how to make objective comparisons and choices between them for different applications. There is a need for a standard framework for the microarray community to compare and improve analytical and statistical methods.

Results: We report on a microarray data set comprising 204 *in-situ* synthesized oligonucleotide arrays, each hybridized with two-color cDNA samples derived from 20 different human tissues and cell lines. Design of the ~24 000 60mer oligonucleotides that report ~2500 known genes on the arrays, and design of the hybridization experiments, were carried out in a way that supports the performance assessment of alternative data processing approaches and of alternative experiment and array designs. We also propose standard figures of merit for success in detecting individual differential expression changes or expression levels, and for detecting similarities and differences in expression patterns across genes and experiments. We expect this data set and the proposed figures of merit will provide a standard framework for much of the microarray community to compare and improve many analytical and statistical methods relevant to microarray data analysis, including image processing, normalization, error modeling, combining of multiple reporters per gene, use of replicate experiments, and sample referencing schemes in measurements based on expression change.

Availability/Supplementary information: Expression data and supplementary information are available at http://www.rii.com/publications/2003/HE_SDS.htm.

Contact: yudong_he@merck.com

INTRODUCTION

DNA microarray technologies are now frequently used to measure genome-wide mRNA expression and its changes for diverse biological samples across conditions such as developmental stages, drug treatments, disease states, and gene disruptions (see review papers, e.g. Brown and Botstein, 1999; Friend, 2000; Young, 2000; Lockhart and Winzeler, 2000). Microarray applications include identification of disease-associated genes (Chee *et al.*, 1996; Heller *et al.*, 1997; Zhang *et al.*, 1997; Khan *et al.*, 1999; DeRisi *et al.*, 2000), drug target validation (Marton *et al.*, 1998; Huang *et al.*, 2000), biological pathway dissection (Spellman *et al.*, 1998; Holstege *et al.*, 1998; Iyer *et al.*, 1999; Roberts *et al.*, 2000; Gasch *et al.*, 2000), discovery of gene functions (Cho *et al.*, 1998; Walker *et al.*, 1999; Hughes *et al.*, 2000), experimental annotation of the human genome (Shoemaker *et al.*, 2001), compound toxicity studies (Waring *et al.*, 2001), tumor classifications (Alizadeh *et al.*, 2000; Golub *et al.*, 1999; Perou *et al.*, 2000), diagnostic and prognostic predictions for various types of cancer patients (Khan *et al.*, 2001; van 't Veer *et al.*, 2002; Van de Vijver *et al.*, 2002), and other biomarker identifications (Amundson *et al.*, 2000, 2001; Curto *et al.*, 2002).

Despite these great advances, the effective use of DNA microarrays still presents challenges for much of the community in areas such as experiment design (Kerr and Churchill, 2001; Yang and Speed, 2002), image processing (Yang *et al.*, 2001; Brown *et al.*, 2001; Li and Wong, 2001), normalization (Bilban *et al.*, 2002;

*To whom correspondence should be addressed.

[†] A wholly owned subsidiary of Merck & Co., Inc.

Yang *et al.*, 2002), supervised and unsupervised learning algorithms (Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Brown *et al.*, 2000; Roberts *et al.*, 2000; Khan *et al.*, 2001; van 't Veer *et al.*, 2002), and oligonucleotide probe design (Hughes *et al.*, 2001). These challenges include many issues in analytical and statistical methods for data analysis (Tseng *et al.*, 2001; Wang *et al.*, 2001; Yue *et al.*, 2001). In addition, differences in microarray types, tissue harvesting, amplification protocols, sample labeling techniques, and hybridization conditions make *post hoc* sharing of microarray data a great challenge (e.g. Bassett *et al.*, 1998; Brazma *et al.*, 2001; Gardiner-Garden and Littlejohn, 2001). This chain of processes determines how biological information is extracted from a set of microarray experiments. Optimization of the links in the chain is normally impractical because of the difficulties inherent in quantifying the impact of improvements in any one link. For example, adjusting image background subtraction so that the background-subtracted intensities no longer correlate with the background estimates has a sound statistical motivation, but what is its quantitative impact on false positives, on false negatives, and on finding interesting expression patterns? Different measures of correlation and similarity distance may be used to derive clustering or grouping of genes or experiments, but how does one quantify the success of the resultant cluster groups? Another difficulty with optimization is that optimization of each link in the chain *sequentially* does not in general provide a global optimum. On the other hand, optimization *simultaneously* over many aspects of data analysis in combination with experimental protocol and chip platform, is generally beyond the resources of any single research group. Many researchers have proposed their own algorithms in specific areas in order to extract more reliable and reproducible information (e.g. Lockhart *et al.*, 1996; Chen *et al.*, 1997; Roberts *et al.*, 2000; Schadt *et al.*, 2000, 2001; Li and Wong, 2001; Kerr *et al.*, 2000; Lee *et al.*, 2000; Newton *et al.*, 2001; Troyanskaya *et al.*, 2001; Tusher *et al.*, 2001). However, direct unbiased comparisons of different analysis methods are rare because different groups generally do not attack the *same* data set. The above problems indicate the need for a flexible data set that can be used to benchmark analysis methods and a set of standard figures of merit that define the benchmark. To be useful as a set for benchmarking analysis methods, the standard data set should have relevance to multiple microarray technology platforms, should possess built-in redundancy, should represent multiple experimental designs, and finally, should be comprised of a diverse set of conditions with variable signal strengths to support the evaluation of alternative data processing approaches and of alternative experiment and array designs. The standard figures of merit should be chosen so that they are defensible as robust evaluators

of relative performance. Our present work is in alignment with recent efforts in the microarray community to work together on microarray data analysis (Johnson and Lin, 2001a,b).

