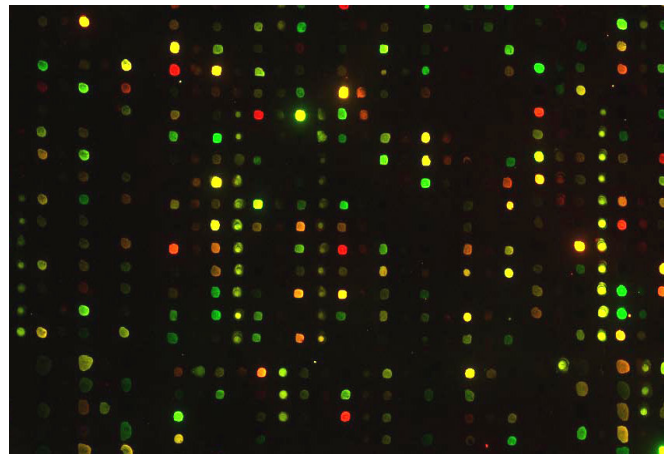
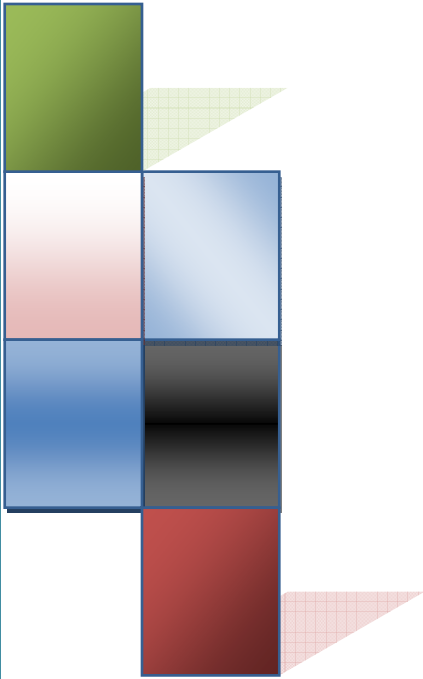


NanoCinna

Pharmacogenomics Center

Microarray, an Overview

Norman H. Lee and Alexander I. Saeed



Reference:

1. Elena Hilario and John Mackay. Protocols for Nucleic Acid Analysis by Nonradioactive Probes. Book 2007, eISBN 1-59745-229-7 • 978-1-59745-229-8.

5th Floor, No.56, Azimi St. Phase 1, Shahrak Ekbatan. Tehran-Iran
Tel: + 98-2144654896, +98-21-44654490
Fax: +98-21-44654896

18

Microarrays

An Overview

Norman H. Lee and Alexander I. Saeed

Summary

Gene expression microarrays are being used widely to address a myriad of complex biological questions. To gather meaningful expression data, it is crucial to have a firm understanding of the steps involved in the application of microarrays. The available microarray platforms are discussed along with their advantages and disadvantages. Additional considerations include study design, quality control and systematic assessment of microarray performance, RNA-labeling strategies, sample allocation, signal amplification schemes, defining the number of appropriate biological replicates, data normalization, statistical approaches to identify differentially regulated genes, and clustering algorithms for data visualization. In this chapter, the underlying principles regarding microarrays are reviewed, to serve as a guide when navigating through this powerful technology.

Key Words: Amplification; ANOVA; cDNA; clustering; dye-swap; expression matrix; expression vector; fluorescence; gene expression; hierarchical; hybridization; *k*-means; long oligomers; microarrays; mRNA; oligonucleotides; probe; RNA; target; *t*-test.

1. Introduction

In the past several years, we have witnessed remarkable progress in the completion of draft genome sequences for human, mouse, rat, *Drosophila*, *Arabidopsis*, and hundreds of microbial species. Couple this sequence information with complementary methodologies (e.g., expressed sequence tags sequencing projects for eukaryotic organisms) to catalog the expressed gene repertoire of an organism, and it becomes clear that we are arriving at an increasingly impressive “parts list” (1). However, interpreting these gene lists in terms of an organism’s underlying biology remains a challenge. This is compounded by the fact that more than half of the identified genes of most organisms have no

From: *Methods in Molecular Biology*, vol. 353:
Protocols for Nucleic Acid Analysis by Nonradioactive Probes, Second Edition
Edited by: E. Hilario and J. Mackay © Humana Press Inc., Totowa, NJ

obvious biological function. Moreover, predicted gene function based on sequence homology to genes with known cellular roles remains to be validated in the laboratory. Hence, DNA microarrays have become a universal tool to link gene expression with biological consequence (2). The power of microarrays is in their ability to simultaneously assess the expression of thousands of genes in a massively parallel fashion and across numerous conditions (e.g., comparing across time, treatment groups, genetic strains, and so on) to uncover higher-order organization of gene transcriptional behavior and, ultimately, to understand biology. The advantage of microarray analysis is not in viewing the expression of genes as individual components, but rather in visualizing the data as a “composite image” to understand biological processes (2).

2. Microarray Platforms

2.1. Overview

The fundamental steps behind a DNA microarray experiment are as follows: obtain RNA from two or more experimental groups that are being compared, convert the RNA to antisense RNA (aRNA) or complementary DNA (cDNA; herein referred to as the target), label the target with a fluorophore, hybridize the labeled target against thousands of DNA probes/elements (representing genes) immobilized on a solid support surface, and measure the relative expression (i.e., fluorescence) of each gene in each of the groups. However, how these steps are accomplished can be quite varied. There is no single “microarray platform” and new technologies are being introduced in an attempt to improve throughput and sensitivity, such as the Universal Hexamer Array from Agilix (3) and Illumina Beadarray (4). That being said, the most established microarray platforms in use are the commercially available Affymetrix GeneChips (Santa Clara, CA, <http://www.affymetrix.com/index.affx>), in-house manufactured PCR amplicon-based cDNA arrays, and, more recently, long oligomer arrays (e.g., 70-mer oligonucleotides) that are manufactured in-house or commercially available (*see* Agilent, Palo Alto, CA, <http://we.home.agilent.com>; Sigma-Genosys, Jamesburg, NJ, <http://www.sigma-genosys.com/oligonucleotide.asp>; Illumina, San Diego, CA, <http://www.oligator.com>; Operon, Huntsville, AL, <http://www.operon.com>; and MWG Biotech, High Point, NC, <http://www.mwg-biotech.com>).

The 25-mer short oligonucleotide probes contained in the Affymetrix GeneChips are synthesized directly onto a solid matrix using a proprietary photolithographic technology (5,6). The probes are “freely” moving, being tethered at one end to the solid support surface. A key advantage of this platform over nonsynthetic methods (mechanically spotted arrays; *see* **Subheading 2.1.**) is that the burdensome aspects of probe handling and tracking are eliminated. Another advantage of this high-precision photolithographic approach is that the use of synthetic reagents minimizes chip-to-chip variation. However, one

drawback is the expense incurred by these arrays, which can be 5- to 10-fold higher than in-house manufactured arrays.

For PCR amplicon and long oligomer arrays, the probes are either mechanically “spotted” onto modified glass slides by direct contact printing or deposited onto the slide surface by ink jet printing (7,8). After spotting, the probes are subjected to UV irradiation to covalently attach the probes onto the surface of the slide. Typically, the spotted PCR amplicons are 500 to 1000 bases in length, whereas the long oligomers are 55 to 70 nucleotides long. Alternatively, long oligomers can be synthesized *in situ* onto the glass slide surface (*see* Agilent). In the case of the Agilent long 70-mer arrays, similar to the Affymetrix GeneChips, one end of the probe is tethered to the solid support surface. It is thought that this design strategy, in contrast to UV-crosslinked probes that lie flat on the slide surface, increases the available probe surface area for probe–target hybridization and, hence, provides greater signal-to-noise benefits. To mimic this design of freely moving probes, both PCR amplicons and long oligomers can be modified at one end by the addition of a 5′ amino group for covalent attachment onto preactivated slides (9,10). Long oligomer arrays are gaining widespread popularity among the spotted array platforms. Problems associated with PCR amplicon-based arrays, such as clone tracking, handling of glycerol stocks, and failed PCR amplifications, are avoided when using long oligomers. With the completion of numerous microbial, plant, and eukaryotic genomes, as well as extensive expressed sequence tags data, there is sufficient sequence information to design unique long oligonucleotide probes capable of distinguishing homologous genes, alternative splice variants, and partially overlapping genes found on opposite DNA strands of compact genomes. As such, long oligomer probes have an added flexibility over PCR amplicons.

2.2. Comparative Analysis of Different Platforms

Inherent across all of these platforms is a multitude of different methodologies for generating labeled target from starting RNA, various hybridization and wash conditions, different microarray scanners (which are not necessarily interchangeable across platforms), a host of image segmentation and quantification techniques, and a multitude of approaches for background noise estimation and normalization. For a more thorough discussion of these differences and their impact on expression measurements, we refer the reader to more specialized reviews (11–13). Comparative analysis of the various platforms has been reported in the literature, with somewhat varying outcomes. Although there seems to be good concordance of gene expression measurements between long oligomer and PCR amplicon-based cDNA arrays (14), there have been conflicting reports regarding the correlation between Affymetrix and amplicon-based measurements (15–17). A careful analysis of the probe sets from each platform

contributing to the discordant data is warranted in the very near future. For example, are the discrepancies caused, in part, by unwanted cross-hybridization properties exhibited by probes in one platform but not by the analogous probes in another platform, or are the conflicting results a reflection of an inability of the longer PCR amplicon-based probes to differentiate alternatively spliced transcripts? Ultimately, it will be up to the user to identify the best platform for their particular application and to validate the microarray results using an alternative RNA quantification method (e.g., Northern blot, real-time reverse transcriptase PCR, and so on; [ref. 18](#)).

3. RNA-Labeling Strategies

3.1. Affymetrix Arrays: One-Color Scheme

The RNA-labeling protocol in use for the Affymetrix GeneChip is vastly different from the approach favored by microarrays comprising the longer probes (i.e., 70-mer oligonucleotides and PCR amplicons). Affymetrix arrays use a one-color scheme by using a single fluorescent label (i.e., phycoerythrin). Expression profiles of each sample are generated on separate chips, and the different fluorescent images are compared against one another for determination of differential gene expression ([5](#)). Using the Eberwine method ([19](#)), total RNA (5 to 10 μg) from an individual sample is primed with an oligo-dT primer with a T7 promoter sequence at the 5'-end ([Fig. 1](#)). Reverse transcriptase is added to generate single-stranded cDNA, followed by addition of RNase H, DNA polymerase, and DNA ligase to synthesize double-stranded cDNA containing the T7 promoter. In the presence of biotin-labeled rNTPs, T7 RNA polymerase uses the double-stranded cDNA as a template to synthesize multiple copies of biotinylated aRNA. The biotinylated aRNA is fragmented, hybridized onto the GeneChip, and stained with a streptavidin-phycoerythrin conjugate before scanning to generate a fluorescent image.

3.2. PCR Amplicon and Long Oligonucleotide Arrays: Two-Color Scheme

In the two-color format, specifically developed for long oligomer and PCR amplicon-based arrays, the two RNA samples being compared are reverse transcribed into cDNA (in separate reaction tubes), and different fluorescent tags (typically Cy3 and Cy5) are incorporated into the two cDNA molecules. Next, the Cy3- and Cy5-labeled cDNA targets are co-hybridized overnight onto a single array, and scanned the following day to generate a two-channel fluorescent image ([7](#)). The relative Cy3 and Cy5 fluorescent intensities associated with each probe on the array provide comparative gene expression information in the two samples.