In this paper, we suggest two easily computed standard figures of merit that should be good predictors of information quality in the vast majority of microarray applications to gene expression. We illustrate application of these figures of merit to a new data set constructed with the benchmarking ideas just discussed in mind. To date, the microarray expression analysis community generally has explored two application types in a broad sense. The *Type 1* application is directed toward highly parallel measurements of individual expression changes or individual expression levels for many genes in response to specific conditions. Applications of this type include comparing diseased and normal tissue to identify disease-specific gene candidates (Pietu *et al.*, 1996; Heller *et al.*, 1997; Welsh *et al.*, 2001), comparing expression in different tissues to find out what genes are specific to a given tissue (Sturniolo *et al.*, 1999), or comparing compound treated samples and vehicle controls to identify genes that responded to the compound treatment (Waring *et al.*, 2001). A primary task here is to develop a reliable method for assigning statistical significance to the individual measurements of expression change, or error bars to the individual measurements of expression level.

The *Type 2* applications treat the entire set of expression measurements from a microarray as a pattern indicative of the state of the cell at the time it was profiled (Eisen *et al.*, 1998; Shoemaker *et al.*, 2001), or as a set of measurements of a given gene across many conditions as a pattern indicative of the role of that gene (Schena *et al.*, 1995; Ross *et al.*, 2000; Hughes *et al.*, 2000). This pattern may be matched or clustered with other patterns to obtain inferences about the cell state or gene role. In order to support this type of application, the reproducibility of a measured pattern becomes the key issue, while error outliers and detailed error modeling become relatively less important.

The standard data set presented here is based on the *in situ* ink-jet synthesizer (*IJS*) oligo array technology (Hughes *et al.*, 2001). This technology has many aspects in common with both spotter (Schena *et al.*, 1995) and Affymetrix GeneChip (Lockhart *et al.*, 1996) technologies. The data set contains built-in redundancy that supports evaluation of experiment designs and analysis methods for microarray profiling. The ink-jet oligonucleotide arrays and the pooling and pairings of samples for all hybridizations were specially designed to support multiple analysis objectives. In the Results, we first introduce non-parametric figures of merit appropriate to these two application types. After describing the standard data set, we then show the results from our baseline

processing of the standard data set for the purpose of illustrating the use of our figures of merit. We finally discuss implications of our results. Detailed information about experimental protocols and analysis methods are collected in the Supplementary information. We expect that a subsequent publication will report on more results from advanced processing methods on the same data set.

RESULTS

Figures of merit for microarray performance evaluation

The *Type 1* applications succeed to the extent we can report expression, or significant changes in expression, when they exist (true positives) and report them as absent when they are absent (true negatives). The other two possibilities of course are false positives and false negatives. With the obvious notation, there are two constraints on these four quantities,

$$TP + FN = N_1 \quad (1)$$

$$TN + FP = N_0 \quad (2)$$

where N_1 and N_0 denote the fraction that were truly positive or negative, so that $N_1 + N_0 = 1$. This means that the performance of our detection and reporting system can be characterized in the space of the remaining two free dimensions. We can use any two of the quantities, or for example two independent linear combinations of them, to coordinatize this performance. Traditionally, detection performance is expressed as the parametrized curve $TP(\lambda)$ versus $FP(\lambda)$; that is, true detections versus false detections, parametrized by the stringency of the detection threshold (see e.g. Egan, 1975). For historical reasons, this display is usually referred to as a Receiver Operating Characteristic (ROC) curve because it was first adopted for radar system applications during WWII. In *Type 1* applications we wish to minimize FP and FN while TP and TN are maintained, or alternatively we maximize TP and TN at fixed FP and FN.

In the case of expression profiling, the variable λ is a level of expression, or a level of logarithmic expression ratio, or a P -value output from an error model. It is somewhat problematic defining which reporters reflect actual change or real expression, since in real experiments there is a continuum of expression or change, with the largest number of genes exhibiting the smallest expression or smallest change. We therefore have chosen for our two-dimensional performance characterization the pair of quantities $(TP + FP)$ and FP ; that is, the total declared positive versus the false positives. Clearly, we can agree on the definition of $(TP + FP)$, since this is just the set of values above our detection threshold. FP we quantify by performing null experiments. In the case of expression change, the null experiment is the

comparison of nominally identical biological samples that have been put independently through the expression profiling assay. In the case of expression level, the null experiment might be a mock experiment with mRNA from the genes of interest omitted. This latter is itself a problematic experiment and we report specifically on the former in this paper. Thus the prototype experiment for the *Type 1* performance characterization involves the biological sample comparison A versus A to obtain the FP behavior versus detection threshold, and the comparison A versus B to obtain the behavior of $(TP + FP)$ versus threshold. Any change to the array technology platform, the data processing, or error modeling which, for these biological samples, improves the trade-off between $(TP + FP)$ and FP , is a change for the better. Most often the curves resulting from different processing methods in this parameter space nest rather than cross. When they do cross, it is necessary to fix one of the dimensions, for example FP as an operating point, or to use the area under the curves, in order to uniquely declare a winner.

Variations on this experimental approach are of course possible. We could use spiked-in known concentrations of mRNA sequences to provide a population of measurements that are known to be either positive or negative. However, in practice, two objections to this method are (1) the subtle biases between the way spike-ins are amplified and labeled, and the way the sequences in the actual biological samples are treated, and (2) the difficulty in providing a large variety of known sequences. The null experiment can be refined by averaging the results of A versus A and B versus B.

Results of this analysis still rest on an assumption of stationarity. For example, if the error behavior systematically changes between the time the A versus A experiment is conducted and the time the A versus B experiment is conducted, then the derived performance will be misleading. In our experiments, multiple null experiments were done interleaved in time with the others.

The *Type 2*, or pattern matching, application succeeds in principle to the extent that biologically similar patterns are declared similar and biologically different ones are declared different. We propose a figure of merit based on the ability to match an expression profile with one from a nominally biologically identical, but experimentally independently derived profile, in the presence of a large set of alternative competing patterns from other biologically relevant samples. Suppose we have done a set of microarray experiments for N biological samples in duplicates, yielding $2N$ profiles. All $2N(2N - 1)/2$ pairwise similarity metrics are computed as either correlation coefficients, Euclidean distance, or some other metric. Let r_{12i} be the rank of the similarity of the duplicate profile p_{i1} with respect to its true mate p_{i2} , among all pairs containing p_{i1} . The best outcome is that $r_{12} = 1$ and the worst possible

outcome is that $r_{12} = 2N - 1$. Let r_{21_i} be the rank of the similarity of p_{i2} with respect to p_{i1} among all pairs containing p_{i2} . The figure of merit is defined by

$$FOM2 = 2N / \sum_{i=1}^N (r_{12} + r_{21})_i. \quad (3)$$

If pattern matching succeeds in pairing up all of the nominally identical profile pairs, $FOM2 = 1$. If pattern matching performs at a level similar to random matching, we would expect $FOM2 \sim 1/N$ with a wide distribution. Note that this framework handles expression ratio profiles and expression level profiles. It also applies to the other dimension: similarity of the expression of gene i across a set of conditions with that of gene j across the same set of conditions. This dimension is the one relevant, for example, to grouping probes into exons, exons into genes, or genes into gene sets by expression co-regulation. No matter what image processing, normalization, detrending, correlation, or other similarity metrics have been defined, Equation (3) summarizes the success of pattern matching with that particular processing stream relative to alternative streams, and it has no free parameters.

In the *Type 1* and *Type 2* figures of merit, the independence of the repeated experiments is key. Processed microarray data values have an accumulation of uncertainties from all the experimental stages including the handling of the tissues, cultures, or animals that were the source of the RNA. For the standard data set described in this work, the duplicate experiments started with the same cRNA samples post-amplification, with everything downstream independently processed (see Methods). The different mRNA samples, however, came from different biological sources, human tissues, and cell lines (Table S1 in Supplementary information). Thus, this data set tests the ability to handle sources of error from all parts of the array manufacture and hybridization assay, certain enzymatic, fluorescent labeling, and purification steps, but not due to biological variations in the organism or tissue, a complication that has been removed from the present study.

Experiment design, chip design, sample preparation and hybridization protocols

Our experimental design is shown in Figure 1 in which each connector chord or spoke indicates a two-color sample pairing. See Supplementary information for details.

Overview of results from baseline processing

After hybridization, slides were scanned using a confocal laser scanner (Agilent Technologies). Fluorescence intensities on scanned images were quantified, corrected for background, and normalized using our baseline processing (see Supplementary information). A P -value for each ratio measurement was also assigned based

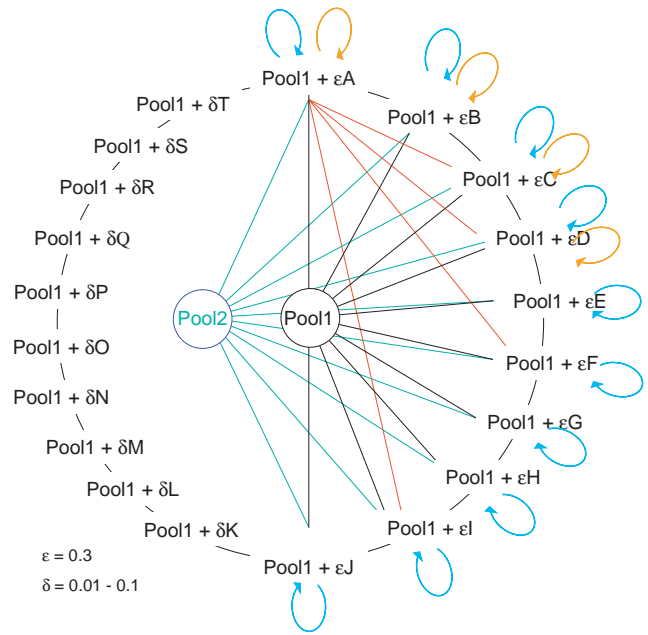


Fig. 1. Experiment design. Each connecting line or arc represents a redundant pair of fluor-reversal array pairs (total four hybridizations). Virtual ratios generated by traversing multiple arcs can be compared with actual two-color ratio profiles done as indicated by the chords. Pool 1 is an average of human tissue or cell line samples A + B + ... + T and so provides a near reference to the condition pairs on the ring. Pool 2 is an average of 10 tissue or cell line samples other than A through T and so provides a distant far-reference, such as would occur if cancer profiles were compared to normal rather than to a common pool of affected individuals. Loops indicate 'same versus same' control experiments (cRNA-to-end in blue color and hybridization-to-end in brown) which can be used to generate estimates of false positive rates and more detailed error models. For details of sample information and ϵ and δ see Table S1.