There are two main protocols for target labeling, and both require 10 to 15 μg total RNA as the starting material. In the direct-labeling approach ([7](#)), a Cy3

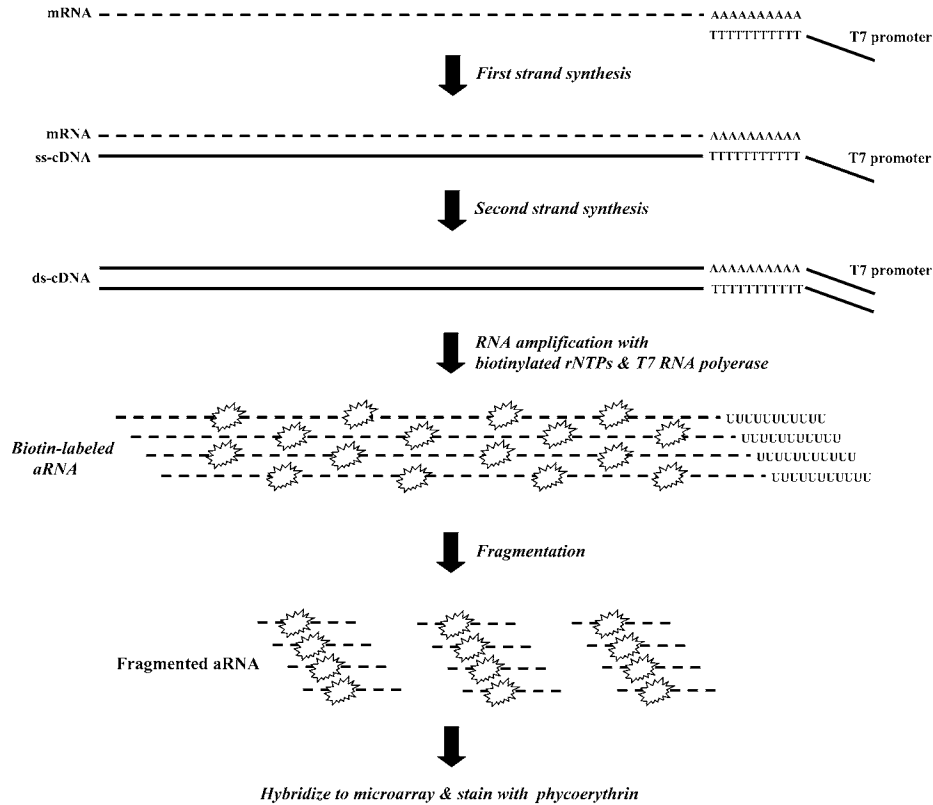
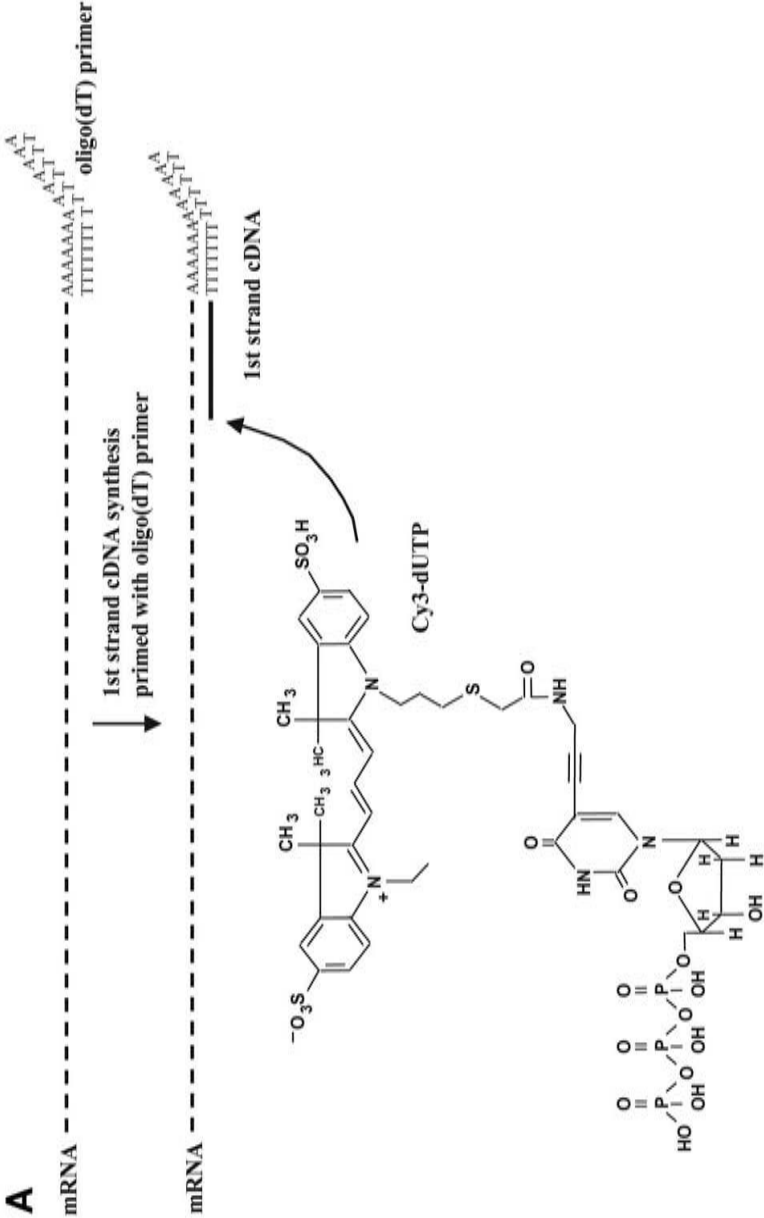


Fig. 1. T7 RNA polymerase amplification method to generate biotinylated aRNA for hybridization onto Affymetrix GeneChips.

or Cy5 fluorescently conjugated nucleotide (e.g., Cy3-dUTP or Cy5-dUTP) is directly incorporated into the first-strand cDNA synthesis (Fig. 2A). This particular labeling strategy was used in many of the early microarray publications. However, the large bulk of the fluorescent dye presents significant steric hindrance problems for the reverse transcriptase, resulting in low incorporation of Cy dye molecules into the cDNA target. Moreover, differences in the sizes of the Cy3 and Cy5 fluorophores resulted in unevenly labeled cDNA target (i.e., the cDNA labeled with the larger Cy5 molecule typically exhibits poorer dye incorporation compared with the cDNA labeled with the smaller Cy3 molecule). Hence, the indirect-labeling approach was developed to eliminate these problems (Fig. 2B; refs. 20 and 21). Instead of using Cy dye-conjugated dUTP, an aminoallyl dUTP is used in the reverse transcription reaction, followed by chemical coupling of the aminoallyl-labeled cDNA with an NHS ester of the Cy dye. There is less global dye bias between the two labeled targets because both



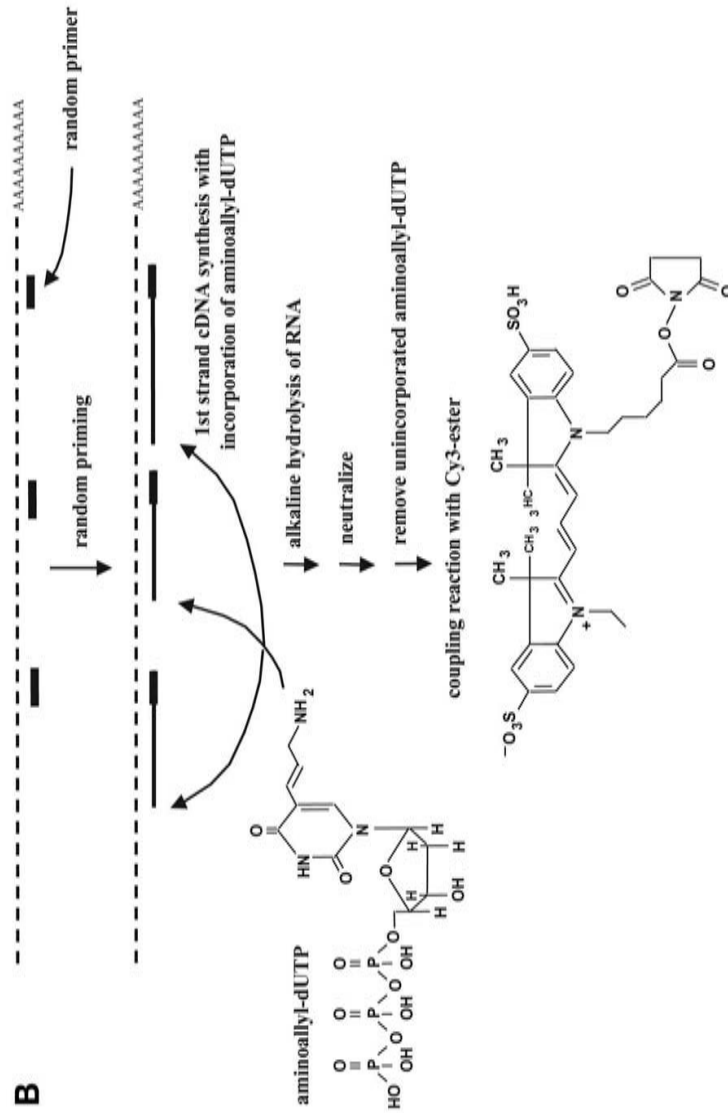


Fig. 2. (A) Direct labeling of cDNA target for hybridization onto two-color arrays. In separate reaction tubes, RNA samples A and B are reverse transcribed into cDNA that has been labeled with Cy3 and Cy5, respectively. Shown is the reaction for sample A. Cy3- and Cy5-labeled targets are co-hybridized onto the microarray. Bulky Cy dye coupled dUTP is inefficiently incorporated into first-strand cDNA during the reverse transcriptase reaction, leading to chain termination and low Cy dye specific activity. (B) Indirect labeling of cDNA target for hybridization onto two-color arrays. Aminoallyl-coupled dUTP is more efficiently incorporated into first-strand cDNA by reverse transcriptase. Random hexamer priming of mRNA for first-strand cDNA synthesis is depicted, but oligo-dT priming works as well.

RNA samples are reverse transcribed into cDNA using aminoallyl dUTP, and the labeling of one cDNA target with the Cy3 ester and the second cDNA target with the Cy5 ester involves a chemical and not an enzymatic reaction. The density of labeling with the indirect method can be as high as approx 1 fluorophore per 6 nucleotides, compared with 1 fluorophore for every 20 to 50 nucleotides in the direct method.

3.3. Signal Amplification for Two-Color Arrays

In the labeling strategies described thus far, the amount of total RNA required typically ranges between 10 to 15 μg . However, what happens if there is less RNA available? One possibility is to pool RNA samples from multiple biological replicates, and the relative merits and statistical implications of such an approach have been described (22). If pooling is not a viable option, there are a number of alternative labeling strategies specifically designed to handle total RNA yields ranging from 0.5 to 2 μg . In the case of the Affymetrix GeneChip, the Eberwine T7 RNA polymerase method can readily handle this amount of starting RNA because it is, in essence, a linear RNA amplification scheme (19). For long oligomer and PCR amplicon-based cDNA arrays, the tyramide signal amplification and dendrimer schemes are available. In the tyramide signal amplification method MICROMAX TSA Labeling and Detection Kit (Perkin Elmer Life Sciences, Boston, MA), biotin-conjugated dNTPs (used as a hapten for subsequent Cy5 labeling) and fluorescein-conjugated dNTPs (used as a hapten for subsequent Cy3 labeling) are incorporated in the first-strand cDNA synthesis reaction of samples A and B, respectively (Fig. 3A). The two hapten-coupled cDNA targets are co-hybridized onto the microarray overnight. After stringent washing, the microarray is incubated with an antibody against fluorescein, conjugated to horseradish peroxidase (HRP), which binds specifically to fluorescein-labeled cDNA target and enzymatically catalyzes the deposition of Cy3-labeled tyramide. After HRP inactivation, streptavidin-HRP binds to biotin-labeled cDNA and catalyzes the deposition of Cy5-labeled tyramide. The resulting 100-fold signal amplification relative to the direct-labeling protocol allows for the use of much less RNA starting material, although the major disadvantage is the high labor cost.