on an error model (see Supplementary information). Figure 2 shows $\log(\text{ratio})$ versus $\log(\text{intensity})$ for three experiment types each obtained from a fluor-reversal pair (FRP) of hybridization: Pool 1 + 0.3A versus Pool 1 + 0.3A (control), Pool 1 versus Pool 1 + 0.3A (A against near-reference), and Pool 2 versus Pool 1 + 0.3A (A against far-reference). The high quality of the data is indicated by the small number of false positives in Figure 2A compared to the positives in Figures 2B and C.

Figure 3 displays a two-dimensional clustering of the logarithmic expression ratio data matrix of 2424 reporters that were significantly differentially expressed with a P -value < 0.01 and more than 2-fold change ($|\log(\text{ratio})| > 0.3$) in more than 20 of the 88 experiments. We first note that each duplicate pair of experiments clusters together with a high similarity measure, leading to a perfect

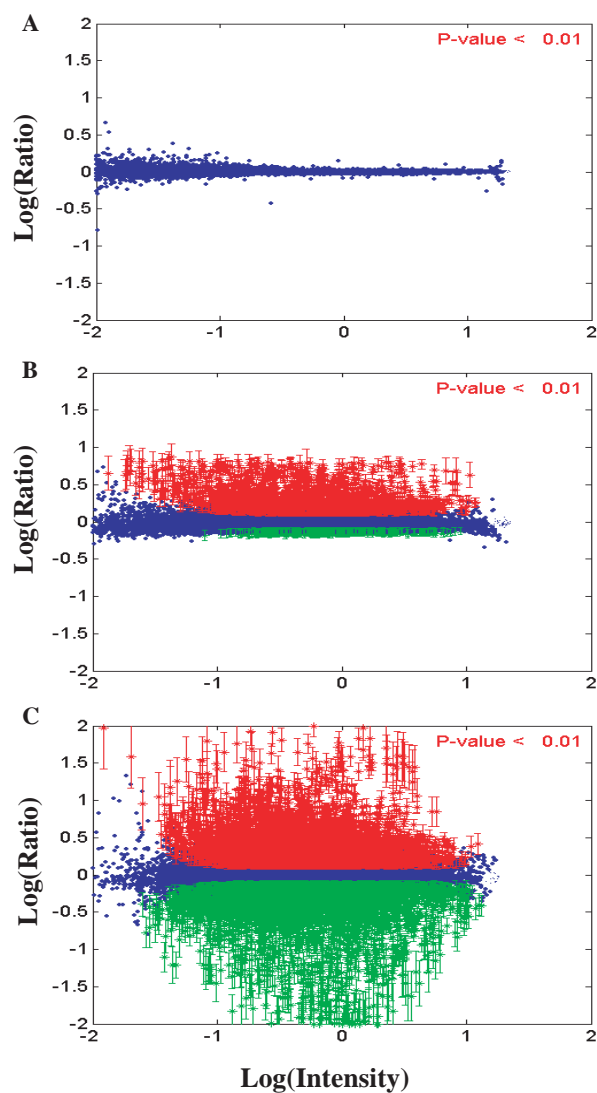


Fig. 2. Sample signature plots for three different experiment groups. (A) Control experiment [Pool 1 + 0.3A versus. Pool 1 + 0.3A], (B) Profile against near-reference pool [Pool 1 versus Pool 1 + 0.3A], (C) Profile against far-reference pool [Pool 2 versus Pool 1 + 0.3A]. In all plots, we show data for all the probes and flag signatures with P -value < 0.01 as red points for up-regulation and green points for down-regulation with errorbars.

score $FOM2 = 1$ for this data set from our baseline processing (see Fig. 3). This includes experiments with very subtle differential expression, such as those with $\alpha = 0.01$ and 0.02 . We find that we also obtain $FOM2 = 1$ when greater or smaller subsets (other than 2424) of reporters are selected while clustering the experiments, and when a Euclidean distance metric is used instead of the correlation-based metric. To provide a more challenging test of pattern matching, a mask is provided

with the published data set that will deselect the brightest reporters. With this mask in place, $FOM2$ drops to 0.87 for the whole data set of 102 experiments when using our baseline processing.

The striking rectangular pattern with strong differential expression in Figure 3 is formed by experiments where individual samples A through J were paired against the far-reference pool (Pool 2); the systematic difference between these samples and Pool 2 dominates the pattern, as expected, and ratio profiles are similar at a correlation level of 0.93 or above. This leaves a narrow dynamic range from 0.93 to 1 for differentiating individual samples A through J. Nevertheless, we find all duplicates paired up and $FOM2 = 1$ for this experiment subset. We return below to more extensive application of $FOM2$.

Illustration of figures of merit for *Type 1* application

Figure 4 shows the fraction of probes that were false positive (FP) versus the fraction what were positives (FP + TP), as the threshold P -value for declaring expression differences is changed. Given the diversity of our samples A through T, we averaged the number of positives for a given P -value threshold over all 88 experiments. We also averaged the number of false positives for the same P -value threshold over all 14 control-type experiments. The upper curve corresponds to baseline processing where a fluor-reversed pair (FRP) of arrays is optimally averaged (see Methods). The other curves show the benefits that were obtained by combining the 2 two-color arrays, and by using the two-color protocol rather than a single color protocol with this technology platform. It is worth noting that at a fixed normalized number of false positives 0.0001, the normalized number of positives is found to be 0.03 from a single two-color slide and 0.09 from an FRP. This factor-of-three improvement in the detection sensitivity at the same specificity may be compared to the rough doubling of cost associated with performing the FRP. We also observed that the variance reduction factor gained by using an FRP over a single slide is larger than the $\sqrt{2}$ that would be expected for uncorrelated errors, consistent with systematic biases that depend on the gene and the fluorescent label. These results confirm that the FRP is cost effective for this technology.

Figure 5 is similar to Figure 4 but compares performance obtained with probe level, exon level, and transcript level detection. The lower curve is the same as the upper curve in Figure 4. When hybridization data for multiple probes representing the same exon are ‘combined’ into one data point (see Supplementary information for details), the resultant average is a more robust detection statistic, and similarly for combining of exons into transcripts. We note that alternative splicing can affect the ‘averaging’ from exon level to transcript level. Figure 5 shows

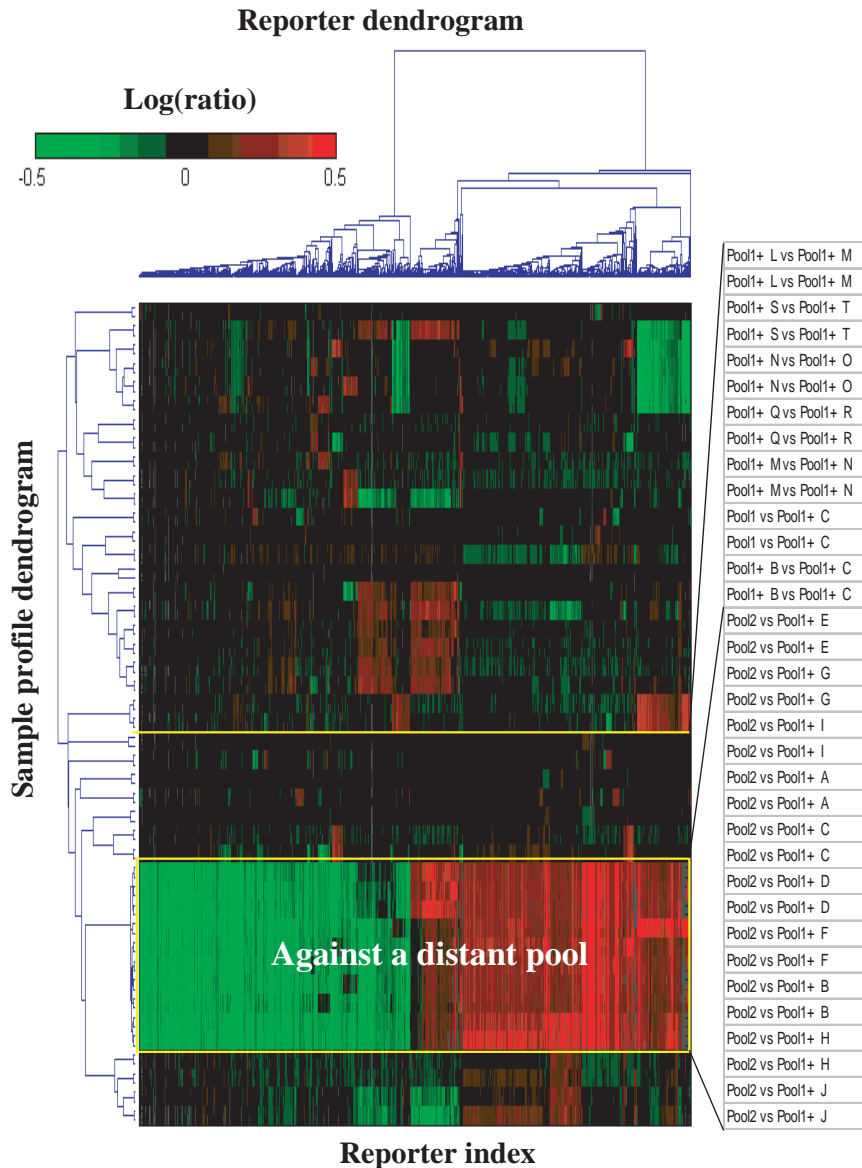


Fig. 3. Two-dimensional clustering results. Logarithmic ratios of gene expression measured across probes (horizontal coordinate) and tissue mixture pairs (vertical coordinate) with rows and columns reordered using hierarchical agglomerative clustering with correlation-based similarity metric. As shown in color bar, red represents $\log(\text{ratio}) > 0$, green $\log(\text{ratio}) < 0$, and black no change. Gray denotes no valid data points. Only the significantly differentially expressed 2424 reporters (columns) are shown. At the right side, we provide sample labels for a subset of the experiments.

that at fixed sensitivity, specificity improves from probe to exon level detection by a factor of 2, and from exon to transcript level detection by a factor of 10 (normalized number of false signatures is 0.002, 0.001, and 0.0001 respectively for probe, exon, and transcript level detection). We used fairly simple algorithms for both levels of combining (see Supplementary information); more sophisticated methods will be the subject of a future paper.