In the signal amplification method using the dendrimer technology (also referred to as the three-dimensional [3D] multilabeled structure protocol; ref. 23), first-strand cDNA synthesis is performed using an oligo-dT primer with a short “dendrimer capture sequence” at the 5'-end (Fig. 3A). After cDNA synthesis, a dendrimer (oligonucleotides crosslinked into a 3D structure) containing approx 300 Cy3 fluorophores is annealed to the cDNA via a complimentary sequence to the dendrimer capture sequence found on the 5'-end of each cDNA molecule. cDNA from a second sample is coupled with Cy5-dendrimers and

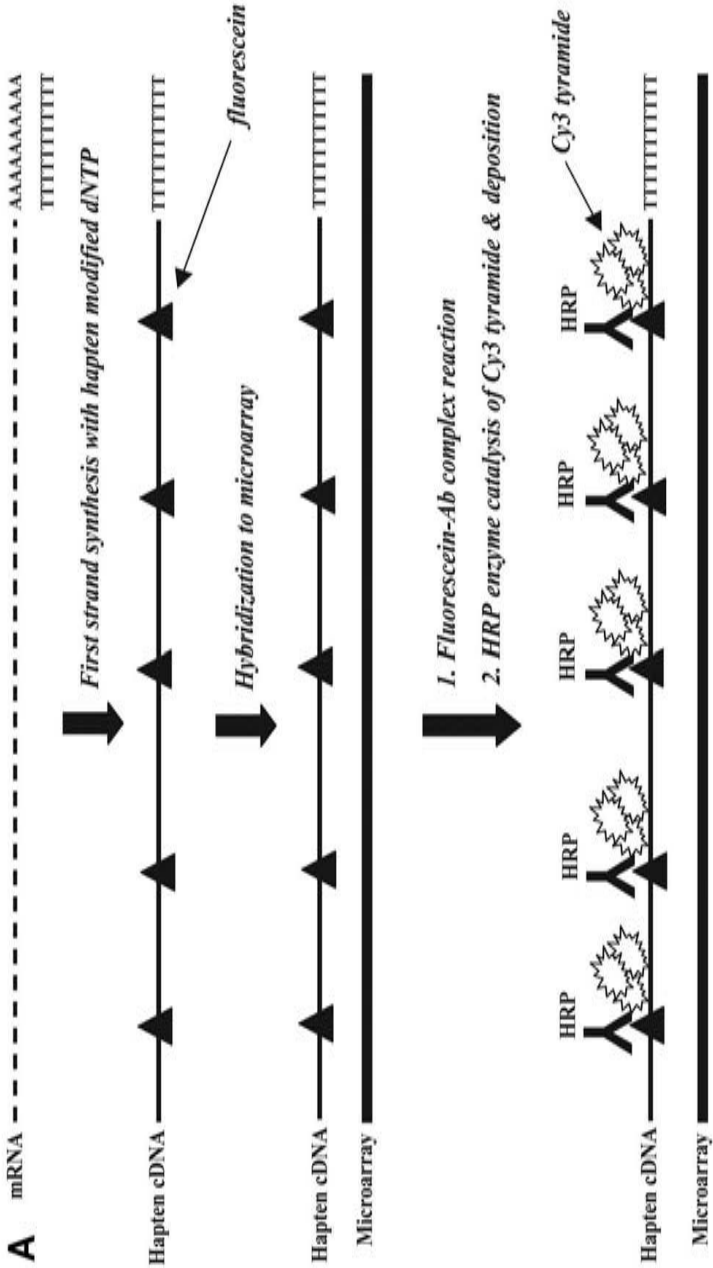


Fig. 3. (Continued)

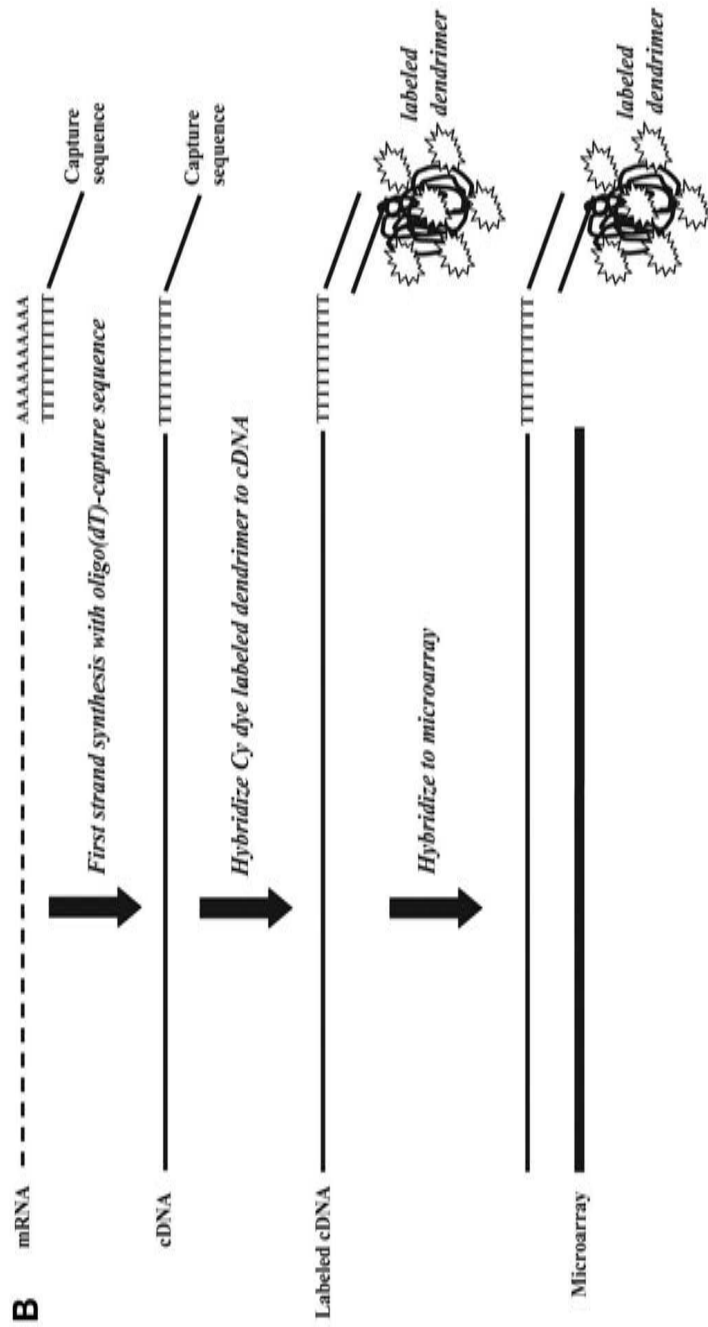


Fig. 3. (A) Tyramide signal amplification scheme for two-color arrays. In separate reaction tubes, RNA samples A and B are reverse transcribed into cDNA that has been labeled with fluorescein and biotin, respectively. Shown is the reaction for sample A. Fluorescein- and biotin-labeled cDNA targets are co-hybridized onto the microarray. The array is treated sequentially with an antibody (Ab) to fluorescein, conjugated with HRP and Cy3 tyramide, followed by streptavidin-HRP and Cy5 tyramide. (B) Dendrimer signal amplification scheme for two-color arrays. In separate reaction tubes, RNA samples A and B are primed with an oligo-dT primer with capture sequence, and reverse transcribed into the cDNA target. Cy3 and Cy5 dendrimers are annealed to the sample A and B cDNAs, respectively, and co-hybridized onto a microarray.

co-hybridized with the cDNA coupled to Cy3-dendrimers. By using fluorescent dendrimers, a 10- to 100-fold signal enhancement relative to the direct-labeling method is achieved.

3.4. RNA Amplification for Affymetrix and Two-Color Arrays

But what happens when only a very limited amount of RNA (10–100 ng and less) is available, as in the case of a laser capture microdissected sample? In this scenario, many laboratories have relied on the Eberwine T7 RNA polymerase amplification approach or modifications thereof. The first round of amplification yields an approx 1000-fold increase of the original amount of starting messenger RNA (mRNA), whereas two rounds will yield an approx 10,000-fold enrichment (24). For PCR amplicon-based microarrays, the aRNA resulting from the first round of amplification is not used for direct hybridization onto the array, but, instead, is used as a template in the indirect-labeling protocol for the generation of fluorophore-containing cDNA target (Fig. 4). If a second round of amplification is necessary, the aRNA from the first round is converted to first-strand cDNA using random hexamer primers (Fig. 4). Next, the first-strand cDNA is primed with the oligo-dT–T7 primer for generation of the second-strand cDNA. The resulting double-stranded cDNA template with T7 promoter is used to run-off a second round of aRNA, which serves as the template for the generation of fluorophore-labeled cDNA via the indirect-labeling approach. For the Affymetrix GeneChip, the user bypasses the direct/indirect-labeling step.

It is important to note that this particular RNA amplification scheme can only be used for Affymetrix GeneChip and PCR amplicon-based arrays and is not applicable to the long oligomer arrays. The orientation of the probes in the long oligomer microarray is in the sense direction, as is the case of the labeled cDNA target derived from standard Eberwine RNA amplification coupled with the indirect-labeling approach. Hence, a modification of the Eberwine method is necessary if first-strand cDNA synthesis is initiated with an oligo-dT primer (no T7 RNA promoter sequence); second-strand cDNA is primed with random nonamers containing a T7 RNA promoter sequence appended to the 5'-end. The resulting double-stranded cDNA is used in a one-round RNA amplification reaction to synthesize sense-strand RNA, which serves as the template for generation of fluorophore-labeled (–) cDNA target. If two or more rounds of amplification are required, the T3 RNA polymerase amplification scheme described by Xiang et al. (25) is a viable option (Fig. 5). Lastly, these so-called “linear” RNA amplification schemes can still introduce biases in the amplified product, especially as the number of rounds increases; hence, important safeguards and quality control assurances must be in place to ensure fidelity of the gene expression measurements (24).