Illustration of figures of merit for *Type 2* application

Table 1 lists FOM2 calculated at probe level detection for two-color hybridization data from an FRP, for slide #1 and slide #2 separately, for Cy3 channel data from an FRP, and for Cy5 channel data from an FRP, based on all probes on the array (upper part of table, 23 841 probes). We calculated FOM2 based on pattern matching across the whole data set of 102 independent experiments. Since it is

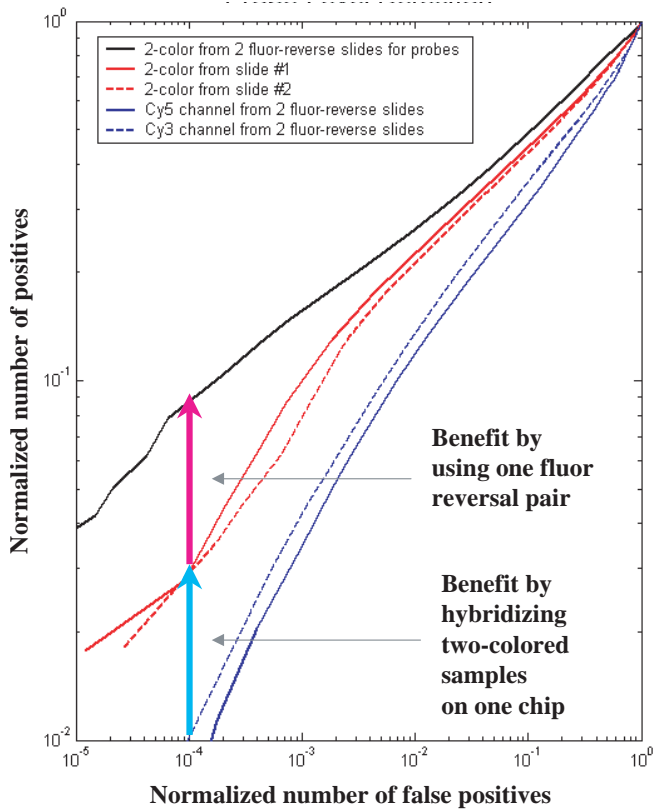


Fig. 4. FOM1 for type 1 application. Normalized number of detected expression changes versus normalized number of false positives, for probe level detection. The horizontal coordinates are based on the average over 14 control ‘same versus same’ experiments, while the vertical coordinates are based on the average over 88 condition pairs (44 done in duplicate) that were biologically different. Each curve is parametrized by the confidence threshold for detection. Upper curve: results based on measurements after combining each fluor-reversed pair of arrays. Middle curves: results based on the individual members of the fluor-reversed pair. Bottom curves: results when using the two-color data as if it were single-color. In this case expression differences were formed by comparing the Cy5 channel of one member of the fluor-reversed pair with the Cy3 channel of the other, and vice versa. Vertical arrows indicate the improvement in sensitivity at fixed false positive rate achieved in going from a one-color approach to the two-color approach (lower arrow) and in combining the fluor-reversed pairs (upper arrow).

an average over 102 similarity searches each of which can have 101 different rank outcomes, it is a robust statistic. We also computed FOM2 separately within each subset of experiments: the near-reference pool group (10 duplicated experiments), the far-reference pool group (10 duplicated experiments), the ring (20 duplicated experiments) and chord (four duplicated experiments) group, and the control group (14 experiments). In these cases, there are fewer profiles to produce false matches, so FOM2 cannot be directly compared across the experiment groups.

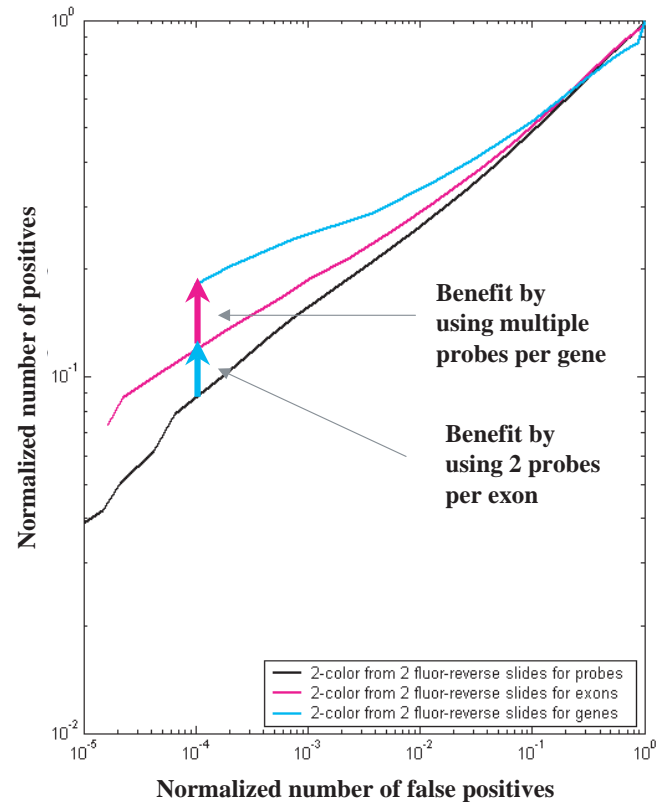


Fig. 5. FOM1 for probe combining. Similar to Figure 4, but showing the gains from probe combining. Probe level data were after FRP combining. Bottom curve is the same as the upper curve in Figure 4. Middle curve is after combining the ~2 probes for each exon. Upper curve is after combining the exons according to their associated transcripts.