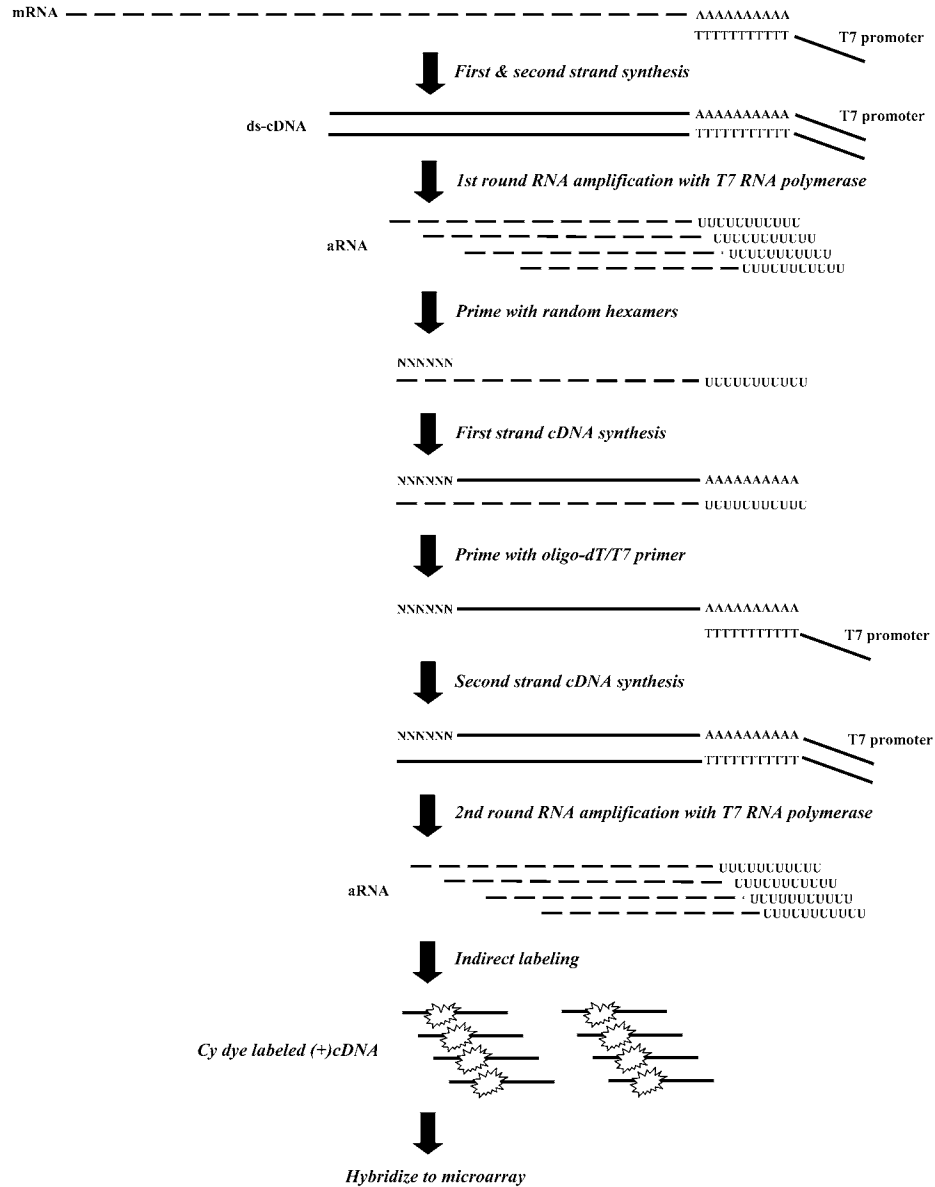


Fig. 4. Two rounds of T7 RNA polymerase amplification for extremely low amounts of starting RNA. Biotinylated aRNA generated from the second round of amplification can be hybridized onto Affymetrix GeneChips, or nonbiotinylated aRNA from the second round can be converted into Cy dye-labeled cDNA target via the indirect-labeling method and hybridized onto two-color arrays. Note that the orientation of the cDNA is sense (+) and, hence, can only be hybridized onto PCR amplicon-based arrays and not long oligomer arrays.

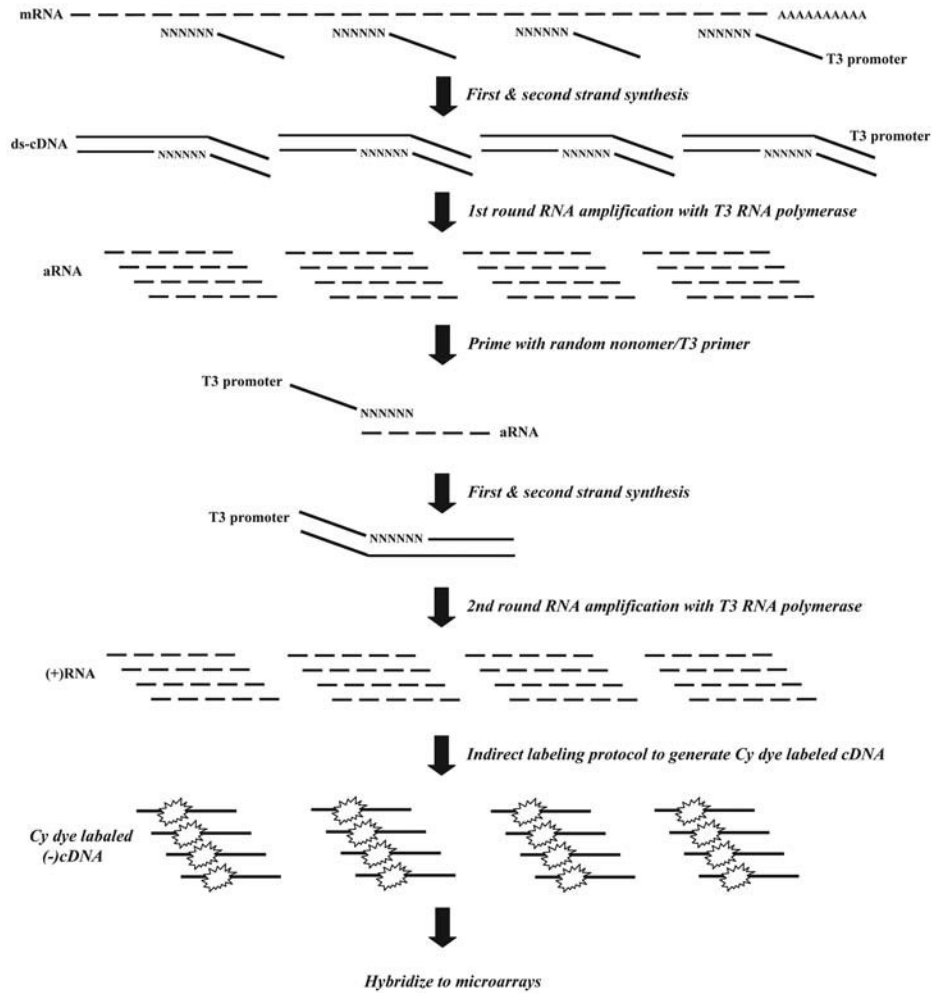


Fig. 5. RNA amplification strategy to generate Cy dye-labeled (-) cDNA target for hybridization onto long oligomer arrays.

4. Study Design and Objectives

4.1. Sample Allocation Overview

The abundance of gene expression data in the literature has sometimes fostered a false sense that microarray data can be collected in a relatively spontaneous or unplanned manner. This could not be further from the truth. Because PCR amplicon and long oligomer microarrays both use a two-color system, an investigator has to decide how to allocate samples to labels and to microarrays. With the Affymetrix GeneChip, RNA samples are individually labeled and

hybridized to individual chips, thereby obviating the need to contend with sample allocation issues. There have been a number of recent reviews discussing sample allocation strategies for microarrays using the two-color system (26–29). The most commonly used of the design strategies can be broken down into four major categories: direct comparison with dye-swap, balanced block design, reference sample design, and loop design. Although more extravagant and complicated schemes have been put forward, these four design strategies are conceptually straightforward and easier to implement, and, in the majority of instances, have sufficient statistical power to identify differentially regulated genes.

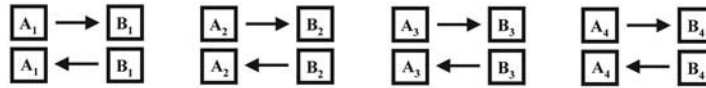
4.2. Direct Comparison Design

The direct comparison is favored if an investigator wishes only to compare the relative gene expression of two classes of samples or experimental groups (Fig. 6). The groups being compared can be untreated and drug-treated cell lines, gene-knockout mouse strain and wild-type mouse strain, chromosome-substituted rat strain and parental rat strain, and so on. On each microarray, RNA samples from the two experimental groups are labeled and co-hybridized onto the same gene chip. The major advantage of the direct comparison design can be found in its name, that is to say, “differential expression of genes in samples A and B is more efficiently measured (and hence more accurate) when comparisons are made on the same array.” This is in contrast to indirect approaches, such as the reference and loop design (see their description in Subheadings 4.4. and 4.5.), in which differential gene expression in the two experimental groups must be inferred because interrogations are performed on separate microarrays, leading to greater variance in the determinations (30). If using the direct-labeling approach (and to a lesser extent with the indirect-labeling approach) we, as well as others, have noted that a small proportion of mRNA species will label with one dye preferentially over the other, leading to a gene-specific dye bias effect. This phenomenon persists even after data normalization to account for global dye biases (i.e., less efficient incorporation of Cy5 vs Cy3 during direct labeling); the reasons for this gene-specific bias are not known, but may be related to distinct physiochemical properties of individual mRNA species. To account for gene-specific dye bias effects, a dye-swap (also referred to as flip-dye or dye reversal) hybridization is performed (Fig. 7). This replicate hybridization, although requiring more starting material RNA, serves two important functions:

1. Identifies gene-specific dye bias artifacts and allows the investigator to exclude such genes from further downstream analysis.
2. Increases the precision of the gene expression measurement because dye-swap hybridizations are akin to performing a technical replicate.

The direct comparison design with dye-swap hybridization uses n arrays for n samples in which each sample is divided into two aliquots.

Direct Comparison with Dye Swap:



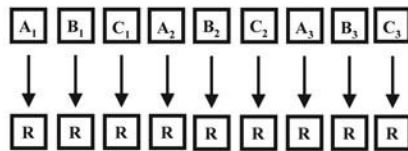
- RNA sample is not limiting
- Flip dyes account for any gene-specific dye bias effects

Balanced Block Design:



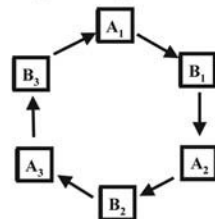
- RNA sample is limiting
- Balanced blocking accounts for any gene-specific dye bias effects

Reference Design (Indirect Comparison):



- More than two samples are compared (e.g. time course; tumor classification)
- Flip dyes are not necessary but can be done to increase precision
- Ratio values are inferred (indirect)
 $(A_1/R_1)/B_1/R_1 = A_1/B_1$
- Requires common RNA reference for all hybs

Loop Design:



- Suitable for 2 or more experimental groups
- Flip dyes are not necessary but can be done to increase precision
- Loop design ‘unravels’ with a single poor hybridization
- ANOVA analysis is performed to estimate \log_2 ratios

Fig. 6. RNA sample allocation strategies for two-color microarrays. RNA samples are represented as boxes. Each arrow represents a single hybridization assay (hyb), with the “head” of each arrow representing Cy5 labeling and the “tail” representing Cy3 labeling. Experimental groups are indicated by letters in boxes and biological replicates are indicated by numbered subscripts. For example, there are two experimental groups A and B and each experimental group is comprised of four biological replicates.

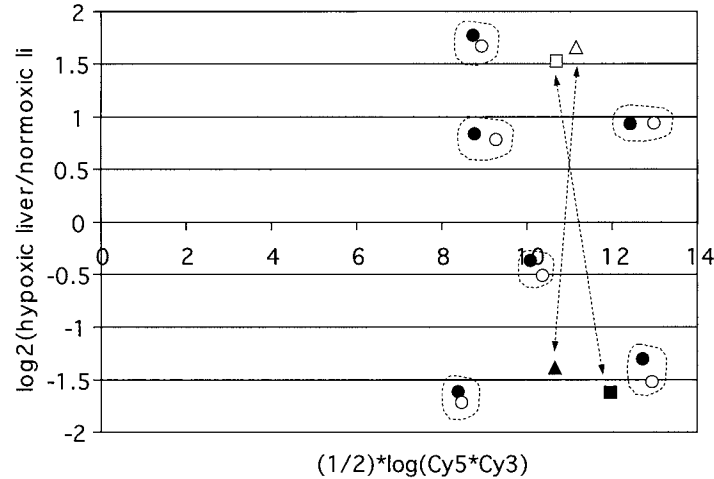


Fig. 7. R-I plot revealing gene-specific dye labeling bias. Three hybridizations (biological replicates) were performed comparing the livers from hypoxic and normoxic Dahl Salt-Sensitive rats. A direct comparison (hypoxic vs normoxic) with dye-swap hybridization strategy was implemented, and the averaged results for eight genes are plotted. For six genes, plotted as open circles representing “forward” hybridizations and closed circles representing dye-swap hybridizations (labeling in the opposite direction), there was no evidence for gene-specific dye labeling bias. This was the case for the majority of genes surveyed. However, the plots of two genes are shown (triangle and square), exhibiting gene-specific dye labeling bias. Forward hybridizations indicated that these two genes were upregulated in hypoxic liver, but dye-swap hybridizations showed the opposite trend.

4.3. Balanced Block Design

If a direct comparison-type strategy is warranted, but the RNA samples are limiting (and the investigator wishes to avoid RNA amplification), and there are concerns of gene-specific dye bias, then a balanced block design is a good alternative (Fig. 6). Here, the direction of the labeling is switched back and forth among a prescribed number of hybridizations; each hybridization represents a biological replicate. In this scheme, mRNAs suffering from labeling bias with a particular fluorophore can be accounted for without resorting to technical replicates. This design uses $n/2$ arrays for n samples.

4.4. Reference Sample Design

If more than two classes of samples or experimental groups are being compared, such as a time-course experiment or classifying multiple tumor types, it is usually more efficient to design array experiments in which each experimental

group is compared back to a common reference RNA sample (**Fig. 6**). The efficiency of using a reference design becomes readily apparent as more and more experimental groups are added, resulting, if one chooses a direct comparison design, in a bewildering number of hybridizations (e.g., experimental group A vs experimental group B, B vs C, C vs D, D vs E, A vs C, A vs D, A vs E, B vs D, and so on), which becomes too cumbersome, cost prohibitive, and/or RNA sample limiting. There is an additional advantage to the application of the reference design, the presence of any gene-specific dye bias affects all arrays similarly and, hence, does not confound experimental group comparisons (29). As a result, dye-swap experiments are not necessary, but can be included to increase the precision of the gene expression measurements. In the absence of dye-swap experiments, the reference design uses n arrays for n nonreference samples; or if dye-swap hybridizations are included, $2n$ arrays for each nonreference sample (two aliquots are required for each nonreference group). The reference design is particularly well-suited for class discovery using cluster analysis, but any relative gene expression comparisons between groups must be inferred, leading to greater variance in expression measurements. For any gene, the relative expression measurement $\log_2(A/B)$ is estimated by taking the difference $\log_2(A/R) - \log_2(B/R)$, where A, B, and R are experimental group A, experimental group B, and reference, respectively. The reference sample RNA may be biologically relevant, such as a time-zero sample from a time-course experiment, but the reference need not have any biological relevance whatsoever. In this case, the reference RNA can be derived from a mixture of RNA from multiple cell lines or tissues. It is only important that the reference RNA sample is able to hybridize to the majority of probes on the microarray (typically >90% of the probes have a positive signal), and is available in a sufficient quantity to cover all microarray experiments, because even different batches of reference RNA may have quite different expression profiles. Hence, care must be taken in planning experiments and estimating the amount of reference RNA required. Interestingly, genomic DNA has recently been advanced as a better alternative, because this nucleic acid is not subject to biological variance associated with different batches of RNA isolations, and, as a result, represents an “inexhaustible reference source” (31).

4.5. Loop Design

In the loop design approach, which may serve as an alternative strategy to the reference design, samples representing two or more experimental groups are compared with one another in a “head-to-tail” fashion, resulting in the formation of a loop (**Fig. 6**; **ref. 26**). This design uses n arrays for n samples, using two aliquots of each sample. By using this configuration, gene-specific dye bias effects are accounted for because each RNA sample in the loop is used once as

the head (i.e., labeled with Cy3) in one hybridization, and as the tail (i.e., labeled with Cy5) in a second hybridization. Analysis of variance (ANOVA) techniques have been developed that allow \log_2 ratio values (relative expression values between two samples) to be estimated for each sample comparison (26). A drawback to this approach can be observed if samples number four or more in a loop, namely, gene expression comparisons between samples not directly connected to each other must be inferred (27). Moreover, samples at opposite ends of the loop require the greatest inferences, resulting in the least accurate gene expression measurements. This becomes a very distinct disadvantage as loops become larger and larger. Lastly, the loop design is less robust against the presence of poor-quality hybridizations in which a single bad array can unravel the loop.

4.6. Defining the Number of Biological Replicates Needed

A common question raised by investigators, regardless of the microarray platform used, is “How many biological replicates are needed?” The number of independent biological replicates needed depends on such factors as the objectives of the experiment and the inherent noise of the biological system. Because gene expression measurements from microarrays can be rather variable, it is important to have some type of assurance that our determinations are not false positives. It is important to distinguish biological replication from technical replication. The dye-swap hybridizations represent a form of technical replication, whereby the precision of our measurements is increased by repeated hybridizations with the same RNA samples. By comparison, biological replication is essential to draw conclusions that are valid beyond the scope of the tested samples (e.g., is there a statistically significant difference between treatments?). To estimate the sample size required to achieve the aims of our study, a power calculation is applied. It takes into account the variance of individual measurements, the acceptable false-positive rate, and the desired discriminatory power of the microarray. Simon and Dobbin (29) described a relatively simple power calculation that can be applied to both two-color microarrays using a reference design and single-color Affymetrix GeneChips if comparing two experimental groups/classes. This approach assumes that the gene-specific expression measurements (e.g., \log_2 values) are approximately normally distributed for each class. We let σ denote the standard deviation of the log expression level among samples within each class, and suppose that the means of the two classes differ by δ for a particular gene. For \log_2 values, a $\delta = 1$ would correspond to a twofold difference in gene expression between classes. We assume that the two classes are compared at the level of expression of each gene, and that a statistically significant difference occurs on rejection of the null hypothesis at a significance level α . Because thousands of genes are analyzed simultaneously on

an array, the significance level α can be set stringently to limit the number of false positives. The statistical power of our calculation is defined by $1 - \beta$, where β is the false-positive rate. Under these parameters, the approximate number of independent biological samples is n , where

$$n = 4(z_{\alpha/2} + z_{\beta})^2 / (\delta/\sigma)^2 \quad (1)$$

and $z_{\alpha/2}$ and z_{β} denote the corresponding percentiles of the standard normal distribution (32). It has been suggested that a good general guideline is to choose $\alpha = 0.001$ and $\beta = 0.05$ (29). For a 10,000-element array, $\alpha = 0.001$ results in an average of 10 false-positive genes and $\beta = 0.05$ provides a 95% probability of detecting a significant change in gene expression. Using $\alpha = 0.001$ ($z_{\alpha/2} = 3.29$), $\beta = 0.05$ ($z_{\beta} = 1.645$), $\delta = 1$, and $\sigma = 0.35$ in **Eq. 1**, we find that a total of 12 samples, 6 for each of the two classes, are required for comparing the two classes and identifying genes exhibiting a significant twofold change. For one-color Affymetrix GeneChips, n = number of arrays; and for the two-color format, $n/2$ = number of arrays.

A second simple approach to estimate the adequacy of the number of biological replicates in a microarray experiment is based on determining the degrees of freedom (27). This can be determined by counting the number of independent biological replicates (e.g., independent animals, independent cell line cultures, or independent pools of microdissected tissues) and subtracting the number of distinct treatments from the number of independent biological replicates. If $df = 0$, there may be no information available to estimate the biological variance, and, hence, the scope of one's conclusions will be limited to the samples themselves. A good guide at the experimental design stage is to have $df = 5$ or greater (27).

5. Systematic Assessment of Microarray Performance

How do we assess the performance of our microarrays? This is an especially relevant question to the novice beginning their first microarray hybridization, and to the experienced user interested in testing a new labeling protocol, RNA extraction method, or RNA amplification scheme. An approach to monitor microarray performance that is gaining widespread popularity is the adoption of external RNA controls, also referred to as spike-in controls or exogenous controls (5,14,33–36). External controls help to identify systematic problems associated with target labeling, array hybridization, and scanning. Typically, external controls are RNA molecules that are synthetically manufactured by *in vitro* transcription. The essential feature of external RNA controls is that the user can introduce predefined amounts to the biological RNA sample. Several external controls are recommended to cover a broad range of expression levels (1–5 copies per cell for rare transcripts, and ~100–300 copies per cell

for moderately expressed transcripts in mammalian organisms). If they are spiked differentially (e.g., two- and threefold differences) into the two RNA samples that are being compared, the external controls mimic differentially expressed genes. Hence, the external controls provide an important benchmark for quality control assessment, and, in many laboratories, the external controls are routinely used in all microarray experiments.

The critical requirement for external RNA controls is that they are representative of the endogenous biological mRNAs in terms of length and sequence characteristics (e.g., GC content, and secondary structure). In addition, cross-hybridization toward the endogenous transcripts should be avoided. Hence, for example, plant-specific RNA external controls can be used when interrogating mammalian RNA samples (14). In terms of the microarrays, the probe sets that recognize the fluorophore-labeled external targets are frequently printed across different sectors of the microarray glass slide, thereby allowing assessment of intraslide variability, whereas interslide variability is assessed across multiple independent hybridizations and target labeling. The external probe elements can also serve as negative controls if the external RNA is not added to the labeling reaction. In the absence of external RNA spiking, nonspecific hybridization of fluorophore-labeled sample target should be negligible, otherwise, it may be an indication that wash conditions are not sufficiently stringent.

6. Identifying Differentially Regulated Genes

The study of gene expression with microarrays has evolved from a qualitative endeavor during its early years to a more quantitative pursuit in more recent years. Statistical procedures for determining differentially regulated genes are just one aspect of this evolution. Even if data mining analysis is going to be performed using one or more of the widely used visualization tools (e.g., cluster analysis; *see Heading 7.*), it is frequently useful to reduce the data set to those genes that can best distinguish between the experimental groups. The earliest microarray papers typically used an *ad hoc* approach to define differentially regulated genes. For example, all genes exhibiting a twofold difference in expression (up or down) between experimental groups were deemed interesting, thus, ignoring biologically relevant genes exhibiting smaller changes. Furthermore, with this approach, there is no associated value that indicates the level of confidence in the designation of genes as differentially regulated.

The *t*-test is a simple, statistically based method for detecting differentially regulated genes (37). This statistical approach can be used for both Affymetrix GeneChips and two-color arrays using a reference or balanced block design. However, a drawback to using the *t*-test on microarray data is the resulting phenomenon known as the *multiple testing problem* (38). Consider a cut-off for differential expression of $p < 0.05$. We would expect 5% of the nondifferentially regulated genes on the array to reach “statistical significance” (false-positives).