We can purposely make the task of pattern matching more difficult by masking off the brightest probes in the data set. Keeping only the bottom ~4% based on the sum of Cy3 and Cy5 channel intensities on an FRP across all 88 experiments, we are left with 889 probes. We found that FOM2 degrades as shown in the lower part of Table 1 after the masking. Nevertheless, it is noted that FOM2 = 1 for the near-reference group based on FRPs or either slide of the FRP. For the ring and chord group, only FRPs lead to FOM2 = 1. For the far-reference group, none of the cases leads to perfect pattern matching. The overall success of pattern matching based only on ~4% of weak reporters is somewhat surprising, but shows the power of the highly reproducible profiles from the *IJS* platform. Similar calculations can be carried out for ratio profile data at the exon level and transcript level, and for expression level profiles appropriately normalized (see Supplementary information).

Table 1. Values of FOM2 pattern-matching performance obtained for various cases

	All profiles	Near-ref	Far-ref	Ring+chord	Control
# of hybridizations	204	40	40	96	28
All probes on array					
1-color Cy3	0.49	1	0.95	0.70	0.12
1-color Cy5	0.37	1	1	0.43	0.13
2-color slide #1	1	1	1	1	0.11
2-color slide #2	1	1	1	1	0.12
2-color FRP	1	1	1	1	0.12
Probes after masking					
1-color Cy3	0.06	0.33	0.16	0.10	0.17
1-color Cy5	0.03	0.19	0.17	0.05	0.14
2-color slide #1	0.73	1	0.54	0.89	0.13
2-color slide #2	0.57	1	0.83	0.64	0.13
2-color FRP	0.87	1	0.61	1	0.12

Values were estimated when applied to two-color ratio profiles from the standard data set at probe level. FOM2 = 1 is a perfect score. FOM2s may be compared within columns, but it is not meaningful to compare across columns.

Ring versus spoke sample referencing schemes

See Supplementary information for calculations showing the loss of accuracy associated with *in-silico* generation of all pairwise profile comparisons, and how this loss affects the choice of sample reference scheme.

DISCUSSION

The two figures of merit were carefully chosen to align with the two ways in which researchers approach expression profiling: as highly parallel northern blotting of many individual genes, or as patterns indicating cell state. The *Type 1* figure of merit focused on the trade off between sensitivity and specificity. At a fixed (tolerable) false positive rate, it can be argued that the number of detections is a rough measure of the information value in the data in this *Type 1* context. We found, for example, that repeating a two-color experiment with fluor reversal delivered more than twice the information value when measured this way.

The *Type 2* figure of merit focused on the ability to distinguish different expression patterns and identify similar ones. The similarity rank of the nominally identical profiles is roughly proportional to the time that will be spent tracking down false leads generated by pattern similarity observations, thus this figure of merit is closely related to the cost effectiveness of the profile data in this *Type 2* context. It was somewhat surprising how well the pattern matching performed: biological sample mixtures differing by only 1% were reliably distinguished. The robustness of pattern matching seems to be a fundamental property achieved from averaging over a large number of reporters.

We have presented just a few examples illustrating the application of the standard data set and figures of merit, including comparing performance of ring and common reference experiment designs, showing the gains achieved with replicates and fluor-reversal pairs in a two-color protocol, and demonstrating the gains associated with averaging multiple probes per exon, and multiple exons per transcript. Although we concentrated on expression differences, analogous analyses could be performed on expression levels obtained from single channel data.

The objective comparison methodology and data presented here can support a wide range of methods development tasks, including the improvement of probe sequence design, probe-specific error models and probe combining schemes, detection of alternative splicing, definition of similarity metrics, and image processing of the original array images including background subtraction and normalization. The data set is relevant to spotter technologies and to Affymetrix chips since the *IJS* technology and the array design and hybridization protocols chosen here have many aspects of both. We hope the microarray community can use this framework to generate illuminating head-to-head comparisons of promising analysis methods and technology variations.

ACKNOWLEDGEMENTS

We thank Stephen H. Friend and Peter S. Linsley for encouragement and support; Chris Roberts and Kellye Daniels for expert management of the hybridization experiments; Chris Armour, Phil Garrett-Engle, and Patrick Loerch for sample preparation protocol; Deborah Kessler, Tom Fare, Matthew Kidd, Yanqun Wang, and Lee Weng for discussions; Matthew Marton, Mollie McWhorter, Tricia Bishop, Kim Howe, and others at Rosetta Gene Expression Laboratory for execution of the sample preparation and microarray experiments.

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.
- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.
- Amundson, S.A., Do, K.T., Shahab, S., Bittner, M., Meltzer, P., Trent, J. and Fornace, Jr, A.J. (2000) Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. *Radiat. Res.*, **154**, 342–346.
- Amundson, S.A., Bittner, M., Meltzer, P., Trent, J. and Fornace, A.J. (2001) Induction of gene expression as a monitor of exposure to ionizing radiation. *Radiat. Res.*, **156**, 657–661.