Because large numbers of tests are being conducted on a single array, this is equivalent to saying that we expect 500 genes to be identified as significant on a 10,000-element array, when, in fact, they are not differentially regulated. To control for these false-positives resulting from the multiple testing problem, a Bonferroni correction is commonly implemented. The nominal false-positive rate is divided by the number of tests (in this case 10,000) to yield the effective rate. For a 10,000-element array, the Bonferroni-corrected p value is reduced to $p < \alpha/N$ array elements or, in our example, $p < 0.000005$. In practice, this correction is too severe, and typically leads to very few identified differentially regulated genes. There are, however, less conservative corrections that can be applied, including the *adjusted Bonferroni correction*, which ranks gene by their t statistic and then applies increasingly less stringent criteria to subsequent genes in the list until an appropriate threshold p value is reached. Alternatively, the Westfall and Young stepdown p values rely on permutation testing to select appropriate significance cutoffs. Both of these approaches, along with the more conservative Bonferroni technique, correct for multiple testing by controlling the familywise error rate, which is the probability of accumulating one or more false-positive errors over a number of statistical tests (37).

For two or more experimental groups assayed on Affymetrix GeneChips or two-color arrays, significance analysis of microarrays (SAM) is a popular approach to identify differentially regulated genes (39). SAM uses an adjusted t statistic along with permutation testing to estimate the false-discovery rate in any user-defined set of significant genes. Alternatively, ANOVA techniques have been described for microarray experiments assaying three or more experimental groups (37). Finally, a one-sample t -test with a multiple testing correction, or a variant, such as one-sample SAM, can be implemented for two-color arrays using a direct comparison or balanced block design. Designs of this type involve the co-hybridization of two experimental groups on the same array, and the primary question is whether the \log_2 expression ratio values are consistently significantly different from zero.

It is important to note that a good foundation in statistics is increasingly critical in microarray applications, but it is not a substitute for good experimental design. For example, in the absence of dye-reversal hybridizations to account for gene-specific dye bias effects, no amount of statistical gymnastics will rescue an investigator from potentially pursuing these false-positive genes in downstream functional analysis.

7. Visualizing Expression Data

7.1. Getting Started

The starting point in the analysis of expression data is the collection of raw expression measurements. For two-color arrays, these measurements are typically performed by image analysis software, such as TIGR Spotfinder (<http://www.tm4.org/spotfinder.html>) or ScanAlyze (<http://rana.lbl.gov/EisenSoftware.htm>),

that detects the fluorescence intensity of each fluorophore (Cy3 and Cy5) on each array spot. After taking into account and correcting for factors such as local background estimates and spot morphology, the software will output a pair of intensity values for each spot—an estimate of the expression level for both conditions in the hybridization.

Before any biologically relevant expression analysis can take place, these raw intensity values must first be normalized. Normalization algorithms can help to reduce the effects of systemic biases, such as differences in labeling efficiencies and spatial variation across the array, and to facilitate comparisons between data sets. Data filtering techniques are often applied near the normalization steps to reduce the complexity of the data set by removing data that are of questionable or poor quality. There are many normalization algorithms available, ranging from simple scaling techniques, such as total intensity normalization (12), to the advanced lowess normalization (40). Other algorithms exist to deal with the rationalization of dye-swap (41) and replicate data. Examples of available normalization packages include MIDAS (<http://www.tm4.org/midas.html>) and ArrayNorm (<http://genome.tugraz.at/Software/>).

7.2. Working With Expression Data

The relationship between the expression measurements for a particular array element in a two-color array can be summarized by the ratio of its intensity values. This expression ratio is calculated by dividing one intensity value (associated with one dye) for a given element by the other intensity value for that same element (associated with the second dye): intensity1 = 20,000 and intensity2 = 10,000.

$$\text{expression ratio} = \frac{\text{intensity1}}{\text{intensity2}} = \frac{20,000}{10,000} = 2.0$$

The community standard has been to use \log_2 ratios instead of basic (nonlog) ratios. Using \log_2 ratios to represent relative expression levels offers several advantages over basic expression ratios. Consider a fivefold change in expression: intensity1 = 50,000 and intensity2 = 10,000.

$$\text{basic ratio} = \frac{\text{intensity1}}{\text{intensity2}} = \frac{50,000}{10,000} = 5.0$$

$$\log \text{ ratio} = \log_2 (\text{basic ratio}) = \log_2(5.0) = 2.32$$

In the opposite case, intensity2 is five times larger than intensity1: intensity1 = 10,000 and intensity2 = 50,000.

$$\text{basic ratio} = \frac{\text{intensity1}}{\text{intensity2}} = \frac{10,000}{50,000} = 0.2$$

$$\log \text{ ratio} = \log_2 (\text{basic ratio}) = \log_2(0.2) = -2.32$$

A comparison of the basic ratios from these examples reveals two arithmetically accurate results that are reciprocals of each other. The basic ratios, 5.0 and 0.2, are asymmetrically distant from the basic ratio that represents a lack of expression change, 1.0. On the other hand, the corresponding \log_2 ratios, 2.32 and -2.32 , are equally distant from the \log_2 ratio that represents a lack of expression change, 0.0. The nature of the logarithm is such that an n -fold change in expression will result in a log ratio that is equal in magnitude to another n -fold change in the opposite “direction.” This trait makes comparisons involving log ratios more intuitive than their basic counterparts, because overexpressed and underexpressed elements are treated symmetrically. From this point on, expression values will be represented by log ratios.

The expression level of an element in a specific hybridization experiment can be summarized by its log expression ratio. One technique to compare the expression levels of an element across experiments is to examine the corresponding series of log ratios. For example, the expression of element A across four experiments (numbered 1–4) can be represented by the four ratios shown in **Table 1**.

This sequence of log ratios is known as an *expression vector*. It represents the expression of an element across multiple experiments. The expression vector can serve as a profile of the specified array element; such profiles are necessary if determining the similarity of expression levels between multiple elements. It is also possible to generate expression vectors to represent the profiles of experiments instead of array elements. Many clustering and classification techniques will operate on expression vectors of experiments as well as vectors of elements.

A natural extension in working with expression vectors is to evaluate, in tandem, those vectors that cover the same series of experiments. “Stacking” such expression vectors produces a structure known as an *expression matrix*. In the example in **Table 2**, note the addition of the expression vectors representing the expression levels of elements B, C, and D across the same set of experiments as the expression vector corresponding to element A. Each intersection of an element and an experiment is a matrix cell that contains a log ratio; this value represents the expression of the specified element in the specified experiment.

One way to think about an expression vector is as a series of Cartesian coordinates that define an element’s location in n -dimensional expression space, where n is equal to the number of experiments in the vector. In the example in **Table 3**, there are four expression vectors, each with three data points (i.e., \log_2 ratios). Each of these vectors can be represented as a triad of Cartesian coordinates: element A = (3.0, 4.0, 5.0); element B = (2.0, 3.0, 4.5); and element C = (–3.0, –2.0, 3.0).

Table 1
Expression of Element A Across Four Experiments

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Element A	Log ratio A1	Log ratio A2	Log ratio A3	Log ratio A4

Table 2
Expression Matrix

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Element A	Log ratio A1	Log ratio A2	Log ratio A3	Log ratio A4
Element B	Log ratio B1	Log ratio B2	Log ratio B3	Log ratio B4
Element C	Log ratio C1	Log ratio C2	Log ratio C3	Log ratio C4
Element D	Log ratio D1	Log ratio D2	Log ratio D3	Log ratio D4

Table 3
Expression Vectors

	Experiment 1	Experiment 2	Experiment 3
Element A	3.0	4.0	5.0
Element B	2.0	3.0	4.5
Element C	-3.0	-2.0	3.0

With these sets of coordinates, each vector can now be plotted on a 3D graph (Fig. 8). Note that elements A and B have similar log ratios in each of the three experiments, both in terms of magnitude and sign. These two elements appear near each other on the graph, but substantially further from element C, whose log ratios are less similar to those of elements A and B. Mathematical formulas called *distance metrics* will be used (see **Subheading 7.3.**) to quantify these observations and to facilitate the analysis of expression vector relationships (i.e., clustering).

7.3. Clustering: An Overview

One branch of microarray data analysis is the exploration of the expression patterns that arise for array elements within a series of experiments. Identifying array elements with similar expression patterns may provide evidence of a biological relationship between the represented genes. The use of clustering algorithms is a common method of evaluating these patterns of expression and organizing related elements.

Clustering algorithms can be divided into a few functional categories. Agglomerative methods, such as hierarchical clustering (42,43), start with

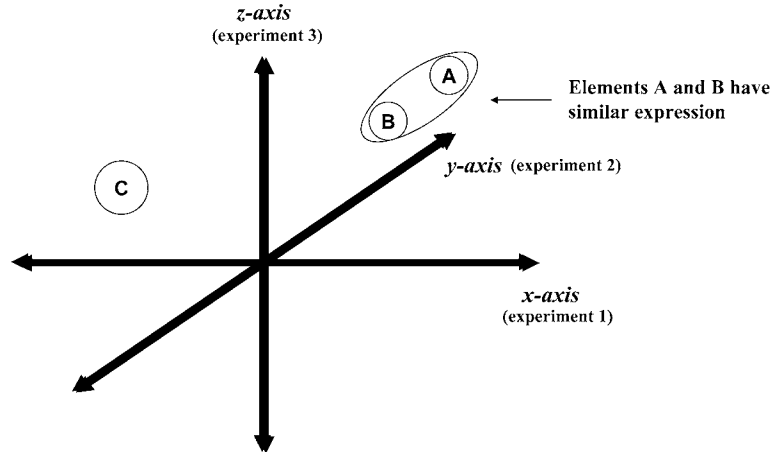


Fig. 8. Expression vectors as points in a three-dimensional (3D) expression space. Microarray data mining involves looking for genes with “similar” patterns of expression. If three hybridization experiments are considered, the expression vector for each gene is represented by a point in 3D space, where the expression measure (\log_2 ratio value) for gene i in experiment 1 is its x coordinate, the expression measure for gene i in experiment 2 is its y coordinate, and the expression measure for gene i in experiment 3 is its z coordinate. In such a geometric representation, expression vectors for gene elements A and B have similar expression patterns.

individual elements and iteratively build up larger structures by associating similar elements with each other. Divisive methods, such as k -means clustering (44,45), seek to take a large collection of elements and segregate them into groups containing elements with similar expression patterns. Other algorithms transform the input expression matrix to facilitate user-defined element groupings or may incorporate approaches from multiple algorithm categories. The hierarchical clustering and k -means clustering algorithms are described in **Subheadings 7.4** and **7.5**.

Clustering algorithms rely on mathematical formulations to determine how similar elements are to each other. The terms similarity and distance are inversely related; two elements are considered similar if the distance between their expression vectors is low. Conversely, a larger distance between a pair of expression vectors indicates a lower level of similarity between the associated elements. It is this measured distance between expression vectors that is used when decisions are made to cluster elements.

There are many methods available to measure the distance between expression vectors; these are collectively known as *distance metrics*. Each distance metric uses a formula that can take two expression vectors and compute a numeric distance measurement. Some clustering algorithms were designed with

a particular distance metric in mind, whereas others are compatible with several metrics. The selection of the distance metric to use is an important decision, because each metric is capable of uncovering different features of the data set.

Two common distance metrics are Euclidean distance and centered Pearson correlation coefficient. Euclidean distance is based on the two-dimensional Pythagorean theorem:

$$d_{A:B} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

where $d_{A:B}$ is the distance between points A and B, point A is defined by the coordinates (x_1, y_1) and point B is defined by the coordinates (x_2, y_2) . In the Euclidean distance metric, each experiment in the expression vector is treated as a dimension. If the expression vectors each contain two ratios (i.e., there are two experiments) then the distance formula could be written in a form similar to **Eq. 2**. The distance between two expression vectors, each containing n ratios, can be calculated using the general equation:

$$d_{A:B} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where $d_{A:B}$ is the distance between expression vectors A and B, x_i is the log ratio from expression vector A and y_i is the log ratio from expression vector B, both from the experiment at position i . The Euclidean distance metric exhibits a commutative behavior; the distance between expression vectors A and B is equal to the distance between expression vectors B and A. The smallest distance possible is the distance between an expression vector and itself, 0. There is not a defined upper limit for distance.

To measure the similarity of the shapes of two expression vectors, a centered Pearson correlation coefficient distance metric is used. The shape of an expression vector is most apparent by graphing experiments on the x -axis, and component log ratio values on the y -axis (**Fig. 9A**). The value of the centered Pearson correlation coefficient, r , for two expression vectors each containing n log ratios, is calculated as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where x_i is the log ratio from expression vector A, and y_i is the log ratio from expression vector B, both for the experiment at position i , \bar{x} is the mean log ratio from expression vector A, and \bar{y} is the mean log ratio from expression vector B.

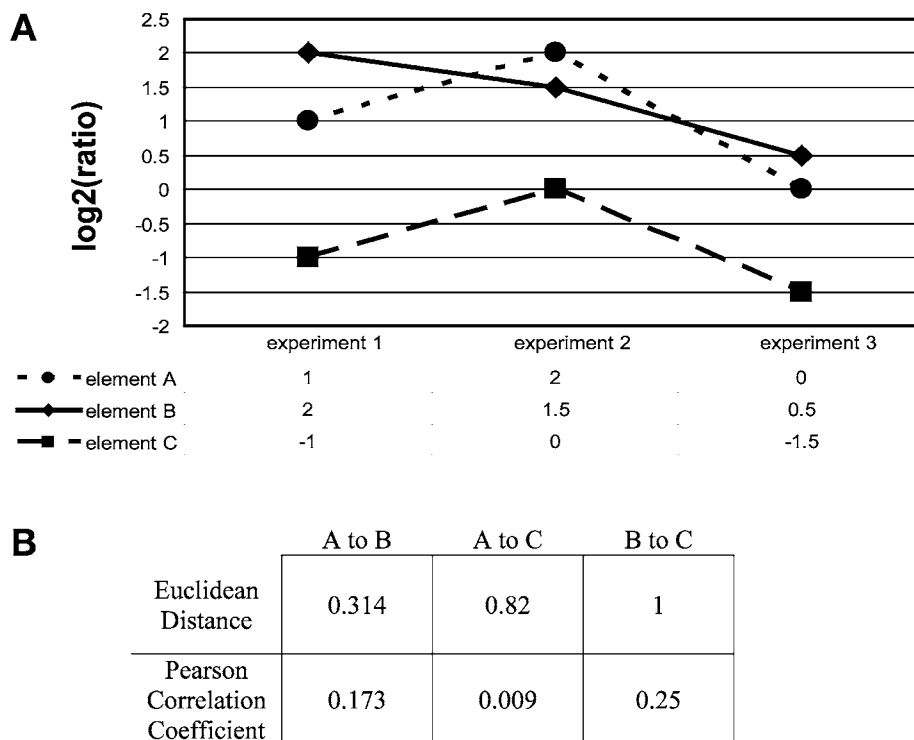


Fig. 9. (A) The expression vectors A, B, and C are shown in both tabular and graphical form. (B) Distances between each of the vectors were calculated using Euclidean distance and Pearson Correlation Coefficient. The distances have been scaled such that the minimum distance is 0 and the maximum distance is 1. Note that the most similar vectors are A and B if using the Euclidean Distance metric, whereas the Pearson Correlation Coefficient metric shows A and C to be most similar.

Values for r range from -1 to 1 . The magnitude of the r -value indicates the strength of the correlation, and the sign indicates whether the correlation is direct or inverse. Expression vectors with strong direct correlation (i.e., similar shapes) will have an r -value close to 1 . In the case of vectors with a strong inverse correlation (i.e., opposite shapes), the r -value will be close to -1 . A pair of expression vectors with weak correlation (i.e., neither directly nor inversely correlated) and independent shapes will have an r -value close to 0 . Two other forms of the Pearson correlation coefficient are also widely used. An uncentered version takes into account the magnitude of expression changes within each vector when calculating r . The Pearson squared form treats pairs of correlated vectors in the same manner as anticorrelated vectors.

Similar to the Euclidean distance metric, the centered Pearson correlation coefficient also exhibits commutative behavior. A comparison of these two metrics is illustrated in **Fig. 9B**.

7.4. Hierarchical Clustering

Hierarchical clustering is an agglomerative clustering method that offers an intuitive visual result in the form of a tree diagram and provides insight into the degree of relationship that elements have with each other. The algorithm takes a collection of independent elements and progressively joins them into increasingly larger clusters.

The preliminary step in creating a hierarchical tree is the calculation of the pairwise distances between every element and every other element to determine which elements are most closely related. A *distance matrix* can be constructed to store all of the calculated distance values as an $n \times n$ grid, where n is the number of elements involved in the analysis. Each row and column represent an element and the matrix cells contain the pairwise distance between the row element and the column element, each calculated using the same distance metric (**Table 4**).

If the distance metric used has a commutative behavior (i.e., the distance from element A to element B is equal to the distance between element B and element A) then the distance matrix will be symmetrical about the diagonal (upper left to lower right). From a computational perspective, this reduces the total number of distance calculations by approximately half.

Once the distance matrix has been constructed, the algorithm will enter an iterative stage in which the following steps will be performed a number of times equal to $n - 1$. At the start, each element in the distance matrix is treated as a “cluster.” As the algorithm progresses, these single-element clusters will be combined to form progressively larger nested clusters.

The steps in the algorithm are:

1. Determine which two clusters are the most similar by finding the smallest distance value from the distance matrix.
2. Combine these two clusters together to form a larger cluster.
3. Recalculate only the distances between this cluster and all other clusters. A predetermined *linkage method* will dictate the procedure to use when calculating the distance between clusters that contain more than one element.
4. Continue with the next iteration at **step 1**. Look for the next smallest distance value from the distance matrix.

There are several linkage methods to choose from in deciding how to measure distances between clusters. Consider two clusters, A and B, each with five member elements. The *single linkage* method sets the distance between clusters

Table 4
Distance Matrix

	Element A	Element B	Element C	Element D
Element A	Distance AA	Distance AB	Distance AC	Distance AD
Element B	Distance BA	Distance BB	Distance BC	Distance BD
Element C	Distance CA	Distance CB	Distance CC	Distance CD
Element D	Distance DA	Distance DB	Distance DC	Distance DD

A and B to equal the smallest distance between any element contained in cluster A and any element contained in cluster B. The opposite approach, *complete linkage*, uses the largest distance between any element contained in cluster A and any element contained in cluster B as the intercluster distance. *Average linkage* calculates the average distance between elements in both clusters and sets this value as the intercluster distance.

The result of this algorithm is a series of progressively larger nested clusters, and a table of relevant intercluster distances. It is a relatively straightforward task to use this data to create a graphical depiction, known as a dendrogram (Fig. 10). The clusters that were joined together in the earliest iterations are connected by short branches (i.e., elements with the most similar patterns of expression), whereas clusters that were joined later are connected by increasingly longer branches (i.e., less similar). A color scheme is applied to the dendrogram to provide an intuitive representation of overexpressed and underexpressed genes. A color gradient running from black to red represents log ratios from zero to a positive end-point value, respectively, along with a color gradient running from black to green to represent log ratios from zero to a negative end-point value, respectively.

7.5. k-Means Clustering

If there is an *a priori* hypothesis regarding the number of clusters into which the elements in the data set should be partitioned, the divisive *k*-means clustering method can be used to perform the partitioning. The goal of the algorithm is to divide the elements into *k* distinct clusters; each cluster should end up containing elements that are more similar to each other than to elements in other clusters. The value for *k* must be set by the user before the start of the algorithm.

The *k*-means algorithm consists of the following steps:

1. Each element is assigned randomly to one of the *k* clusters.
2. An expression vector is used to represent each cluster by computing the mean expression vector of all elements in that cluster. If the median expression vector is used instead, this method is called *k*-medians clustering.

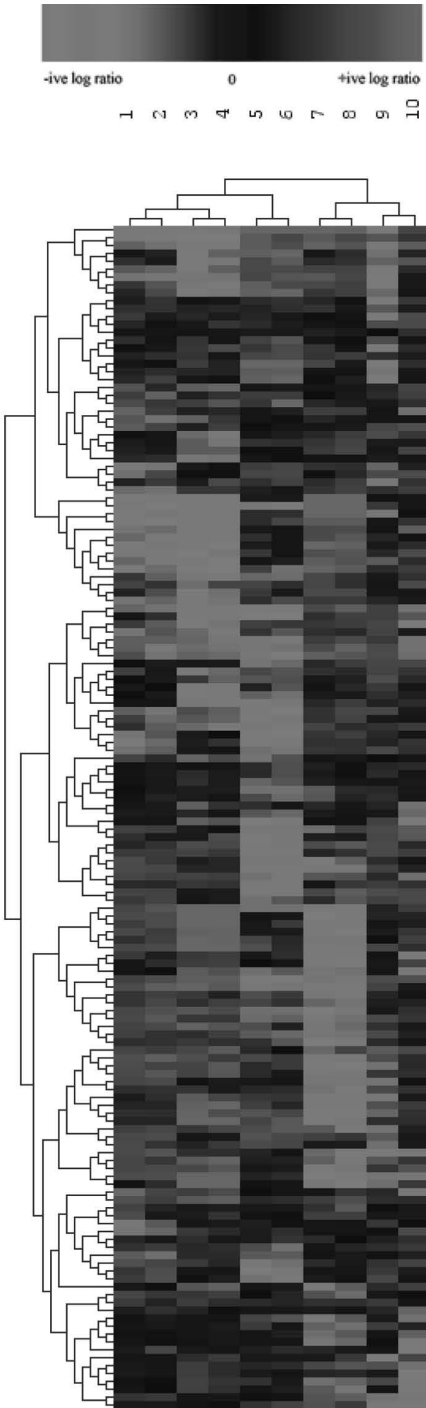


Fig. 10. A dendrogram derived from hierarchical clustering. Hierarchical trees have been constructed for both gene elements (rows) and experiments (columns). Shorter branches indicate smaller distances between the expression vectors, and a closer relationship.