- Bassett,D.E., Eisen,M.B. and Boguski,M.S. (1998) Gene expression informatics—it's all in your mine. *Nat. Genet.*, **21**, 51–55.
- Bilban,M., Buehler,L.K., Head,S., Desoye,G. and Quaranta,V. (2002) Normalizing DNA microarray data. *Curr. Issues Mol. Biol.*, **4**, 57–64.
- Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. and Causton,H.C. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
- Brown,C.S., Goodwin,P.C. and Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 8944–8949.
- Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
- Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, **21**, 33–37.
- Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P. (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
- Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA micro-array images. *J. Biomed. Opt.*, **2**, 364–374.
- Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
- Curto,E.V., Lambert,G.W., Davis,R.L., Wilborn,T.W. and Dooley,T.P. (2002) Biomarkers of human skin cells identified using DermArray DNA arrays and new bioinformatics methods. *Biochem. Biophys. Res. Commun.*, **291**, 1052–1064.
- DeRisi,J., Penland,L., Brown,P.O., Bittner,M.L., Meltzer,P.S., Ray,M., Chen,Y., Su,Y.A. and Trent,J.M. (2000) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.*, **14**, 457–460.
- Egan,J.P. (1975) *Signal Detection Theory and ROC Analysis*. Academic Press, New York.
- 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*, **95**, 14 863–14 868.
- Friend,S.H. (2000) Genomic approaches to drug discovery. *Adv. Oncology*, **16**, 3–11.
- Gardiner-Garden,M. and Littlejohn,T.G.A. (2001) Comparison of microarray databases. *Brief Bioinform.*, **2**, 143–158.
- Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
- Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Heller,R.A., Schena,M., Chai,A., Shalon,D., Bedilion,T., Gilmore,J., Woolley,D.E. and Davis,R.W. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl Acad. Sci. USA*, **94**, 2150–2155.
- Holstege,F.C., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
- Huang,P., Feng,L., Oldham,E.A., Keating,M.J. and Plunkett,W. (2000) Superoxide dismutase as a target for the selective killing of cancer cells. *Nature*, **407**, 390–395.
- Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H. He,Y.D. et al. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
- 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.
- Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,Jr.,J. Boguski,M.S. et al. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
- Johnson,K. and Lin,S. (2001a) Call to work together on microarray data analysis. *Nature*, **411**, 885.
- Johnson,K.F. and Lin,S.M. (2001b) Critical assessment of microarray data analysis: the 2001 challenge. *Bioinformatics*, **17**, 857–858.
- Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays, preprint, the Jackson Laboratory.
- Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
- Khan,J., Bittner,M., Chen,Y., Meltzer,P.S. and Trent,J.M. (1999) DNA microarray technology: the anticipated impact on the study of human disease. *Biochimica et Biophysica Acta*, **1423**, M17–M28.
- Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.
- Lee,M.L.T., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.
- Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
- Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.
- Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Marton,M.J., DeRisi,J.L., Bennett,H.A., Iyer,V.R., Meyer,M.R., Roberts,C.J., Stoughton,R., Burchard,J., Slade,D. Dai,H. et al. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.*, **4**, 1293–1301.

- Newton, M.A., Kendzioriski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.
- 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.
- Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Samson, R., Houlgatte, R., Soularue, P. and Auffray, C. (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of high density cDNA array. *Genome Res.*, **6**, 492–503.
- Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L. Hughes, T.R. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.
- Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S. Van de Rijn, M. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
- Schadt, E.E., Li, C., Su, C. and Wong, W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.*, **80**, 192–202.
- Schadt, E.E., Li, C., Ellis, B. and Wong, W.H. (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem.*, **Suppl. 37**, 120–125.
- Schena, M.D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
- Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y. Cavet, G. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.
- Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Sturniolo, T., Bono, E., Ding, J., Radrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P. Sinigaglia, F. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555–561.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
- Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.
- Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
- 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. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
- Van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C. Marton, M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.
- Walker, M.G., Volkmoth, W., Sprinzak, E., Hodgson, D. and Klingler, T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.*, **9**, 1198–1203.
- Wang, X., Ghosh, S. and Guo, S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, E75.
- Waring, J.F., Ciurlionis, R., Jolly, R.A., Heindel, M. and Ulrich, R.G. (2001) Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. *Toxicol. Lett.*, **120**, 359–368.
- Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A. and Hampton, G.M. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA*, **98**, 1176–1181.
- Yang, G.P., Ross, D.T., Kuang, W.W., Brown, P.O. and Weigel, R.J. (1999) Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acids Res.*, **27**, 1517–1523.
- Yang, Y.H., Buckley, M.J. and Speed, T.P. (2001) Analysis of cDNA microarray images. *Brief Bioinform.*, **2**, 341–349.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, E15.
- Yang, Y.H. and Speed, T.P. (2002) Design issues for cDNA microarray experiments. *Nat. Genet. Rev.*, **3**, 579–588.
- Young, R.A. (2000) Biomedical discovery with DNA arrays. *Cell*, **102**, 9–15.
- 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. *et al.* (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, E41–1.
- Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.
- Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. and Sakaki, Y. (1995) High-density cDNA filter analysis: a novel approach for large-scale quantitative analysis of gene expression. *Gene*, **156**, 207–213.