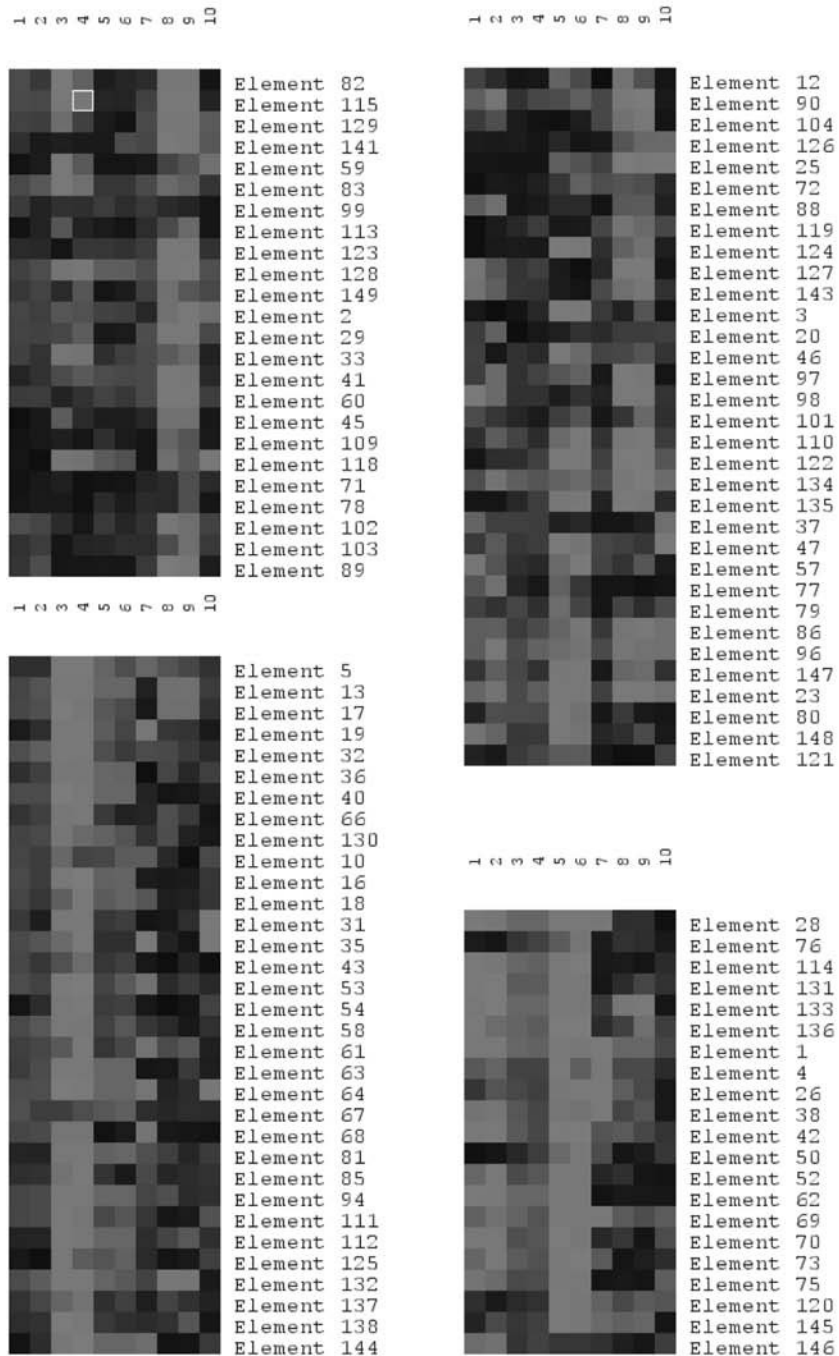


Fig. 11. (Continued)

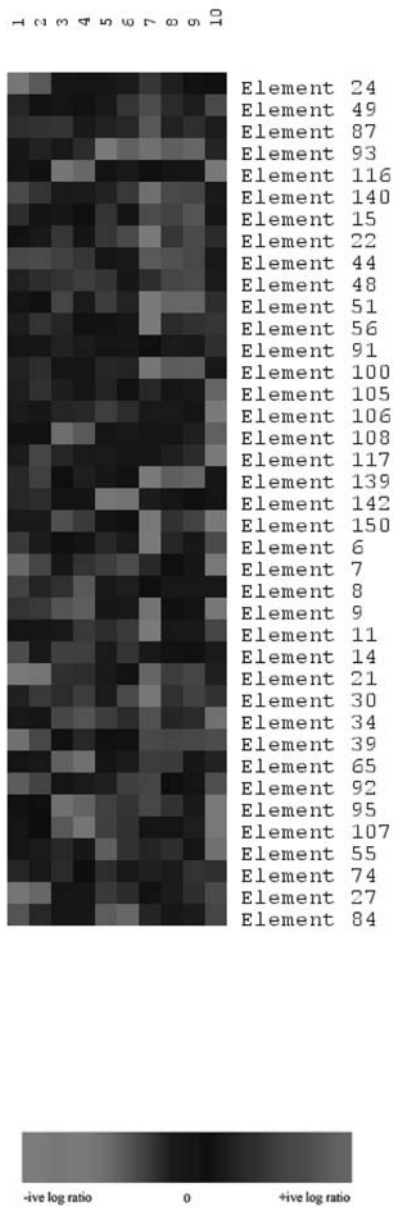


Fig. 11. *k*-Means clustering. A *k* of five was chosen, and five clusters were produced. Two elements from within the same cluster have a similar appearance, whereas two elements from different clusters will look less alike.

3. Perform **steps 3a,b** once for every element, in turn. A single iteration of this step involves the evaluation all elements.
 - a. Select an element and find the cluster with which it has the most similarity (i.e., into the cluster with a mean expression vector least distant from the element's own). If the element is not already a member of this cluster, move it there.
 - b. If the element was moved into a different cluster, recalculate the mean expression vector for the cluster it was moved from and for the cluster to which it was moved. Continue to **step 3a**.
4. If no elements were moved during the most recent iteration of **step 3**, then all elements are currently in their most ideal clusters and the algorithm is finished. Otherwise, begin the next iteration of **step 3**.

The result of this algorithm is a collection of k clusters (**Fig. 11**), each containing the elements that most closely matched the cluster's mean expression vector at the time each element was assigned.

There are many software packages available that give the user the ability to perform analyses similar to those described in **Subheadings 7.4** and **7.5**. Recommended open-source systems (**46**) include TM4 (<http://www.tm4.org>; **ref. 47**), BioConductor (<http://www.bioconductor.org>; **ref. 48**), and BASE (<http://base.thep.lu.se>; **ref. 49**). Each of these systems are available free of charge and source code is provided (additional visualization and analysis schemes are available, and we refer the reader to the pertinent reviews in **refs. 50** and **51**).

It is important to note that the results of clustering algorithms are merely mathematical interpretations of the data and may not necessarily correlate with biological organizations. The algorithms that are chosen in the course of the analysis of a data set, as well as the specific parameter settings used, will have a significant effect on the conclusions that can be drawn from the analysis. Such conclusions should not be taken as absolute facts but rather as hypotheses that can be further examined.

References

1. Liang, P. and Pardee, A. B. (2003) Analysing differential gene expression in cancer. *Nat. Rev. Cancer* **3**, 869–876.
2. Miller, L. D., Long, P. M., Wong, L., Mukherjee, S., McShane, L. M., and Liu, E. T. (2002) Optimal gene expression analysis by microarrays. *Cancer Cell* **2**, 353–361.
3. Roth, M. E., Feng, L., McConnell, K. J., et al. (2004) Expression profiling using a hexamer-based universal microarray. *Nat. Biotechnol.* **22**, 418–426.
4. Fan, J. B., Yeakley, J. M., Bibikova, M., et al. (2004) A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome Res.* **14**, 878–885.
5. Lockhart, D. J., Dong, H., Byrne, M. C., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680.

6. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* **21(Suppl)**, 20–24.
7. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.
8. Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E., and Davis, R. W. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**, 301–306.
9. Watson, A., Mazumder, A., Stewart, M., and Balasubramanian, S. (1998) Technology for microarray analysis of gene expression. *Curr. Opin. Biotechnol.* **9**, 609–614.
10. Southern, E., Mir, K., and Shchepinov, M. (1999) Molecular interactions on microarrays. *Nat. Genet.* **21**, 5–9.
11. Hess, K. R., Zhang, W., Baggerly, K. A., Stivers, D. N., and Coombes, K. R. (2001) Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol.* **19**, 463–468.
12. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32(Suppl)**, 496–501.
13. Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., and Simon, R. (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics* **4**, 33.
14. Wang, H. Y., Malek, R. L., Kwitek, A. E., et al. (2003) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biol.* **4**, R5.
15. Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., and Sealfon, S. C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48.
16. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412.
17. Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr., et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684.
18. Chuaqui, R. F., Bonner, R. F., Best, C. J., et al. (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.* **32(Suppl)**, 509–514.
19. Eberwine, J., Yeh, H., Miyashiro, K., et al. (1992) Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. USA* **89**, 3010–3014.
20. Randolph, J. B. and Waggoner, A. S. (1997) Stability, specificity and fluorescence brightness of multiply-labeled fluorescent DNA probes. *Nucleic Acid Res.* **25**, 2923–2929.
21. Hughes, T. R., Mao, M., Jones, A. R., et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347.
22. Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W., and Stromberg, A. J. (2003) Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* **4**, 26.

23. Stears, R. L., Getts, R. C., and Gullans, S. R. (2000) A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol. Genomics* **3**, 93–99.
24. Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T., and Marincola, F. M. (2000) High-fidelity mRNA amplification for gene profiling. *Nat. Biotech.* **18**, 457–459.
25. Xiang, C. C., Chen, M., Ma, L., et al. (2003) A new strategy to amplify degraded RNA from small tissue samples for microarray studies. *Nucleic Acids Res.* **31**, e53.
26. Kerr, M. K. and Churchill, G. A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**, 123–128.
27. Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32(Suppl)**, 490–495.
28. Dobbin, K., Shih, J. H., and Simon, R. (2003) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J. Natl. Cancer Inst.* **95**, 1362–1369.
29. Simon, R. M. and Dobbin, K. (2003) Experimental design of DNA microarray experiments. *Biotechniques (Suppl)*, 16–21.
30. Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3**, 579–588.
31. Talaat, A. M., Howard, S. T., Hale, W., 4th, Lyons, R., Garner, H., and Johnston, S. A. (2002) Genomic DNA standards for gene expression profiling in *Mycobacterium tuberculosis*. *Nucleic Acids Res.* **30**, e104.
32. Desu, M. M. and Raghavarao, D. (eds.) (2003) *Nonparametric Statistical Methods for Complete and Censored Data*. Chapman & Hall/CRC Press, Boca Raton, FL.
33. Eickhoff, B., Korn, B., Schick, M., Poustka, A., and van der Bosch, J. (1999) Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res.* **27**, e33.
34. Yue, H., Eastman, P. S., Wang, B. B., et al. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.* **29**, E41-1.
35. Badiie, A., Eiken, H. G., Steen, V. M., and Lovlie, R. (2003) Evaluation of five different cDNA labeling methods for microarrays using spike controls. *BMC Biotechnol.* **3**, 23.
36. Benes, V. and Muckenthaler, M. (2003) Standardization of protocols in cDNA microarray analysis. *Trends Biochem. Sci.* **28**, 244–249.
37. Cui, X. and Churchill, G. A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210.
38. Nadon, R. and Shoemaker, J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271.
39. 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.
40. Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.* **74**, 829–836.

41. Yang, Y. H., Dudoit, S., Luu, P., et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.
42. 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.
43. Wen, X., Fuhrman, S., Michaels, G. S., et al. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* **95**, 334–339.
44. Soukas, A., Cohen, P., Succi, N. D., and Friedman, J. M. (2000) Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev.* **14**, 963–980.
45. Aronow, B. J., Toyokawa, T., Canning, A., et al. (2001) Divergent transcriptional responses to independent genetic causes of cardiac hypertrophy. *Physiol. Genomics* **6**, 19–28.
46. Dudoit, S., Gentleman, R. C., and Quackenbush, J. (2003) Open source software for the analysis of micorarray data. *BioTechniques* **34(Suppl)**, 45–51.
47. Saeed, A. I., Sharov, V., White, J., et al. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378.
48. Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.
49. Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., and Peterson, C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3**, SOFTWARE0003.
50. Sherlock, G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* **12**, 201–205.
51. Hughes, T. R. and Shoemaker, D. D. (2001) DNA microarrays for expression profiling. *Curr. Opin. Chemical Biol.* **5**, 21–25.